DOI: http://dx.doi.org/10.7551/978-0-262-31709-2-ch135

A hierarchical support vector machine based on feature-driven method for speech emotion recognition

Lingli Yu[1], Binglu Wu[1] and Tao Gong[2,3,4]

[1] School of Information Science and Engineering, Central South University, Changsha, 410083, P.R. China
[2] College of Information S. & T., Donghua University, Shanghai 201620, China
[3] Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education, Donghua University, Shanghai 201620, China
[4] Department of Computer Science, Purdue University, West Lafayette 47907, USA
llyu@csu.edu.cn taogong@dhu.edu.cn

Abstract

Through the analysis of one-vs.-one, one-vs.-rest and the decision tree mechanism of binary support vector machine emotion classifiers, a method based on feature-driven hierarchical support vector machine is proposed for speech emotion recognition. For each layer, classifier used different feature parameters to drive its performance, and each emotion is subdivided layer by layer. This method did not rely entirely on the activity-valance dimensional emotion model, but relied on the type of emotion to distinguish. Furthemore, classifications are constructed by appropriate characteristic parameters ultimately. Experiments on the Chinese-speaker-dependent and Berlin-speaker-independent corpus reached conclusions as follows, Chinese-speaker-dependent recognition rate is relatively higher than Berlin-speaker-independent. feature-driven hierarchical support vector machine in the case driven by effective features improves the speech emotion recognition performance. Meanwhile applying the mean of the log-spectrum to this method can identify high-activity and low-activity emotion effectively.

Keywords: Feature-driven, Speech emotion recognition, Support vector machine (SVM), Hierarchical Classifier, Mean of the Log-Spectrum (MLS)

1. Introduction

Speech emotion recognition is not only used for human-computer interaction, but also applied in speech synthesis, artificial counseling, polygraph, telephone banking, driverless system and so on[1]. Nowadays, the field of emotion recognition is facing to huge challenges. The main difficulty [2] is that we could not extract a most effective universal phonetic feature for various kinds of emotion. Furthermore, one sentence may contain several kinds of emotions at the same time, and emotions may be associated with just parts of a sentence. There is no clear boundary for each complex emotion. Sometimes even humans are not capable of distinguishing them. Moreover, their cultural backgrounds and the environments also affect the emotion expression. Speech emotion recognition is a hot research issue in natural computing area, some researchers and institutions have done many works for emotion recognition [3].

Emotion recognition system consists of three parts: the module for extracting feature parameters, the module for reducing feature parameters' dimensions and the module for emotion recognition. Most researchers mainly utilized prosody features [4-6] like pitch period, short time energy, duration of voice, and their relative statistics as feature parameters. Besides, MFCC (Mel Frequency Cepstrum Coefficient) was also used for emotion recognition. This coefficient has more information when it is extracted from voiced sound rather than from unvoiced sound. Yang[7] proposed a coordinate feature set based on music theory for emotion recognition. They considered that the acoustic and semantic features were useful for the recognition. Another problem in speech emotion recognition is how to reduce the dimensions of the features in order to simplify the calculation. There were several common ways to do it: Sequential Floating Forward Selection (SFFS), Forward Feature Selection (FFS), Backward Feature Selection (BFS), Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA). After the simplified features obtained, we needed to build the effective classifiers. In 1990s, most of the emotional models were built on Maximum Likelihood Bayes (MLB) and Linear Discriminant Classification (LDC). Now, there are much more kinds of emotional models being used for emotion classification, like Hidden Markov Model (HMM), Gaussian Mixture Models (GMM), Artificial Neural Network (ANN), Support Vector Machine (SVM) etc. [8], and sometimes we use multiple methods simultaneously. However, we cannot know which classifier is the best at the current time.

SVM stemmed from statistical learning theory proposed by Vapnik and others, as a classifier. SVM could yield good results even from small test samples. So it was widely used for speech emotional recognition. Because of the Structural Risk Minimization, SVM classifier usually had better performance than others [9]. In this paper, an improved method based on feature driven is proposed in order to perfect speech emotional recognition performance. Feature-driven hierarchical SVM does not completely depend on the emotion dimensional model of activation-valence, and it adjusts the feature parameters of each layer gradually according to the property recognized by lots of experiments.

