1

MACHINE LEARNING METHODS IN THE APPLICATION OF SPEECH EMOTION RECOGNITION

Ling Cen[1], Minghui Dong[1], Haizhou Li[1]
Zhu Liang Yu[2] and Paul Chan[1]
[1]*Institute for Infocomm Research*
Singapore
[2]*College of Automation Science and Engineering,*
South China University of Technology,
Guangzhou, China

1. Introduction

Machine Learning concerns the development of algorithms, which allows machine to learn via inductive inference based on observation data that represent incomplete information about statistical phenomenon. Classification, also referred to as pattern recognition, is an important task in Machine Learning, by which machines "learn" to automatically recognize complex patterns, to distinguish between exemplars based on their different patterns, and to make intelligent decisions. A pattern classification task generally consists of three modules, i.e. data representation (feature extraction) module, feature selection or reduction module, and classification module. The first module aims to find invariant features that are able to best describe the differences in classes. The second module of feature selection and feature reduction is to reduce the dimensionality of the feature vectors for classification. The classification module finds the actual mapping between patterns and labels based on features. The objective of this chapter is to investigate the machine learning methods in the application of automatic recognition of emotional states from human speech.

It is well-known that human speech not only conveys linguistic information but also the paralinguistic information referring to the implicit messages such as emotional states of the speaker. Human emotions are the mental and physiological states associated with the feelings, thoughts, and behaviors of humans. The emotional states conveyed in speech play an important role in human-human communication as they provide important information about the speakers or their responses to the outside world. Sometimes, the same sentences expressed in different emotions have different meanings. It is, thus, clearly important for a computer to be capable of identifying the emotional state expressed by a human subject in order for personalized responses to be delivered accordingly.

Speech emotion recognition aims to automatically identify the emotional or physical state of a human being from his or her voice. With the rapid development of human-computer interaction technology, it has found increasing applications in security, learning, medicine, entertainment, etc. Abnormal emotion (e.g. stress and nervousness) detection in audio surveillance can help detect a lie or identify a suspicious person. Web-based E-learning has prompted more interactive functions between computers and human users. With the ability to recognize emotions from users' speech, computers can interactively adjust the content of teaching and speed of delivery depending on the users' response. The same idea can be used in commercial applications, where machines are able to recognize emotions expressed by the customers and adjust their responses accordingly. The automatic recognition of emotions in speech can also be useful in clinical studies, psychosis monitoring and diagnosis. Entertainment is another possible application for emotion recognition. With the help of emotion detection, interactive games can be made more natural and interesting. Motivated by the demand for human-like machines and the increasing applications, research on speech based emotion recognition has been investigated for over two decades (Amir, 2001; Clavel et al., 2004; Cowie & Douglas-Cowie, 1996; Cowie et al., 2001; Dellaert et al., 1996; Lee & Narayanan, 2005; Morrison et al., 2007; Nguyen & Bass, 2005; Nicholson et al., 1999; Petrushin, 1999; Petrushin, 2000; Scherer, 2000; Ser et al., 2008; Ververidis & Kotropoulos, 2006; Yu et al., 2001; Zhou et al., 2006).

