Digital Signal Processing 22 (2012) 1154–1160

Contents lists available at SciVerse ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Speech emotion recognition: Features and classification models

Lijiang Chen [a], Xia Mao [a,*], Yuli Xue [a], Lee Lung Cheng [b]

[a] School of Electronic and Information Engineering, Beihang University, Beijing, China
[b] Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Article history:
Available online 24 May 2012

Keywords:
Emotion recognition
Speaker independent
Fisher discriminant
SVM

ABSTRACT

To solve the speaker independent emotion recognition problem, a three-level speech emotion recognition model is proposed to classify six speech emotions, including sadness, anger, surprise, fear, happiness and disgust from coarse to fine. For each level, appropriate features are selected from 288 candidates by using Fisher rate which is also regarded as input parameter for Support Vector Machine (SVM). In order to evaluate the proposed system, principal component analysis (PCA) for dimension reduction and artificial neural network (ANN) for classification are adopted to design four comparative experiments, including Fisher + SVM, PCA + SVM, Fisher + ANN, PCA + ANN. The experimental results proved that Fisher is better than PCA for dimension reduction, and SVM is more expansible than ANN for speaker independent speech emotion recognition. The average recognition rates for each level are 86.5%, 68.5% and 50.2% respectively.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

Affective interaction is the high-level phase of human computer interaction (HCI). Picard invented the term *Affective Computing* to describe a newly established field dealing with the automatic sensing, recognition and synthesis of human emotions from any biological modality such as speech or facial expressions [1]. It is an interdisciplinary field widely involving computer science, psychology, and cognitive science [2–4]. In recent years, a great deal of research has been done to automatically recognize emotions from human speech [5–7]. Some of these researches have been further applied to call centers, multi-agent systems or other areas [8–12]. However, most of them relate only to speaker dependent recognition. Speaker independent emotion recognition is a difficult issue. In a survey conducted to measure human performance on emotion recognition, only 60 percent people can correctly determine the expressed emotions of unknown people [13]. There are two key phases to recognize emotions in speech signals, one is to find effective speech emotion features, the other is to establish proper mathematic model for speech emotion recognition.

An important issue in speech emotion recognition is the extraction of speech features that efficiently characterize the emotional content of speech and at the same time do not depend on the speaker or the lexical content. In existing researches, prosody features play an important role in emotion recognition, including pitch, intonation, accent, mute and speed [14]. All of these features are in low-frequency domain. In addition, some high-frequency features can also express some kinds of emotions, such as deep breath sounds for anger and tremble voice for fear. Schuller combined acoustic features and linguistic information for speech emotion recognition [15]. Our research group has adopted multi-time domain features [16]. We also find some new effective features using Parametric Filter and Fractal Dimension [17]. However, all these researches have used the same features for any kind of emotions. In fact, a relevant feature which is good to discriminate a certain pair of emotional classes may fail in classifying another pair of emotional classes. In this paper, we will use Fisher Discriminant to expose the relationship between features and emotions.

Existing speech emotion recognition researches usually divide utterances into several species of emotions. The most famous models are Ekman's six basic emotions model and Fox's multi-level emotional model [18,19]. This paper will combine these two emotion models to establish a three-level emotion recognition model. Each level of this model contains one or two classifiers for pairwise emotion classification. The lower levels carry out rough classification, while the higher ones complete precise classification. Ultimately, the six basic emotional categories are to be distinguished.

The paper is structured as follows: Section 2 gives the description of the databases we use; Section 3 illustrates the feature extraction; Section 4 introduces the dimension reduction method including modified Fisher discriminator and PCA; Section 5 introduces the structure of proposed multi-level system; finally, we present the results of comparative experiments and the conclusions.

* Corresponding author.
E-mail address: moukyoucn@yahoo.com.cn (X. Mao).

1051-2004/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.dsp.2012.05.007

2. Database description

Beihang University Database of Emotional Speech (BHUDES) is used in our research [20]. It contains Mandarin utterances of six emotions including sadness (sad.), anger (ang.), surprise (sur.), fear (fea.), happiness (hap.) and disgust (dis.). All these utterances are performed by seven actors and eight actresses whose ages are between 20 and 25. Each speaker repeats twenty neutral texts by three times with each emotion, meaning that $15 \times 20 \times 3 \times 6 = 5400$ utterances are obtained. All these utterances have a sample frequency of 11 025 Hz and a mean duration of 1.2 s.