This paper is organized as follows. In section 2, several speech corpuses are described, and then Chinese Speaker-dependent and Berlin speaker-independent Speech Corpus are used in this paper. In section 3, three kinds of binary SVM emotion classifications are discussed for a multi-category classification problem of Speech emotion recognition. In section 4, Hierarchical SVM emotion recognition method based on feature-driven is proposed to improve the performance of Speech emotion recognition. We extract six feature parameter classes based on prosody affective features and acoustic affective features to achieve the hierarchical SVM classifiers analysis. Experimental comparison between 4 kinds of hierarchical methods and some analysis for improved the parameters in speech feature driven method are discussed, and compared in section 5. Finally, conclusions are summarized in section 6. The page limit is 8 pages for a full paper. Your submission must be converted to Portable Document Format (PDF). Please be sure to use highest portability and quality options. Papers that significantly deviate from these instructions will not be included.

2. Several binary SVM classifications

Support vector machines(SVMs) are one of supervised learning models with associated learning algorithms. Speech emotion recognition of is a multi-category classification problem, here it is converted to binary classification problem to solve one by one, the state-of-the-art including:

2.1 One-Versus-One binary SVM

The hyper planes of binary SVM are built from any two of all categories, so the number of binary SVM classifiers is $k*(k-1)/2$. Here 'max-wins' voting method is used, for One-Versus-One voting strategy, the $k*(k-1)/2$ binary SVM classifiers are trained in parallel, For example, category i and category j trains with classifier C_{ij}, C_{ij} decides whether sample x belong to category i or category j. Therefore the number of i category votes adds one, otherwise j's number of votes adds one. When the process is over, the category with the most voters is the right category that the sample belongs to. The structure of One-Versus-One binary SVM is shown in figure 1. Here 1-5 are set to represent 5 emotion categories of 2 speech corpuses, so 10 classifiers are trained. From the process of category, we know that this method is less effective while the number of the classifiers increases, which will cause the decision speed more slowly.

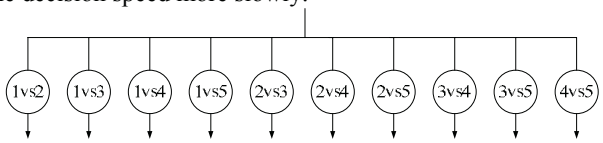


Figure 1: One-Versus-One 'max-wins' voting SVM

2.2 One-Versus-The-Rest binary SVM

One-Versus-The-Rest binary SVM only builds k SVMs, each SVM classifier recognizes one category from all the other categories. The unbalanced decision tree combines the One-Versus-The-Rest with right branch, and it just needs to train (k-1) classifiers, the number of classifiers is less than them in one-versus-one method. Figure 2 shows the structure of this method, the recognized emotion should be the easiest to be distinguished on first layer. Chinese Speaker-dependent speech corpus is used, furthermore anger is chosen on the first layer. For instance, 1 represents anger, 2 represents sadness, 3 represents happiness, 4 represents neutral and 5 represents amazement.

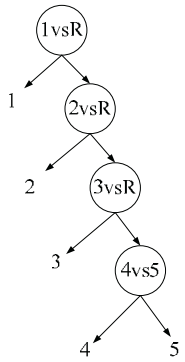


Figure 2: One-Versus-The-Rest unbalanced binary tree SVM

2.3 SVM decision Tree Mechanism

The classification error may occurs on any layer of these nodes, and it will spread to all the successor nodes, to this problem, DAG (directed acyclic graph) is proposed by Platt and others [11]. There are $k*(k-1)/2$ internal nodes, k branches, each node is a SVM binary classifier. For each test sample, each node's binary decision determines the path of the next decision from the root node. Figure 3 demonstrates the five categories' directed acyclic graph. Whatever the emotion of the test sample is, it will always reach to the bottom of the classifiers[12]. Here, the result is right when every classifier's result is right, but because each binary classifier just handles 2 different emotions, the training is simple and effective. where 1 represents anger, 2 represents sadness, 3 represents happiness, 4 represents neutral and 5 represents amazement in figure 3.

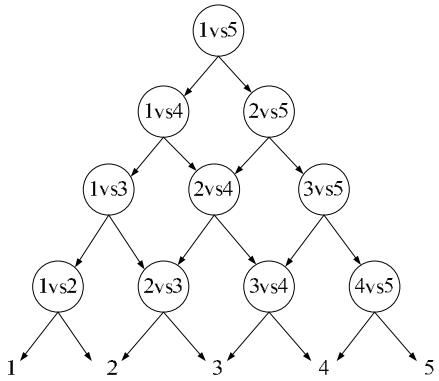


Figure 3: Directed acyclic graph SVM

3 Extracting the speech emotion characteristic parameters

Six kinds of feature parameters are extracted to study the hierarchical SVM classifiers based on prosody affective features and acoustic affective features.