Speech feature extraction is of critical importance in speech emotion recognition. The basic acoustic features extracted directly from the original speech signals, e.g. pitch, energy, rate of speech, are widely used in speech emotion recognition (Ververidis & Kotropoulos, 2006; Lee & Narayanan, 2005; Dellaert et al., 1996; Petrushin, 2000; Amir, 2001). The pitch of speech is the main acoustic correlate of tone and intonation. It depends on the number of vibrations per second produced by the vocal cords, and represents the highness or lowness of a tone as perceived by the ear. Since the pitch is related to the tension of the vocal folds and subglottal air pressure, it can provide information about the emotions expressed in speech (Ververidis & Kotropoulos, 2006). In the study on the behavior of the acoustic features in different emotions (Davitz, 1964; Huttar, 1968; Fonagy, 1978; Moravek, 1979; Van Bezooijen, 1984; McGilloway et al., 1995, Ververidis & Kotropoulos, 2006), it has been found that the pitch level in anger and fear is higher while a lower mean pitch level is measured in disgust and sadness. A downward slope in the pitch contour can be observed in speech expressed with fear and sadness, while the speech with joy shows a rising slope. The energy related features are also commonly used in emotion recognition. Higher energy is measured with anger and fear. Disgust and sadness are associated with a lower intensity level. The rate of speech also varies with different emotions and aids in the identification of a person's emotional state (Ververidis & Kotropoulos, 2006; Lee & Narayanan, 2005). Some features derived from mathematical transformation of basic acoustic features, e.g. Mel-Frequency Cepstral Coefficients (MFCC) (Specht, 1988; Reynolds et al., 2000) and Linear Prediction-based Cepstral Coefficients (LPCC) (Specht, 1988), are also employed in some studies. As speech is assumed as a short-time stationary signal, acoustic features are generally calculated on a frame basis, in order to capture long range characteristics of the speech signal, feature statistics are usually used, such as mean, median, range, standard deviation, maximum, minimum, and linear regression coefficient (Lee & Narayanan, 2005). Even though many studies have been carried out to find which acoustic features are suitable for

emotion recognition, however, there is still no conclusive evidence to show which set of features can provide the best recognition accuracy (Zhou, 2006).

Most machine learning and data mining techniques may not work effectively with high-dimensional feature vectors and limited data. Feature selection or feature reduction is usually conducted to reduce the dimensionality of the feature space. To work with a small, well-selected feature set, irrelevant information in the original feature set can be removed. The complexity of calculation is also reduced with a decreased dimensionality. Lee & Narayanan (2005) used the forward selection (FS) method for feature selection. FS first initialized to contain the single best feature with respect to a chosen criterion from the whole feature set, in which the classification accuracy criterion by *nearest neighborhood* rule is used and the accuracy rate is estimated by *leave-one-out* method. The subsequent features were then added from the remaining features which maximized the classification accuracy until the number of features added reached a pre-specified number. Principal Component Analysis (PCA) was applied to further reduce the dimension of the features selected using the FS method. An automatic feature selector based on a RF2TREE algorithm and the traditional C4.5 algorithm was developed by Rong et al. (2007). The ensemble learning method was applied to enlarge the original data set by building a bagged random forest to generate many virtual examples. After which, the new data set was used to train a single decision tree, which selected the most efficient features to represent the speech signals for emotion recognition. The genetic algorithm was applied to select an optimal feature set for emotion recognition (Oudeyer, 2003).

After the acoustic features are extracted and processed, they are sent to emotion classification module. Dellaert et al. (1996) used K-nearest neighbor (*k*-NN) classifier and majority voting of subspace specialists for the recognition of sadness, anger, happiness and fear and the maximum accuracy achieved was 79.5%. Neural network (NN) was employed to recognize eight emotions, i.e. happiness, teasing, fear, sadness, disgust, anger, surprise and neutral and an accuracy of 50% was achieved (Nicholson et al. 1999). The linear discrimination, *k*-NN classifiers, and SVM were used to distinguish negative and non-negative emotions and a maximum accuracy of 75% was achieved (Lee & Narayanan, 2005). Petrushin (1999) developed a real-time emotion recognizer using Neural Networks for call center applications, and achieved 77% classification accuracy in recognizing agitation and calm emotions using eight features chosen by a feature selection algorithm. Yu et al. (2001) used SVMs to detect anger, happiness, sadness, and neutral with an average accuracy of 73%. Scherer (2000) explored the existence of a universal psychobiological mechanism of emotions in speech by studying the recognition of fear, joy, sadness, anger and disgust in nine languages, obtaining 66% of overall accuracy. Two hybrid classification schemes, stacked generalization and the un-weighted vote, were proposed and accuracies of 72.18% and 70.54% were achieved respectively, when they were used to recognize anger, disgust, fear, happiness, sadness and surprise (Morrison, 2007). Hybrid classification methods that combined the Support Vector Machines and the Decision Tree were proposed (Nguyen & Bass, 2005). The best accuracies for classifying neutral, anger, lombard and loud was 72.4%.