3. Feature extractions

Firstly, we extract instantaneous features. Instantaneous features are based on partitioning utterance into frames. Each frame has the length of 256 samples and frame shift of 128 samples. For each frame, energy (E), zero crossing rate (Z_r), energy times zero crossing rate (E_z), pitch (P), the first to third formants (F_{1-3}), spectrum centroid (S_c), spectrum cut-off frequency (S_f), correlation density (C_d), fractal dimension (F_d), five Mel-frequency bands energy (E_{M1-M5}) are extracted.

Secondly, the first derivative and the second derivative of the 16 features are obtained. Assume 0 as the index of the original features, 1 and 2 as indexes of the first derivative and the second derivative features.

Finally, we get utterance-level features through calculating the statistic of the instantaneous feature, including maximum (max.), minimum (min.), mean (mea.), standard deviation (std.), skewness (ske.) and kurtosis (kur.). Thereby 288 candidate features are obtained.

3.1. Pre-processing

In order to equalize the effect of the propagation of speech through air, a pre-emphasis radiation filter is used to process speech signal before extraction of features. The transfer function is given by (1):

$$H(z) = 1 - 0.97z^{-1} \tag{1}$$

In addition, to reduce ripples in the spectrum of the speech spectrum, each frame is multiplied by a Hamming window before feature extraction.

3.2. Traditional features

Energy, zero crossing rate, pitch, formants are traditional speech signal features. Energy and pitch are prosody features within low-frequency domain. Zero crossing rate and formants are high-frequency features. Energy times zero crossing rate contains information in both energy and zero crossing rate.

3.3. Spectrum centroid and spectrum cut-off frequency

Spectrum centroid and spectrum cut-off frequency are frequency domain features. They both reflect the frequency distribution characteristics of speech signals.

Firstly, carry out Fast Fourier transform for each data frame, as given by Eq. (2):

$$F(k) = \left\| \sum_{i=1}^{N} x(n) \times e^{-i2\pi k \frac{n}{N}} \right\| \quad (k = 1, 2, \ldots, N) \tag{2}$$

where $x(n)$ is the input speech signal with the length N equals to 256. $F(k)$ is the amplitude of spectrum.

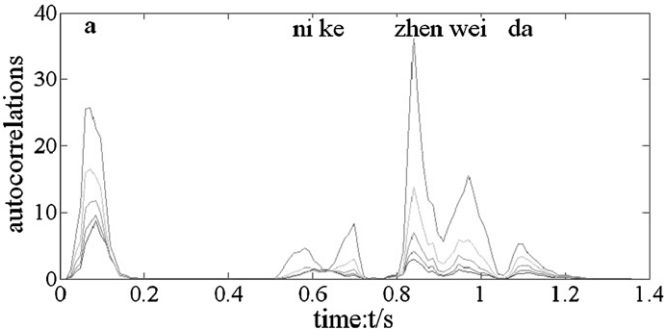


Fig. 1. Frame-level autocorrelation.

Spectrum centroid is defined as (3):

$$S_c = \left(\sum_{k=1}^{N/2} k \times F(k) \right) \Big/ \left(\sum_{j=1}^{N/2} F(j) \right) \tag{3}$$

Spectrum cut-off frequency S_f satisfies (4):

$$\left(\sum_{k=1}^{S_f} F(k) \right) \Big/ \left(\sum_{j=1}^{N/2} F(j) \right) = 0.85 \tag{4}$$

3.4. Correlation density and fractal dimension

Correlation density reflects the short-time speech signal spectral distribution. Fractal dimension reflects the nonlinearity of speech signal [17]. The algorithm to calculate the correlation density is given in detail as follows:

- Design a group of one-pole filters as Eq. (5), each pole has the same angle but different amplitude.

  $$\alpha = \eta \times e^{j\theta} \tag{5}$$

  where $\eta = \{0.125, 0.25, 0.5, 0.75, 0.875\}$, $\theta = 0.05\pi$ (corresponding frequency ≈ 275.6 Hz, which equal to the mean of pitch both of male and female).
- Filter the speech signal $X(i)$ using the designed one-pole filters as Eq. (6). Five output speech signals whose low-frequency components are upgraded to various degrees are obtained.

  $$Y_k(i) = \alpha(k) \times Y(i-1) + X(i) \tag{6}$$

  where i is the index of the speech signal and k is the index of α.
- Five frame-level autocorrelation contours derived from the outputs of the one-pole filters are displayed in Eq. (7) and Fig. 1.

  $$\rho_k(t) = \sum_{i=1}^{L} Y_k(i) \times W(i) \times Y_k(i+1) \times W(i+1) \tag{7}$$

  where t is the index of the current frame, $W(i)$ is a Hanning window whose length is L which is the same as the current frame.