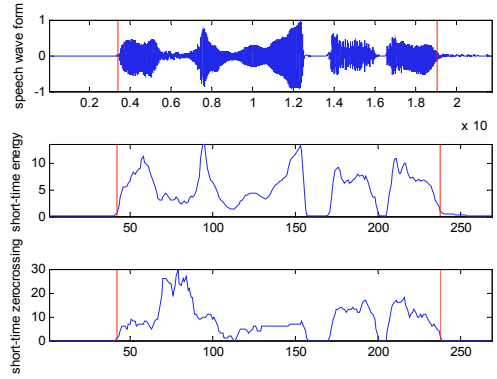
3.1 Prosody affective features

Prosody affective feature usually includes intensity, length or duration, pitch, accent, tone, intonation and rhythm.

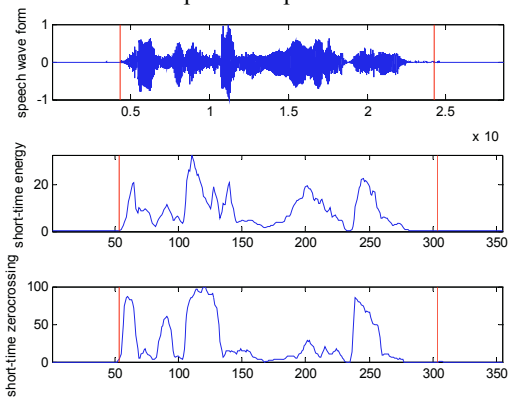
Short time energy and short time amplitude:

$$Energy(n) = \sum_{m=0}^{M-1} | S_n(m) |^2, n = 0,1,...,N-1 \quad (1)$$

Where S_n is the nth frame after enframing and windowing, N is the number of frames, M is the length of frame. From figure 4, we know that different emotion has different short time energy change, the same is true for short time amplitude.



(a) Sadness in Chinese speaker-dependent



(b) Anger in Chinese speaker-dependent

Figure 4. Voice activity detection using double threshold comparison method

Short time zero-crossing rate:

$$Z_n = \sum_{m=-\infty}^{\infty} | \mathrm{sgn}[x(n)] - \mathrm{sgn}[x(n-1)] | \cdot w(n-m) \quad (2)$$

In addition, short time zero-crossing rate combined with short time energy is used in hunting for starting point and end point of the sound, Figure 4 demonstrates this problem, when make final decision, the threshold is needed to set in real situation.

Pitch period:

$$\tau_N(k) = \frac{1}{N} \sum_{n=0}^{N-k-1} x(n)x(n+k), 0 \le k \le M_o - 1 \quad (3)$$

Where t_n is the value of autocorrelation function, $x(n)$ is the recognized voice signal, k is the delay time, M_o is the number of the autocorrelation which is calculated in [13]. After getting the autocorrelation function and detecting its peak, we can get pitch period. Figure 5 are shown for Chinese Speaker-dependent speech corpus's pitch period of the same sentence expressed by anger, sadness, neutral and amazement respectively, and the rate of change in anger and amazement is larger than in sadness and neutral, at the end of the sadness's pitch period envelop curve is cocking up.

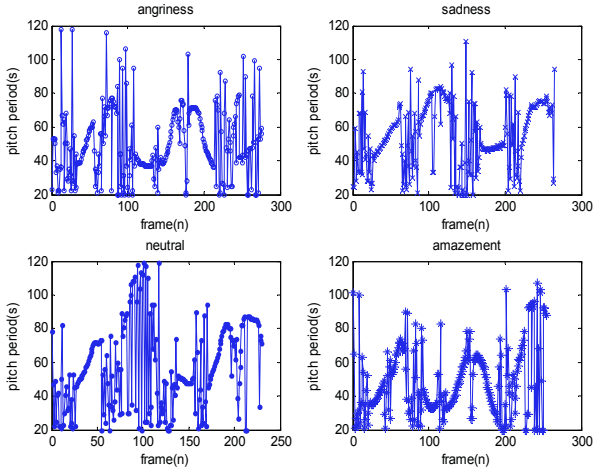


Figure 5. The pitch period of four kind emotions for Chinese speaker-dependent

3.2 Acoustic affective features

Acoustic affective features are generally the feature of tone and speech spectrum.

Formant:

$$P(w) = 20\log(\frac{\hat{1}}{|1+\sum_{k=1}^{p} a_k e^{-jwk}|^2}), k=1, 2, ...,p \quad (4)$$

Levinson-Durbin method is used to calculate the linear prediction coefficient $a_k(\ )$. The LPC spectrum are calculated by formula (4), only the first 3 formants are extracted.

MFCCs:

The extracting processes of Mel Frequency Cepstral Coefficients (MFCCs) are shown in figure 6.