In this chapter, we will discuss the application of machine learning methods in speech emotion recognition, where feature extraction, feature reduction and classification will be covered. The comparison results in speech emotion recognition using several popular classification methods have been given (Cen et al. 2009). In this chapter, we focus on feature processing, where the related experiment results in the classification of 15 emotional states

for the samples extracted from the LDC database are presented. The remaining part of this chapter is organized as follows. The acoustic feature extraction process and methods are detailed in Section 2, where the feature normalization, utterance segmentation and feature dimensionality reduction are covered. In the following section, the Support Vector Machine (SVM) for emotion classification is presented. Numerical results and performance comparison are shown in Section 4. Finally, the concluding remarks are made in Section 5.

2. Acoustic Features

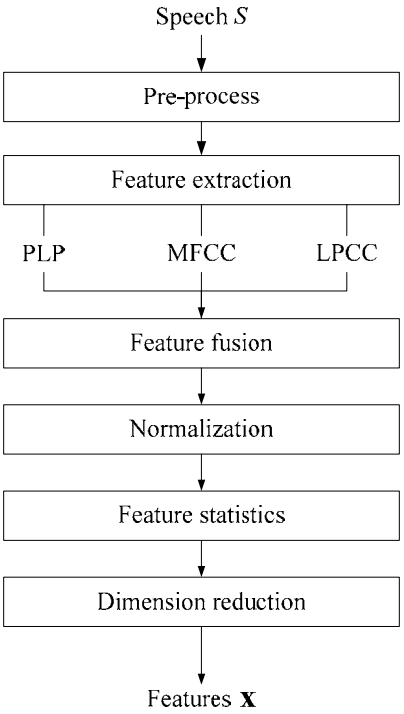


Fig. 1. Basic block diagram for feature calculation.

Speech feature extraction aims to find the acoustic correlates of emotions in human speech. Fig. 1 shows the block diagram for acoustic feature calculation, where S represents a speech sample (an utterance) and $\mathbf{x}$ denotes its acoustic features. Before the raw features are extracted, the speech signal is first pre-processed by pre-emphasis, framing and windowing processes. In our work, three short time cepstral features are extracted, which are Linear Prediction-based Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP) Cepstral Coefficients, and Mel-Frequency Cepstral Coefficients (MFCC). These features are fused to achieve a feature matrix, $\mathbf{x} \in R^{F \times M}$ for each sentence S, where F is the number of frames in the utterance, and M is the number of features extracted from each frame. Feature normalization is carried out on the speaker level and the sentence level. As the features are

extracted on a frame basis, the statistics of the features are calculated for every window of a specified number of frames. These include the mean, median, range, standard deviation, maximum, and minimum. Finally, PCA is employed to reduce the feature dimensionality. These will be elaborated in subsections below.

2.1 Signal Pre-processing: Pre-emphasis, Framing, Windowing

In order to emphasize important frequency component in the signal, a pre-emphasis process is carried out on the speech signal using a Finite Impulse Response (FIR) filter called pre-emphasis filter, given by

$$H_{pre}(z) = 1 + a_{pre} z^{-1}. \tag{1}$$

The coefficient a_{pre} can be chosen typically from [-1.0, 0.4] (Picone, 1993). In our implementation, it is set to be $a_{pre} = -(1 - \frac{1}{16}) = -0.9375$, so that it can be efficiently implemented in fixed point hardware.

The filtered speech signal is then divided into frames. It is based on the assumption that the signal within a frame is stationary or quasi-stationary. Frame shift is the time difference between the start points of successive frames, and the frame length is the time duration of each frame. We extract the signal frames of length 25 msec from the filtered signal at every interval of 10 msec. A Hamming window is then applied to each signal frame to reduce signal discontinuity in order to avoid spectral leakage.