  The six main peaks in Fig. 1 corresponding to six voiced sounds or vowels which are *a*, *i*, *e*, *en*, *ei*, *a*. The first, fourth and fifth vowels are strengthened and the duration of the first vowels is lengthened compared to the average level. These changes help the speaker to express a certain emotion.
- Calculate correlation density using Eq. (8):

  $$C_d(t) = \log \left\{ \sum_{k=1}^{4} \left[\rho_{k+1}(t) - \rho_k(t) \right]^{-2} \right\} \tag{8}$$

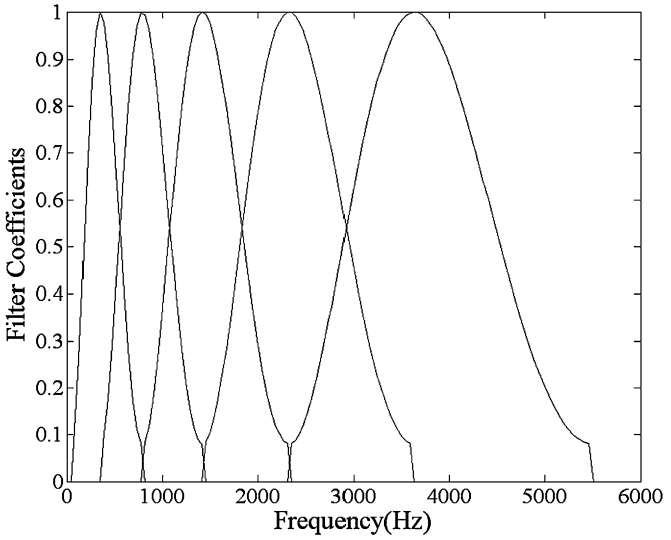


Fig. 2. Mel sub-band filters.

The algorithm to calculate the box-counting dimension is given as follows:

- Suppose $Y_{\min}$ is the minimum of current frame data Y. Move $Y_{\min}$ to zero.

$$Y_{shift} = Y - Y_{\min} \tag{9}$$

- Interpolate Y_{shift} into a fixed length *cellmax*.

$$Y_{interp} = interpl\big((1:L), Y_{shift}, (1:cellmax)\big) \tag{10}$$

where L is the length of Y, *cellmax* equals to 1024 which is lager than L, function *interpl*() uses *Linear interpolation*.

- Normalize Y_{interp} using Eq. (11):

$$YY = \frac{Y_{interp} \times cellmax}{\max(Y_{interp})} \tag{11}$$

- Now we cover YY with a set of nets whose meshes have side lengths as:

$$s = 1, 2, 4, 8, \ldots, cellmax \tag{12}$$

- Suppose $N(s)$ is the number of meshes that covered any part of the waveform, the box-counting dimension is calculated as Eq. (13).

$$D_b = polyfit\big(\log_2(s), \log_2\big(N(s)\big), 1\big) \tag{13}$$

where function *polyfit* is used to find the coefficients of a polynomial $p(x) = D_b \cdot x + C$ of degree 1 that fits the data, $\log_2(s)$ to $\log_2(N(s))$, in a least squares sense.

3.5. Mel-frequency bands energy

Mel frequency is consistent with the human ear perception of sound frequency characteristics. Mel-frequency bands energy is obtained as follows:

Establish Mel sub-band filters as shown in Fig. 2.

Mel-frequency bands energy is defined as:

$$E_{mk} = \sum_{n=1}^{N} \big[Y_{mk}(n)^2\big] \tag{14}$$

where $Y_{mk}(n)$ is the output of the kth Mel sub-band filter.

4. Feature dimension reduction

Dimension reduction is the process of reducing the number of variables under consideration, and can be divided into feature selection and feature extraction.

4.1. Fisher discriminator

Linear discriminant analysis (LDA) [21] selects those vectors in the underlying space that best discriminate among classes (rather than those that best describe the data). LDA has been widely used in pattern recognition [22,23]. Within-class scatter matrix is defined as Eq. (15):

$$S_w = \frac{1}{n} \sum_{j=1}^{c} \sum_{x_{ij} \in X_j} (x_{ij} - m_j)(x_{ij} - m_j)^T \tag{15}$$

where x_{ij} is the ith sample of class j, m_j is the mean of class j, and c is the number of classes.

Define between-class scatter matrix as Eq. (16):

$$S_b = \frac{1}{n} \sum_{j=1}^{c} n_j (m_j - m)(m_j - m)^T \tag{16}$$

where m represents the mean of all classes.

We can obtain Fisher rate as Eq. (17):

$$F_r = diag(S_b ./ S_w) \tag{17}$$

where ./ divides each entry of S_b by the corresponding entry of S_w, and function $diag(A)$ obtain the main diagonal elements of A.