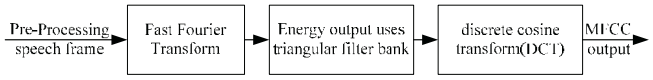
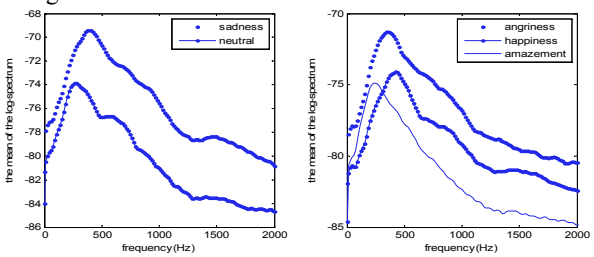


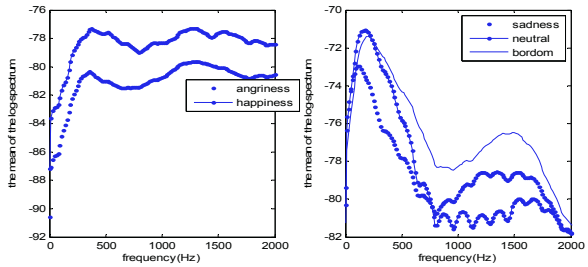
Figure 6: The MFCC extracting process

Mean of log power spectrum (MLS):

MLS is calculated by each frame's of one sentence, which is mainly used to convert time domain signal to frequency domain signal. Thus the average of MLS are calculated as showed in formula (5), where k is spectrum bandwidth, N_l is the lth class emotion's utterance number, $v_{il}(n,k)$ is the discrete Fourier transformation for the nth frame of signal i, the range of k's bandwidth is between 0 to 2000Hz.



(a) The average of the log-spectrum for Chinese speaker-dependent



(b) The average of the log-spectrum for Berlin speaker-independent.

Figure 7: The average of the log-spectrum

$$S_l(k) = \frac{1}{N_l}\sum_{i=1}^{N}\frac{1}{N_i}\sum_{n=1}^{N}\log|v_{il}(n,k)| \quad (5)$$

Figure 7 showed the average of the log-spectrum arranged by the similarity of spectrum. (a) is the 5 categories distributed situations of Chinese speaker-dependent. We can see that the MLS peak of anger, happiness concentrated to 400Hz, while the MLS peak of sadness and neutral concentrate to 200 to 250 HZ, MLS envelope curves are similar with each other, especially between high-activity and non-high-activity speech. (b) is the distributed situations of the log power spectrum mean of Berlin speaker-independent. They have similar envelope curves as they are both high-activity emotion. The MLS peak is also concentrated to 400Hz, while sadness, neutral and boredom are concentrated between 100 to 250 Hz. Now, six kinds of parameters and its different dimensions feature derivatives are extracted as showed in table 1.

Table 1: 247 dimensions global feature parameters

Basic feature parameters	statistical characteristic	feature dimensions
short time energy and amplitude	Mean, standard deviation, minimum, maximum, dynamic range, mean of first difference, standard deviation of first difference	14
short time zero crossing-rate	Mean, standard deviation, minimum, maximum, dynamic range, mean of first difference, standard deviation of first difference	7
pitch period	Mean, standard deviation, minimum, maximum, dynamic range, mean of first difference, standard deviation of first difference	7
MFCC and its first difference	Mean, standard deviation, minimum, maximum, dynamic range, mean of first difference, standard deviation of first difference	168
First three formants	Mean, standard deviation, minimum, maximum, dynamic range, mean of first difference, standard deviation of first difference	21
30 dimensions MLS	mean	30

3.3 Reducing dimensions of feature vector

The more dimensions the feature vector has, the more information it contains. However, the calculated complex also greatly increases with the dimensions increases. When the number of vector dimensions exceeds a certain limit, dimension disaster [14] would appear. Therefore, feature selection in broad definition is one kind of mapping transformation from the high-dimensional vector to low-dimensional vector for the sake of reducing dimensions. For literature [15], principal component analysis (PCA) is contributed for reducing dimensions, when PCA is applied to classifiers. it not only reduces calculated quantity, but also eliminates some interference factors. Here, select several characteristic vectors as a main component vectors that correspond to the first k characteristic value, and d is the vector dimensions, we set k/d equals to 0.95, so the number of feature vector's dimensions are reduced from 247 to 31 using PCA in this paper.