2.2 Feature Extraction

Three short time cepstral features, i.e. Linear Prediction-based Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP) Cepstral Coefficients, and Mel-Frequency Cepstral Coefficients (MFCC), are extracted as acoustic features for speech emotion recognition.

A. LPCC

Linear Prediction (LP) analysis is one of the most important speech analysis technologies. It is based on the source-filter model, where the vocal tract transfer function is modeled by an all-pole filter with a transfer function given by

$$H(z) = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}, \tag{2}$$

where a_i is the filter coefficients. The speech signal, S_t assumed to be stationary over the analysis frame is approximated as a linear combination of the past *p* samples, given as

$$\hat{S}_t = \sum_{i=1}^{p} a_i S_{t-i}. \tag{3}$$

In (3) a_i can be found by minimizing the mean square filter prediction error between $\hat{S}_t$ and S_t. The cepstral coefficents is considered to be more reliable and robust than the LP filter coefficents. It can be computed directly from the LP filter coefficients using the recursion given as

$$\hat{c}_k = a_k + \sum_{i=1}^{k-1}\left(\frac{i}{k}\right)c_i a_{k-i}, \qquad 0<k \le p, \tag{4}$$

where c_k represents the cepstral coefficients.

B. PLP Cepstral Coefficients

PLP is first proposed by Hermansky (1990), which combines the Discrete Fourier Transform (DFT) and LP technique. In PLP analysis, the speech signal is processed based on hearing perceptual properties before LP analysis is carried out, in which the spectrum is analyzed on a warped frequency scale. The calculation of PLP cepstral coefficients involves 6 steps as shown in Fig. 2.

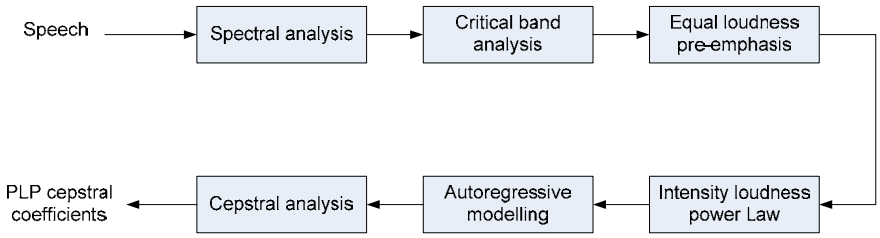


Fig. 2. Calculation of PLP cepstral coefficients.

Step 1 Spectral analysis
- The short-time power spectrum is achieved for each speech frame.

Step 2 Critical-band Spectral resolution
- The power spectrum is warped onto a Bark scale and convolved with the power spectral of the critical band filter, in order to simulate the frequency resolution of the ear which is approximately constant on the Bark scale.

Step 3 Equal-loudness pre-emphasis
- An equal-loudness curve is used to compensate for the non-equal perception of loudness at different frequencies.

Step 4 Intensity loudness power law
- Perceived loudness is approximately the cube root of the intensity.

Step 5 Autoregressive modeling
- Inverse Discrete Fourier Transform (IDFT) is carried out to obtain the autoregressive coefficients and all-pole modeling is then performed.

Step 6 Cepstral analysis
- PLP cepstral coefficients are calculated from the AR coefficients as the process in LPCC calculation.

C. MFCC

The MFCC proposed by Davis and Mermelstein (1980) has become the most popular features used in speech recognition. The calculation of MFCC involves computing the cosine transform of the real logarithm of the short-time power spectrum on a Mel warped frequency scale. The process consists of the following process as shown in Fig. 3.

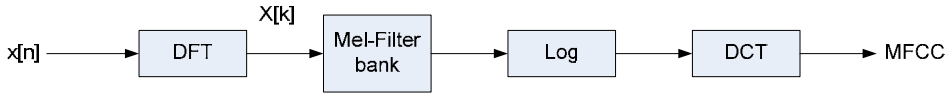


Fig. 3. Calculation of MFCC.