4.2. Principal component analysis

Principal component analysis (PCA) is used to find a subspace whose basis vectors correspond to the maximum-variance directions in the original space.

Calculate the covariance matrix of all the feature vectors, as shown in Eq. (18):

$$\mathbf{S_t} = \frac{1}{n} \sum_{x_{ij} \in \mathbf{X}} (\mathbf{x}_{ij} - \mathbf{m})(\mathbf{x}_{ij} - \mathbf{m})^T \tag{18}$$

Computes the singular value decomposition matrix $\mathbf{S_t}$ as (19):

$$\mathbf{S_t} = \mathbf{U} \times \mathbf{S} \times \mathbf{V}^T \tag{19}$$

where $\mathbf{U}$, $\mathbf{V}$ are unitary matrixes, $\mathbf{S}$ is a diagonal matrix. The main diagonal elements of $\mathbf{S}$ are singular values of $\mathbf{S_t}$, as shown in Eq. (20):

$$\mathbf{F}_p = diag(\mathbf{S}) \tag{20}$$

where $\mathbf{V}$ is the transformation matrix of PCA, and the several big elements of $\mathbf{F}_p$ is called the Principal Component of $\mathbf{S_t}$.

5. Three-level model of speech emotion recognition

Based on Ekman's six kinds of basic emotion model and Fox's multi-level emotional classification model, we establish a three-level emotion recognition model which contains five classifiers for pairwise emotion classification. The structure of the multi-level emotion recognition model is obtained as follows:

- Since there is no emotional information in neutral speech, the basic level contains the six basic emotions only.

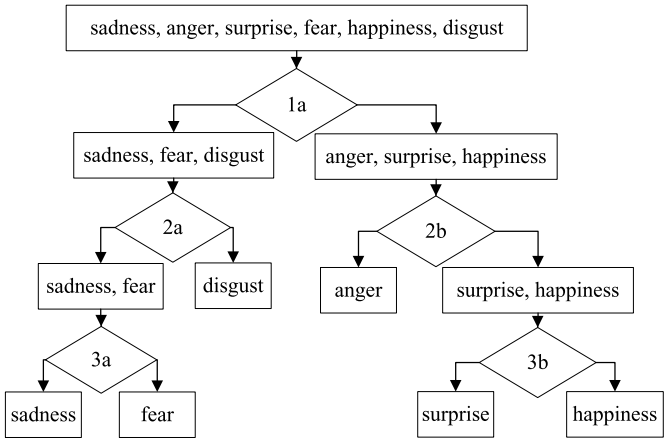


Fig. 3. Structure of three-level model.

Table 1
Fisher rates for 1a classifier.

Index	Fisher rate	Statistic	Derivative	Feature
52	**1.07242**	min.	0	C_d
1	0.96107	max.	0	E
161	0.90994	std.	1	E
177	0.85487	std.	2	E
145	0.83332	std.	0	E
81	0.79404	min.	2	E
17	0.74534	max.	1	E
65	0.74342	min.	1	E
97	0.73744	mea.	0	E
33	0.60945	max.	2	E

Table 2
Fisher rates for 2a classifier.

Index	Fisher rate	Statistic	Derivative	Feature
100	**0.47202**	mea.	0	C_d
52	0.27558	min.	0	C_d
4	0.22875	max.	0	C_d
97	0.22663	mea.	0	E
99	0.22488	mea.	0	E_z
147	0.21139	std.	0	E_z
163	0.21021	std.	1	E_z
3	0.19806	max.	0	E_z
179	0.19530	std.	2	E_z
1	0.17808	max.	0	E

- Suppose there are N kinds of grouping styles to divide the six emotions into two categories:

$$N = C_6^1 + C_6^2 + C_6^3/2 = 31 \tag{21}$$

- Based on BHUDES, calculate Fisher rates of all candidate features for each grouping styles. Suppose F_i $(i = 1, 2, \ldots, 31)$ is the mean of the four largest Fisher rates for the ith grouping styles. The final grouping style for the first level is the one with the biggest F_i.
- Make similar selections for other levels, then the final structure of multi-level emotion recognition model is obtained and shown in Fig. 3.

The model have three levels, the first level contains one classifier which divides utterance into two classes. The next two levels contain two classifiers respectively, which further refine utterance. The Fisher rates of candidate features for each classifier are shown in Tables 1–5.