4. Hierarchical SVM emotion recognition method based on feature-driven

It is shown in Figure 4, A feature-driven hierarchical SVM is proposed for emotion recognition, which demonstrates the structure of this method. In this paper, five kinds of emotions are subdivided for three layers. Especially, Feature-driven hierarchical SVM does not completely depend on the emotion dimensional model of activation-valence, and it adjusts the feature parameters of each layer gradually according to the property recognized by lots of experiments and experience. This feature-driven method is similar with unbalanced decision tree, however, the number of hierarchical layers decreases. A feature-driven hierarchical SVM is strict to each classifier of each layer, generally, we set the two easiest distinguished main categories as the first layer. Therefore, the performance of each classifier of each layer should be well enough to guarantee test samples correctly before it enters the next layer. Meanwhile, linear kernel function is utilized for hierarchical SVM based on feature-driven in this paper.

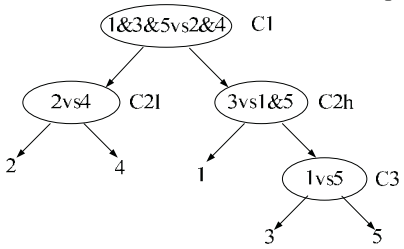


Figure 8: Feature-driven hierarchical SVM

In figure 8, 1 represents anger, 2 represents sadness, 3 represents happiness, 4 represents neutral and 5 represents amazement. The structural parameters are as following for Chinese Speaker-dependent speech corpus. Here,

(1) Classifier C1 distinguishes anger, happiness and amazing emotion from sadness, neutral emotion. The feature parameters are the combination of short time energy, short time amplitude, short time zero-crossing rate, pitch period, and MLS (Mean of the Log-Spectrum).

(2) Classifier C21 distinguishes sadness from neutral. In addition, the feature parameters of MFCC and formant are the combination.

(3) Classifier C2h distinguishes anger from happiness and amazement. Therefore, parameters includes short time energy, short time amplitude, pitch period, short time zero-crossing rate, mean value of 12 dimensions MFCC and formant combine the feature parameter.
(4) Classifier C3 distinguishes happiness from amazement. The feature parameter is all the feature parameters of C2h combined with MLS.
In the Berlin Speaker-independent speech corpus: 1 represents anger, 2 represents neutral, 3 represents happiness and 4 represents boredom. The structural parameters of Berlin Speaker-independent speech corpus follows. Here
(1) Classifier C1 distinguishes sadness, neutral and boredom from anger and happiness. The feature parameters of short time energy, short time amplitude, pitch period, short time zero-crossing rate and MLS (Mean of the Log-Spectrum) are the combination.
(2) Classifier C21 distinguishes anger from happiness. The feature parameters are the combination with short time energy, short time amplitude, pitch period, short time zero-crossing rate, MLS, and the mean of MFCC.
(3) Classifier C2h distinguishes sadness from neutral and boredom. The feature parameters are the combination of short time energy, short time amplitude, pitch period, short time zero-crossing rate, MFCC and formant.
(4) Classifier C3 distinguishes neutral from boredom, so the feature is the same as C2h.

5 Experimental test and result analysis

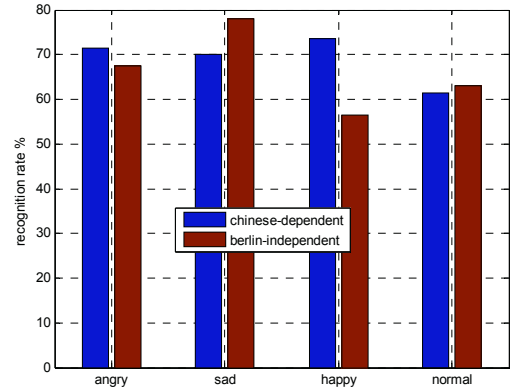
The orthogonal method is adopted to ensure the independent of each training and test samples. For the experimental process, we choose 50 sentences for each emotion, training samples using 30 of them, test samples using the rest 20. 50 sentences are labeled with 1-50, and every 10 sentences as one group, then there are totally C_5^3 =10 situations in the combination of training samples and test samples. Therefore, the 10 independent experiments are conducted to reduce the effect of the unbalance in speech corpus.