1) DFT is applied in each speech frame given as

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi 2\pi / N}, \quad 0 \le k \le N-1. \tag{5}$$

2) Mel-scale filter bank
The Fourier spectrum is non-uniformly quantized to conduct Mel filter bank analysis. The window functions that are first uniformly spaced on the Mel-scale and then transformed back to the Hertz-scale are multiplied with the Fourier power spectrum and accumulated to achieve the Mel spectrum filter-bank coefficients. A Mel filter bank has filters linearly spaced at low frequencies and approximately logarithmically spaced at high frequencies, which can capture the phonetically important characteristics of the speech signal while suppressing insignificant spectral variation in the higher frequency bands (Davis and Mermelstein, 1980).

3) The Mel spectrum filter-bank coefficients is calculated as

$$F[m] = \log\left(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k] \right), \quad 0 \le m \le M. \tag{6}$$

4) The Discrete Cosine Transform (DCT) of the log filter bank energies is calculated to find the MFCC given as

$$c[n] = \sum_{m=0}^{M} F[m] \cos(\pi n(m-1)/2\mathrm{M}), \quad 0 \le n \le M, \tag{7}$$

where $c[n]$ is the n^{th} coefficient.

D. Delta and Acceleration Coefficients

After the three short time cepstral features, LPCC, PLP Cepstral Coefficients, and MFCC, are extracted, they are fused to form a feature vector for each of the speech frames. In the vector, besides the LPCC, PLP cepstral coefficients and MFCC, Delta and Acceleration (Delta Delta) of the raw features are also included, given as

Delta Δx_i :

$$\Delta x_i = \frac{1}{2}(x_{i+1} - x_{i-1}), \tag{8}$$

Acceleration (Delta Delta) $\Delta\Delta x_i$:

$$\Delta\Delta x_i = \frac{1}{2}(\Delta x_{i+1} - \Delta x_{i-1}), \tag{9}$$

where x_i is the i^{th} value in the feature vector.

E. Feature List

In conclusion, the list below shows the full feature set used in speech emotion recognition presented in this chapter. The feature vector has a dimension of R^M , where M = 132 is the total number of the features calculated for each frame.

1) PLP - 54 features
- 18 PLP cepstral coefficients
- 18 Delta PLP cepstral coefficients
- 18 Delta Delta PLP cepstral coefficients.

2) MFCC - 39 features
- 12 MFCC features
- 12 delta MFCC features
- 12 Delta Delta MFCC features
- 1 (log) frame energy
- 1 Delta (log) frame energy
- 1 Delta Delta (log) frame energy

3) LPCC – 39 features
- 13 LPCC features
- 13 delta LPCC features
- 13 Delta Delta LPCC features

2.3 Feature Normalization

As acoustic variation in different speakers and different utterances can be found in phonologically identical utterances, speaker- and utterance-level normalization are usually performed to reduce these variations, and hence to increase recognition accuracy.
In our work, the normalization is achieved by subtracting the mean and dividing by the standard deviation of the features given as

$$x_i = \frac{(x_i - \mu_{ui})/\sigma_{ui} - \mu_{si}}{\sigma_{si}}, \tag{10}$$

where x_i is the i^{th} coefficient in the feature vector, μ_{ui} and σ_{ui} are the mean and standard deviation of x_i within an utterance, and μ_{si} and σ_{si} are the mean and standard deviation of x_i within the utterances spoken by the same speaker. In this way, the variation across speakers and utterances can be reduced.

2.4 Utterance Segmentation

As we have discussed, the three short time cepstral features are extracted for each speech frames. The information in the individual frames is not sufficient for capturing the longer time characteristics of the speech signal. To address the problem, we arrange the frames within an utterance into several segments as shown in Fig. 4. In this figure, f_i represents a frame and s_i denotes a segment. Each segment consists of a fixed number of frames. The *sf* represents the segment size, i.e. the number of frames in one segment, and Δ is the overlap size, i.e. the number of frames overlapped in two consecutive segments.