Data from Tables 1–5 show that, with the classification of layers increasing, Fisher rates become smaller. For example, in 1a classifier, Fisher rate reached the maximum value of 1.07, while the maximum Fisher rate in 3b classifier is less than 0.1. It reveals that the more sophisticated classifier leads to more difficulties for classification. Surprise and joy (3b classifier) are the most difficult to be classified, followed by sadness and fear (3a classification).

Table 3
Fisher rates for 2b classifier.

Index	Fisher rate	Statistic	Derivative	Feature
163	**0.31239**	std.	1	E_z
179	0.29220	std.	2	E_z
161	0.27810	std.	1	E
191	0.26038	std.	2	E_{M4}
67	0.25870	min.	1	E_z
175	0.25285	std.	1	E_{M4}
177	0.25064	std.	2	E
188	0.24621	std.	2	E_{M1}
3	0.24212	max.	0	E_z
47	0.24103	max.	2	E_{M4}

Table 4
Fisher rates for 3a classifier.

Index	Fisher rate	Statistic	Derivative	Feature
81	**0.14235**	min.	2	E
177	0.13919	std.	2	E
65	0.13735	min.	1	E
161	0.13665	std.	1	E
163	0.10373	std.	1	E_z
22	0.10213	max.	1	P
100	0.10115	mea.	0	C_d
99	0.09844	mea.	0	E_z
1	0.09688	max.	0	E
179	0.09324	std.	2	E_z

Table 5
Fisher rates for 3b classifier.

Index	Fisher rate	Statistic	Derivative	Feature
168	**0.08905**	std.	1	S_f
184	0.04373	std.	2	S_f
54	0.04125	min.	0	P
251	0.04092	**kur.**	0	F_3
280	0.04032	**kur.**	2	S_f
264	0.03771	**kur.**	1	S_f
102	0.03736	mea.	0	P
194	0.03532	**ske.**	0	E_z
70	0.03189	min.	1	P
214	0.03046	**ske.**	1	P

In addition, with the classification of layers increasing, we need to use higher order statistics features. For example, in 1a classifier, the maximum, minimum, average and other first order statistics are mostly needed. In 2a and 2b classification, second order statistics such as standard deviation are needed. In 3a and 3b classification, the third and fourth order statistics (skewness and peak) have significant discriminative influence. In summary, the more precise description of emotion, the more fine structure of time-varying characteristics are needed.

In order to obtain the best recognition rate, the number of features used in each classifier is specified by experiments. The values are different for each classifier while they are all less than 10 dimensions.

6. Classifier selection

In the area of emotion classification, many pattern recognition algorithms have been used. Typical examples include: K-nearest neighbors (KNN), hidden Markov model (HMM), Gaussian mixtures model (GMM), support vector machine (SVM) and artificial neural net (ANN). SVM is a novel type of learning machine, which

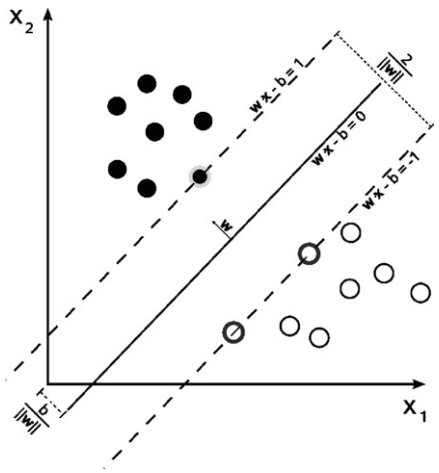


Fig. 4. Maximum-margin hyperplane.

bases on statistical learning theory (SLT). That is, an SVM is an approximate implementation of the method of structural risk minimization. SVM has shown to provide a better generalization performance in solving various classification problems than traditional techniques. Aimed at speaker independent emotion recognition from Mandarin, this paper advises multi-level emotion recognition system based on SVM. In addition, a modified Fisher discriminator is also adopted to choose proper features for each SVM in the proposed multi-level system.

6.1. Support vector machine

There are many hyperplanes in linear classifier. One reasonable choice as the best hyperplane is the one that represents the largest margin between the two classes. Therefore, we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier defined is known as a maximum margin classifier.

We are given a set of n points of the form:

$$D = \left\{ (\mathbf{x}_i, c_i) \,\middle|\, \mathbf{x}_i \in R^L,\ c_i \in \{-1, 1\} \right\}_{i=1}^{n} \tag{22}$$

where c_i indicates the class to which the point $\mathbf{x}_i$ belongs.

Any hyperplane can be written as:

$$\mathbf{w} \cdot \mathbf{x} - b = 0 \tag{23}$$

where $\cdot$ denotes the dot product. The vector $\mathbf{w}$ is a normal vector.