5.1 Speech Corpus

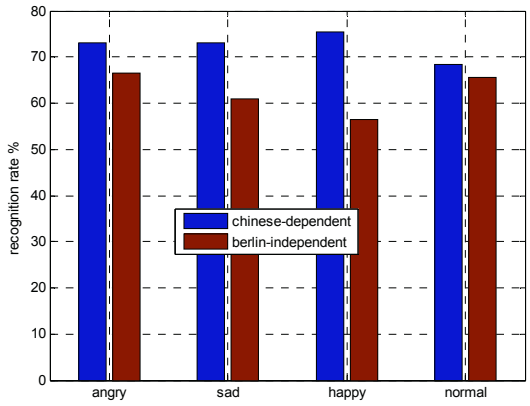
Chinese Speaker-dependent and Berlin Speaker-independent Speech Corpus are selected for our experiments. The term of speaker-dependent indicates that all the speech come from one person, which means tones and pronunciation habits are all the same. The Chinese speech corpus records the voice of a woman and includes 5 types of emotions: anger, sadness, happiness, amazement and neutral. Each type of emotion has 50 sentences. The corpus is saved as .wav format with 16 kHz-16 bit resolution. The Berlin corpus [10] records the voice of five man and five women and includes seven types of emotions: happiness, sadness, anger, boredom, disgust, fear and neutral. The corpus is also saved as .wav format with 16 kHz -16 bit resolution.

5.2 Comparison between Chinese Speaker-dependent and Berlin Speaker-independent speech corpus

We focus on the fallibility between anger, happiness and amazement in Chinese Speaker-dependent speech corpus, and the fallibility between neutral and sadness. In Berlin Speaker-independent speech corpus, the anger is hard to distinguish from happiness. Meanwhile, the difference between sadness, neutral and boredom are also hard to detect. Therefore, we design the experiments to these two kinds of speech corpus using one versus one method and feature driven method respectively, the results are showed in figure 9. As the amazement in Chinese corpus and boredom in Berlin corpus is not on the same feature space in emotional model, so that just four kinds of emotions (anger, sadness, happiness and neutral) are compared in figure 9.



(a)1vs1-voting mechanism



(b) feature-driven method

Figure 9: Comparison results of for Chinese and Berlin Corpus

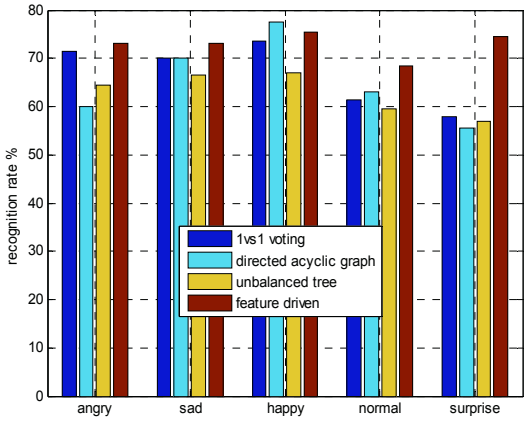
From the results of figure 9, especially for feature driven SVM, we find that the recognition rate for speaker-dependent is obviously higher than for speaker-independent. This is mainly caused by different personal pronunciation habits. Hence, when emotional recognition for speaker-dependent is extended to speaker-independent, the effect of personal pronunciation should be eliminated in feature parameters. As is shown in figure 9, we also know that recognition rate for speaker-independent in one-versus-one mechanism does not decline evidently, which is shown that the 1vs1 recognition method is also available. Combined figure 9 (a) and (b), the emotion recognition rate for speaker-dependent in feature-driven method is almost higher than 1vs1-voting mechanism except for sadness. Meanwhile, the recognition rate for speaker-dependent is also higher than for speaker-independent. Those cause by two reasons. Firstly, there are more feature parameters fused in one-versus-one mechanism each layer which influences the recognition rate. Secondly,

the recognition feature parameters of each layer are more appropriate for Chinese vocal features than German in feature driven SVM.

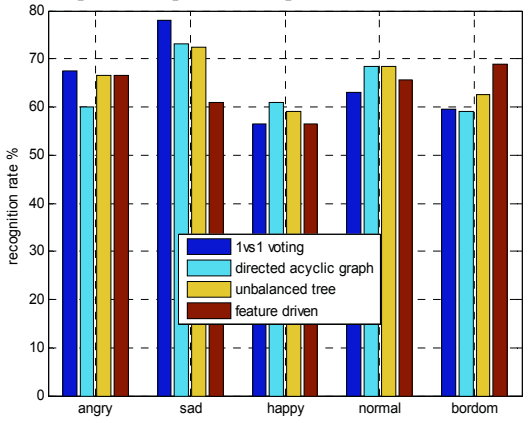
5.3 Experimental comparison between four kinds of hierarchical methods

The recognition results of four kinds of hierarchical structure are shown in figure 10. The feature driven method not only keeps the recognition rate for anger, sadness and happiness, but also improves the recognition performance for neutral and amazement. it proves that feature driven method applied to Chinese Speaker-dependent person speech corpus is more effective.