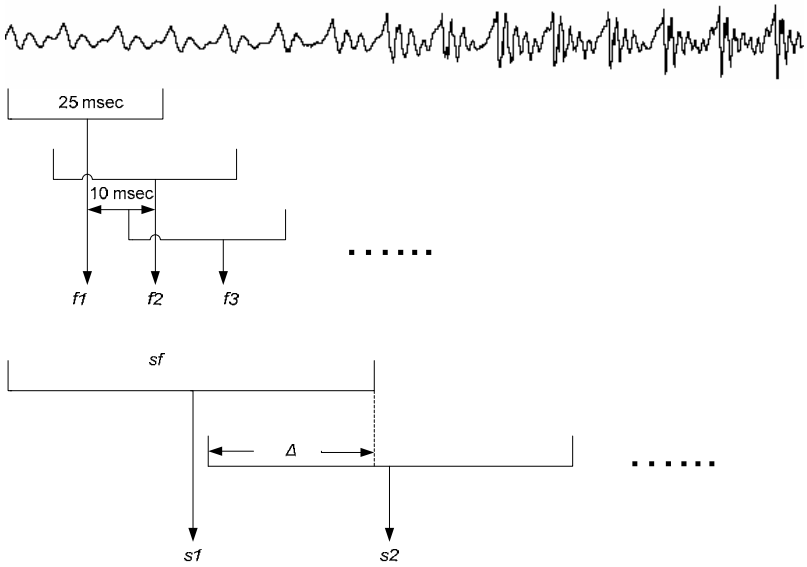


Fig. 4. Utterance partition with frames and segments.

Here, the trade-off between computational complexity and recognition accuracy is considered in utterance segmentation. Generally speaking, finer partition and larger overlap between two consecutive segments potentially result in better classification performance at the cost of higher computational complexity. The statistics of the 132 features given in the previous sub-section is calculated for each segment, which is used in emotion classification instead of the original 132 features in each frame. This includes median, mean, standard deviation, maximum, minimum, and range (max-min). In total, the number of statistic parameters in a feature vector for each speech segment is $132 \times 6 = 792$.

2.5 Feature Dimensionality Reduction

Most machine learning and data mining techniques may not work effectively if the dimensionality of the data is high. Feature selection or feature reduction is usually carried out to reduce the dimensionality of the feature vectors. A short feature set can also improve computational efficiency involved in classification and avoids the problem of overfitting. Feature reduction aims to map the original high-dimensional data onto a lower-dimensional space, in which all of the original features are used. In feature selection, however, only a subset of the original features is chosen.

In our work, Principal Component Analysis (PCA) is employed to reduce the feature dimensionality. Assume the feature matrix, $X^T \in R^{N_S \times M}$, with zero empirical mean, in which each row is a feature vector of a data sample, and N_S is the number of data samples. The PCA transformation is given as

$$Y^T = X^T W = V\Sigma, \tag{11}$$

where $V\Sigma\Sigma^T$ is the Singular Value Decomposition (SVD) of X^T. PCA mathematically transforms a number of potentially correlated variables into a smaller number of uncorrelated variables called Principal Components (PC). The first PC (the eigenvector with the largest eigenvalue) accounts for the greatest variance in the data, the second PC accounts for the second variance, and each succeeding PCs accounts for the remaining variability in order. Although PCA requires a higher computational cost compared to the other methods, for example, the Discrete Cosine Transform, it is an optimal linear transformation for keeping the subspace with the largest variance.

3. Support Vector Machines (SVMs) for Emotion Classification

SVMs that developed by Vapnik (1995) and his colleagues at AT&T Bell Labs in the mid 90's, have become of increasing interest in classification (Steinwart and Christmann, 2008). It has shown to have better generalization performance than traditional techniques in solving classification problems. In contrast to traditional techniques for pattern recognition that are based on the minimization of empirical risk learned from training datasets, it aims to minimize the structural risk to achieve optimum performance.