The parameter $b/\|\mathbf{w}\|$ determines the offset of the hyperplane from the origin along the normal vector $\mathbf{w}$.

When the training set is not linearly separable, we can select the two hyperplanes of the margin in a way that there are no points between them and then try to maximize their distance. Geometrically, we find the distance between these two hyperplanes is $2/\|\mathbf{w}\|$, so we want to minimize $\|\mathbf{w}\|$. As we also have to prevent data points falling out of the margin, the following constraint is added:

$$c_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geqslant 1 \tag{24}$$

Fig. 4 describes the maximum-margin hyperplane in two-dimensional space. In practical applications, $\|\mathbf{w}\|^2/2$ is being minimized. For mathematical simplicity, the factor is equal to $1/2$. Hence, this is a quadratic programming (QP) optimization problem.

When training dates are not linearly separable, the Soft Margin method will choose a hyperplane that splits the examples as clearly as possible, while continue to maximize the distance to the nearest clearly split examples [24]. The method introduces slack variable ξ.

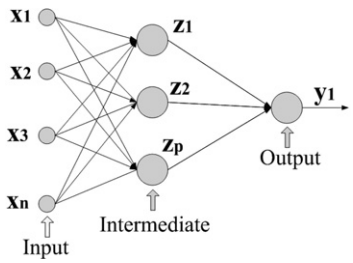


Fig. 5. MLP ANN topology.

The value of slack variable is dedicated to specific applications, the smaller the value, the more accurate the training, it will project the worse classifier, on the contrary, if the value is close to 1, training is more blurred and the classifier has better generalization. In the three-level model of emotion recognition of speech used in each classifier depends on the input vector of slack variable and feature selection criteria.

When Fisher Discriminant method is used, slack variable is:

$$\xi_F = \exp\left(-K_F \times \frac{1}{N_f} \sum_{i=1}^{N_f} \mathbf{F}_r(i) \right) \tag{25}$$

where $\mathbf{F}_r$ is the Fisher rate of each input feature, N_f is the dimension of the input features and K_F is an experimental factor which is equal to 15.

When PCA method is used, slack variable is:

$$\xi_P = \exp\left(-K_p \times \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{F}_p(i) \right) \tag{26}$$

where $\mathbf{F}_p$ is the Fisher rate of each input feature, N_p is the dimension of the input features and K_p is an experimental factor which is equal to 8.

6.2. Artificial neural network

The ANN employed in this paper has a multi-layer perceptron topology, as shown in Fig. 5.

The output node function is a Sigmoid function:

$$S_{ig}(x) = \frac{2}{1 + e^{-2x}} - 1 \tag{27}$$

The number of hidden neuron and training epochs are obtained empirically. The weight and bias values are updated according to Levenberg–Marquardt optimization [25]. In order to make strict comparison with the SVM, the ANN classifiers are defined for pairwise classification, so there is only one output node. The number of input nodes are equal to the dimension of input features. For each classifier (1a, 2a, 2b, 3a, 3b), the input features of ANN are identical to the ones of SVM.

7. Experimental results

Four comparative experiments are designed, including Fisher + SVM, PCA + SVM, Fisher + ANN, PCA + ANN. Two women and two men utterances are selected from BHUDES as training data, while one woman and two men utterances are selected as test data. Tables 6–9 show the classification results of the four experiments. Fig. 6 shows the hierarchical recognition rates for each experiment.

From Tables 6–9 and Fig. 6, we come up with two conclusions.

Table 6
Fisher + SVM recognition rate.

Emo.	Sad.	Ang.	Sur.	Fea.	Hap.	Dis.	Rate
Sad.	**96**	0	0	53	0	31	0.533
Ang.	0	**130**	20	6	23	1	**0.722**
Sur.	4	13	**78**	15	40	30	**0.433**
Fea.	48	0	5	**67**	8	52	0.372
Hap.	1	7	56	3	**95**	18	**0.528**
Dis.	21	3	13	28	39	**76**	0.422

Table 7
PCA + SVM recognition rate.

Emo.	Sad.	Ang.	Sur.	Fea.	Hap.	Dis.	Rate
Sad.	**95**	0	7	40	2	36	0.528
Ang.	1	**102**	40	3	16	18	**0.567**
Sur.	4	3	**63**	6	52	52	**0.350**
Fea.	33	0	30	**58**	7	52	0.322
Hap.	2	8	64	4	**71**	31	**0.394**
Dis.	11	2	30	25	35	**77**	0.428

Table 8
Fisher + ANN recognition rate.