From figure 10, the recognition rate of the one-versus-one mechanism is the highest among all four methods. In addition, recognition rates of unbalanced binary tree and directed acyclic graph are the almost same. The recognition performance of feature driven method is not satisfactory, especially for the sadness and happiness. Therefore, the main reason is that the Chinese pronunciation habit is different from Germany. As we all known that the pause time or silent segments in German is longer than it in Chinese, so the chosen parameters may fit for Chinese but not fit for German.



(a) Chinese Speaker-dependent corpus



(b) Berlin Speaker-independent corpus

Figure 10: Comparison between four kinds of methods

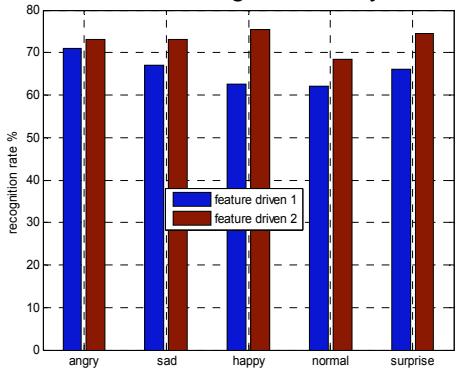
5.4 Experiments after improved the parameters in feature driven method of speech emotion recognition

Feature parameters in every layer of hierarchical SVM based on feature-driven are devised and modulated respectively for extracting those feature parameters fitter, meanwhile, for improving the recognition performance. The results after devised and modulated for parameters in feature driven method are showed in table 2.

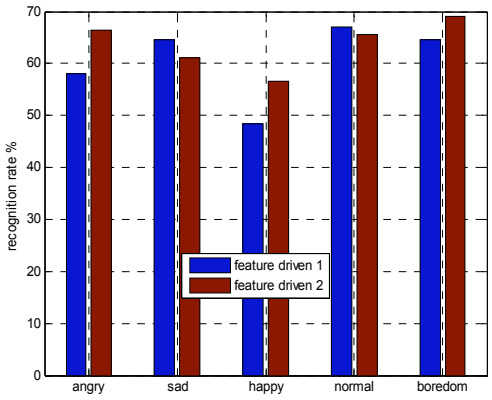
Table 2: The parameters adjustment and the rate of error identification of feature-driven method

classifier		Chinese speaker-dependent corpus		Berlin speaker-independent Corpus	
C1	feature parameters	ZEP	ZEP + MLS	ZEP	ZEP + MLS
	Error recognition rate	8.7%	6.9%	10.3%	8.1%
C2l	feature parameters	MFCC	MFCC + Formant	MFCC	MFCC + ZEP & Formant & MLS
	Error recognition rate	24.34%	22.25%	37.17%	31.09%
C2h	feature parameters	ZEP & MFCC	ZEP & MFCC + Formant	ZEP & MFCC	ZEP & MFCC + Formant
	Error recognition rate	18.57%	16.49%	20.61%	21.35%
C3	feature parameters	ZEP	ZEP & MFCC + Formant	ZEP	ZEP & MFCC + Formant & MLS
	Error recognition rate	19.34%	12.96%	16.24%	15.93%

Here, ZEP represents the combination of short time energy, short time amplitude, pitch period and short zero-crossing rate. Here, given C1 as a instances for Chinese speech corpus to illustrate "Error recognition rate", when high valence emotion (anger, happiness and amazement) is judged to be low valence emotion (neutral, sadness), then it is considered error recognition, and vice versa. A computational method for another classifier is built through similar way.



(a) Chinese Speaker-dependent



(b) Berlin Speaker-independent

Figure 11: Results after improving parameters in Feature-driven methods

Figure 11 is the comparison between before improved and after improved feature driven method. Combining the figure 11 and table 2, we know that the recognition rate is greatly improved after importing the MLS feature parameter. Furthermore, the recognition rate of C21 and C2h classifiers in Chinese Speaker-dependent speech corpus improved after importing the formant, but the recognition rates of the two classifiers are adverse in Berlin Speaker-independent speech corpus. This indicates that a better feature parameters extracted is related to the type of language closely.

6. Conclusion

A new feature driven hierarchical SVM classifier is devised for emotion recognition. Here Chinese Speaker-dependent and Berlin Speaker-independent speech corpuses are used for experimental study. Meanwhile, the mean of the log-spectrum (MLS) is particularly used to improve the feature driven SVM classifier. Since SVM isn't used to recognize multiple emotions directly. Therefore, we set ordinary binary SVM classifier as a contrast experiment to feature driven hierarchical SVM. Then we calculated the recognition rate respectively and analyzed the potential problems. However, the following problems still need further study. (1) All global feature parameters are extracted through the same statistical features, there may be conflict between statistical features and some feature's impaction may be reduced when reducing dimensions through PCA. It is a research direction to use other methods such as SFS algorithm to reduce the dimensions. (2) How to extracts effective feature parameters. The feature parameters extracted in this paper still can't separate anger, happiness from amazement and sadness from neutral very clearly. (3) Linear kernel function is utilized for binary SVM in this paper, for better performance, new kernel paper may be tried.