It is based on the concept of decision planes that separates the objects belonging to different categories. In the SVMs, the input data are separated as two sets using a separating hyperplane that maximizes the margin between the two data sets. Assuming the training data samples are in the form of

$$\{\mathbf{x}_i, c_i\},\ i = 1, ..., N,\ \mathbf{x}_i \in \mathbf{R}^M, c_i \in \{-1, 1\} \tag{12}$$

Where $\mathbf{x}_i$ is the *M*-dimension feature vector of the *i*th sample, *N* is the number of samples, and c_i is the category to which $\mathbf{x}_i$ belongs. Suppose there is a hyperplane that separates the feature vectors $\phi(\mathbf{x}_i)$ in the positive category from those in the negative one. Here $\phi(\bullet)$ is a nonlinear mapping of the input space into a higher dimensional feature space. The set of points $\phi(\mathbf{x})$ that lie on the hyperplane is expressed as

$$\mathbf{w}\cdot\phi(\mathbf{x})+b=0, \tag{13}$$

where $\mathbf{w}$ and b are the two parameters. For the training data that are linearly separable, two hyperplanes are selected to yield maximum margin. Suppose $\mathbf{x}_i$ satisfies

$$\begin{aligned}\phi(\mathbf{x}_i)\cdot\mathbf{w}+b\geq 1,\ \text{for } c_i=1,\\ \phi(\mathbf{x}_i)\cdot\mathbf{w}+b\leq 1,\ \text{for } c_i=-1.\end{aligned} \tag{14}$$

It can be re-written as

$$c_i\left(\phi(\mathbf{x}_i)\cdot\mathbf{w}+b\right)-1\geq 0,\ \forall i=1,2,...,N. \tag{15}$$

Searching a pair of hyperplanes that gives the maximum margin can be achieved by solving the following optimization problem

$$\begin{aligned}&\text{Minimize } \|\mathbf{w}\|^2\\ &\text{subject } c_i\left(\phi(\mathbf{x}_i)\cdot\mathbf{w}+b\right)\geq 1,\forall i=1,2,...,N.\end{aligned} \tag{16}$$

In (16), $\|\mathbf{w}\|$ represents the Euclidean norm of $\mathbf{w}$. This can be formulated as a quadratic programming optimization problem and be solved by standard quadratic programming techniques.

Using the Lagrangian methodology, the dual problem of (16) is given as

$$\begin{aligned}&\text{Minimize } W(\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}c_ic_j\alpha_i\alpha_j\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j),\\ &\text{subject } \sum_{i=1}^{N}c_i\alpha_i=0,\ \alpha_i\geq 0,\ \forall i=1,2,...,N.\end{aligned} \tag{17}$$

Here α_i is the Lagrangian variables.

The simplest case is that $\phi(\mathbf{x})$ is a linear function. If the data cannot be separated in a linear way, non-linear mappings are performed from the original space to a feature space via kernels. This aims to construct a linear classifier in the transformed space, which is the so-called "kernel trick". It can be seen from (17) that the training points appear as their inner products in the dual formulation. According to Mercer's theorem, any symmetric positive semi-definite function $k(\mathbf{x}_i,\mathbf{x}_j)$ implicitly defines a mapping into a feature space

$$\phi:\mathbf{x}\rightarrow\phi(\mathbf{x}) \tag{18}$$

such that the function is an inner product in the feature space given as

$$k(\mathbf{x}_i,\mathbf{x}_j)=\phi(\mathbf{x}_i)\cdot\phi(\mathbf{x}_j) \tag{19}$$

The function $k(\mathbf{x}_i, \mathbf{x}_j)$ is called kernels. The dual problem in the kernel form is then given as

$$\begin{aligned} &\text{Minimize } W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} c_i c_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j), \\ &\text{subject } \sum_{i=1}^{N} c_i \alpha_i = 0, \ \alpha_i \geq 0, \ \forall i = 1, 2, ..., N. \end{aligned} \tag{20}$$