Emo.	Sad.	Ang.	Sur.	Fea.	Hap.	Dis.	Rate
Sad.	**124**	0	8	65	4	39	0.517
Ang.	0	**218**	9	2	0	11	**0.908**
Sur.	1	180	**32**	5	12	10	**0.133**
Fea.	43	13	29	**106**	11	38	0.442
Hap.	2	168	21	2	**23**	24	**0.096**
Dis.	23	55	33	33	12	**84**	0.350

Table 9
PCA + ANN recognition rate.

Emo.	Sad.	Ang.	Sur.	Fea.	Hap.	Dis.	Rate
Sad.	**134**	1	8	64	8	25	0.558
Ang.	0	**222**	2	7	7	2	**0.925**
Sur.	0	201	**17**	5	7	10	**0.071**
Fea.	43	12	22	**101**	15	47	0.421
Hap.	3	183	17	6	**19**	12	**0.079**
Dis.	15	56	36	20	42	**71**	0.296

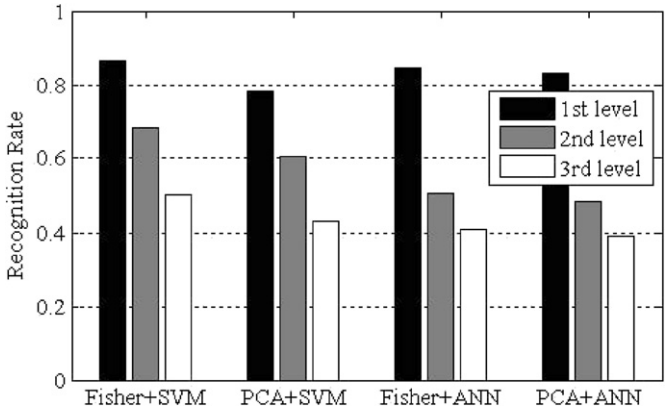


Fig. 6. Hierarchical recognition rates.

Firstly, the optimum combination is Fisher + SVM, whose recognition rate for each level is better than any other combinations. Comparing data from Table 6 and Table 7, Table 8 and Table 9, we can reveal that feature reduction using Fisher criterion is better than PCA. The reason for this improvement can be attributed to the different purposes of the two methods. Fisher criterion is aiming at finding those vectors in the underlying space that are best discriminating among classes rather than those that best describing data. Whereas, PCA is aiming at finding a subspace whose basis vectors correspond to the maximum-variance directions in the original space. In this application, we want to classify emotions rather than describe them. Comparing data from Table 6 and Table 8, Table 7 and Table 9, we can reveal that SVM which is superior to ANN in speaker independent emotion recognition. The reason is due to the high generalization performance of SVM.

Secondly, recognition rate of various emotions are very different. **Angry** is the best emotion to be recognized but **Surprise** and **happy** are the hardest emotions to be recognized. For experiments by using ANN, recognition rates of **surprise** and **happy** are even lower than the random level (1/6). For the SVM conditions, recognition rates of **surprise** and **happy** are about twice higher than the random level.

8. Conclusion

Aimed at improving speaker-independent speech emotion recognition performance, this article tries to explore the intrinsic link between speech features and emotions using Fisher criterion. The establishment of multi-level SVM-based model also helps to improve speech emotion recognition performance.

The experiment results reveal that recognition rate of some emotions, including fear and disgust, still needs to be further improved. Further research should focus on the following aspects: first, the combination of special emotional information should be paid attention to, such as the fundamental frequency rise in the end of surprising sentence, the shaking sound of fear, etc. Second, use fuzzy theory to find the probability of some kind of emotions.

Acknowledgments

This work is supported by the National Research Foundation for the Doctoral Program of Higher Education of China (No. 20070006057), National Nature Science Foundation of China (No. 61103097) and the Fundamental Research Funds for the Central Universities (No. 501LZGF2012102016).