Acknowledgements

This paper was supported by the National Natural Science Foundation of Hunan (13JJ4018, 13JJ4093), the Fundamental Research Funds for the Central Universities (2012QNZT060), and the youth Foundation of education bureau of Hunan province (11B070).

References

[1]N. Kamaruddin, A. Wahab, Ch. Quek (2012). Cultural dependency analysis for understanding speech emotion, *Expert Systems with Applications*, 39: 5115-5133,.

[2]M. E. Ayadi, M. S. Kamel, F. Karray (2011). Survey on speech emotion recognition Features, classification schemes and databases, *Pattern Recognition.*44: 572-587.

[3]W. Han, H. F. Li (2009),Speech emotion recognition based on prosodic segment level features , *Journal of Tsinghua University (Science and Technology)*, S1:1363-1368.

[4]T. Iliou, C.N. Anagnostopoulos (2010). SVM-MLP-PNN classifiers on speech emotion recognition field-A comparative study, *The Fifth International Conference on Digital Telecommunications*, Athens/Glyfada, Greece,

[5]B. Schuller, G. Rigoll, M. Lang (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, *IEEE International Conference on Acoustics, Speech, and Signal Processing.*

[6]E. Bozkurt, E. Erzin, C. E. Erdem, A. T. Erdem (2011), Formant position based weighted spectral features for emotion recognition, *Speech Communication*, 53:1186-1197.

[7]B. Yang, M. Lugger (2010), Emotion recognition from speech signals using new harmony features, *Signal Processing*, 90:1415-1423.

[8]M. Sheikhan, M. Bejani, D. Gharavian (2012), Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method, *Neural Comput & Applic*, 8:1-13.

[9]S. Chandaka, A. Chatterjee, S. Munshi (2009), Support vector machines employing cross-correlation for emotional speech recognition, *Measurement*,42:611-618,.

[10] Burkhardt, F. Paeschke, A Rolfes, M. Sendlmeier, W. Weiss, B. (2005), A database of German emotional speech, *The 9th European Conference on Speech Communication and Technology*, Euro speech Inter speech Lisbon, Portugal.

[11] J. Platt, N. Cristianini, J. Shawe-Taylor (2000). Large margin DAGs for multi-class classification, *Proceedings of Neural Information Processing Systems,* Denver,.

[12] A. Hassan, R. I. Damper(2012). Classification of emotional speech using 3DEC hierarchical classifier, *Speech Communication*,24:1-10.

[13] K. C. Huang, Y. H. Kuo (2010). A Novel Objective Function to Optimize Neural Networks for Emotion Recognition from Speech Patterns, *The Second World Congress on Nature and Biologically Inspired, Computing* , Kitakyushu, Fukuoka, Japan.

[14] E. M. Albornoz, D. H. Milone, H. L. Rufiner (2011), Spoken emotion recognition using hierarchical classifiers, *Computer Speech and Language*, 25:556-570.

[15] H.B. Yao, L. Tian (2003). A genetic-algorithm-based selective principal component analysis (GA-SPCA) method for high-dimensional data feature extraction, *IEEE Transactions on Geoscience and Remote Sensing*, 41: 1469-1478.

Lingli YU is working in the institute of Intelligence Science and Technology (IST), School of Information Science and Engineering of Central South University, Changsha, 410083, P.R. China. Her current research interests include speech emotion recognition, mobile-robot fault diagnosis. E-mail: llyu@csu.edu.cn.

Binglu WU is studying in the institute of Intelligence Science and Technology (IST), School of Information Science and Engineering of Central South University, Changsha, 410083, P.R. China. Her current research interests include speech emotion recognition. E-mail: wubingluqipu@163.com.

Tao Gong is Associate Professor of Immune Computation at the College of Information Science and Technology, Donghua University. He holds Top Award of Baosteel Education Fund for his work on immune computation. Tao is the founder and General Editor-in-Chief (EiC) of Immune Computation, which is the first leading journal in the field of immune computation. Tao also acts as a Vice-Chair of IEEE Computer Society Task Force on Artificial Immune Systems. Tao is a Life Member of Sigma Xi, The Scientific Research Society and Shanghai Chenguang Scholar. Tao has worked as the chairs and committee members of many international conferences on artificial immune systems and computer systems.