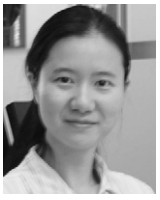
References

[1] R. Picard, Affective Computing, MIT Press, Cambridge, 1997.
[2] J. Tao, T. Tan, Affective computing: A review, in: Affective Computing and Intelligent Interaction, 2005, pp. 981–995.
[3] J. Tao, T. Tan, Affective Information Processing, Springer-Verlag, New York, 2008.
[4] R. Porzel, Contextual Computing: Models and Applications, Springer-Verlag, 2010.
[5] D. Ververidis, C. Kotropoulos, Emotional speech recognition – resources features and methods, Speech Commun. 48 (2006) 1162–1181.
[6] K. Scherer, Vocal communication of emotion: A review of research paradigms, Speech Commun. 40 (1–2) (2003) 227–256.
[7] T. Sobol-Shikler, P. Robinson, Classification of complex information: Inference of co-occurring affective states from their expressions in speech, IEEE Trans. Pattern Anal. Mach. Intell. 32 (7) (2010) 1284–1297.
[8] C. Peter, Affect and Emotion in Human–Computer Interaction: From Theory to Applications, vol. 4868, Springer-Verlag, New York, 2008.
[9] W. Yoon, K. Park, A study of emotion recognition and its applications, in: Modeling Decisions for Artificial Intelligence, vol. 6417, 2007, pp. 455–462.
[10] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, H. Konosu, Being bored? Recognising natural interest by extensive audiovisual integration for real-life application, Image Vis. Comput. 27 (12) (2009) 1760–1774.
[11] K. Van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, S. Baumann, Fully generated scripted dialogue for embodied agents, Artificial Intelligence 172 (10) (2008) 1219–1244.
[12] E. Lorini, F. Schwarzentruber, A logic for reasoning about counterfactual emotions, Artificial Intelligence 175 (3) (2011) 814–847.
[13] B. Schuller, S. Reiter, G. Rigoll, Evolutionary feature generation in speech emotion recognition, in: IEEE International Conference on Multimedia and Expo ICME 2006, IEEE, 2006, pp. 5–8.
[14] T. Bänziger, K. Scherer, The role of intonation in emotional expressions, Speech Commun. 46 (3–4) (2005) 252–267.

[15] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), vol. 1, IEEE, 2004, pp. 577–580.
[16] X. Mao, L. Chen, L. Fu, Multi-level speech emotion recognition based on HMM and ANN, in: Proc. 2009 WRI World Congress on Computer Science and Information Engineering, vol. 7, IEEE, 2009, pp. 225–229.
[17] X. Mao, L. Chen, Speech emotion recognition based on parametric filter and fractal dimension, IEICE Trans. Inf. Syst. 93 (8) (2010) 2324–2326.
[18] P. Ekman, Are there basic emotions, Psychol. Rev. 99 (3) (1992) 550–553.
[19] N. Fox, If it's not left, it's right: Electroencephalograph asymmetry and the development of emotion, Am. Psychol. 46 (8) (1991) 863–872.
[20] X. Mao, L. Chen, L. Fu, Mandarin speech emotion recognition based on a hybrid of HMM/ANN, Int. J. Comput. 1 (4) (2007) 321–324.
[21] R. Fisher, et al., The statistical utilization of multiple measurements, Ann. Eugenics 8 (1938) 376–386.
[22] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (2002) 711–720.
[23] Y. Sun, Y. Zhou, Q. Zhao, Y. Yan, Acoustic feature optimization based on F-ratio for robust speech recognition, IEICE Trans. Inf. Syst. 93 (9) (2010) 2417–2430.
[24] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[25] J. More, The Levenberg–Marquardt algorithm: Implementation and theory, Numer. Anal. (1978) 105–116.

Lijiang Chen received his B.Sc. degree in Electronic and Information Engineering, Beihang University, Beijing, China, in 2007. He is currently pursuing the Ph.D. degree in Electronic and Information Engineering, Beihang University, Beijing, China. His main research interests include speech signal processing, pattern recognition and speech emotion recognition.

Xia Mao was born in Yiwu Zhejiang province China in 1952. She received her M.S. degree and Ph.D. degree from Saga University, Japan in 1993 and 1996 respectively. She is currently a Professor at School of Electronic and Information Engineering, Beihang University, Beijing, China. Her current research interests include affective computing, artificial intelligence, pattern recognition and Human–Computer Interaction. So far, she has published over 140 pieces of papers both domestically and overseas, many of them have been cited by the SCI, EI, ISTP, etc. Dr. Mao is leading several projects supported by the National High-tech Research and Development Program (863 Program), National Natural Science Foundation and Beijing Natural Science Foundation.



Yuli Xue received his B.Sc. degree and Ph.D. degree in Electronic and Information Engineering, Beihang University, Beijing, China, in 2003 and 2009 respectively. Her main research interests include human–computer interaction, pattern recognition and emotion recognition.

Lee Lung Cheng received his B.Sc. and M.Sc. from King's College, University of London in 1976 and 1981, respectively. He received his Doctor of Engineering from the Tsinghua University of China. His recent researches are mainly on security systems, coding, cryptography, GPS, smart cards and computing systems. Dr. Cheng published over 30 papers in international conferences and journals. He has 4 Patents awarded. He also received one Excellent Paper Award in the 2nd Youth Scientist Conference 1995, Beijing, China.