By replacing the inner product in (17) with a kernel and solving for α, a maximal margin separating hyperplane can be obtained in the feature space defined by a kernel. Choosing suitable non-linear kernels, therefore, classifiers that are non-linear in the original space can become linear in the feature space. Some common kernel functions are shown below:

1) Polynomial (homogeneous) kernel: $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}')^d$,
2) Polynomial (inhomogeneous) kernel: $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x} \cdot \mathbf{x}' + 1)^d$,
3) Radial basis kernel: $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2\right), \text{ for } \gamma > 0,$
4) Gaussian radial basis kernel: $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right).$

A single SVM itself is a classification method for 2-category data. In speech emotion recognition, there are usually multiple emotion categories. Two common methods used to solve the problem are called one-versus-all and one-versus-one (Fradkin and Muchnik, 2006). In the former, one SVM is built for each emotion, which distinguishes this emotion from the rest. In the latter, one SVM is built to distinguish between every pair of categories. The final classification decision is made according to the results from all the SVMs with the majority rule. In the one-versus-all method, the emotion category of an utterance is determined by the classifier with the highest output based on the winner-takes-all strategy. In the one-versus-one method, every classifier assigns the utterance to one of the two emotion categories, then the vote for the assigned category is increased by one vote, and the emotion class is the one with most votes based on a max-wins voting strategy.

4. Experiments

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. 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 14 emotional states (including anxiety, boredom, cold anger, hot anger, contempt, despair, disgust, elation, happiness, interest, panic, pride, sadness, and shame) as well as neutral state.

The number of utterances is approximately 2300. The histogram distribution of these samples for the emotions, speakers, and genders are shown in Fig. 5, where Fig. 5-a shows the number of samples expressed in each of 15 emotional states; 5-b illustrates the number of samples spoken by each of 7 professional actors (1st, 2nd, and 5th speakers are male; the others are female); Fig. 5-c gives the number of samples divided into gender group (1-male; 2-female).

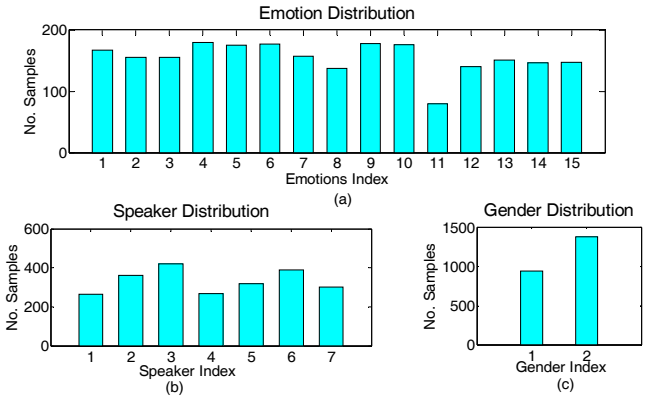


Fig. 5. Histogram distribution of the number of utterances for the emotions, speakers, and genders.

The SVM classification method introduced in Section 3 is used to recognize the emotional states expressed in the speech samples extracted from the above database. The speech data are trained in speaker dependent training mode, in which the different characteristics of speech among the speakers are considered and an individual training process is hence carried out for each speaker. The database is divided into two parts, i.e. the training dataset and the testing dataset. Half of the data are employed to train the classifiers and the remainder are used for testing purpose.

4.1 Comparisons among different segmentation forms

It is reasonable that finer partition and larger overlap size tend to improve recognition accuracy. Computational complexity, however, should be considered in practical applications. In this experiment, we test the system with different segmentation forms, i.e. different segment sizes *sf* and different overlap sizes Δ.

The segment size is first changed from 30 to 60 frames with a fixed overlap size of 20 frames. The numerical results are shown in Table 1, where the recognition accuracy in each emotion as well as the average accuracy is given. A trend of decreasing average accuracy is observed as the segment size is increased, which is illustrated in Fig. 6.

sf / Emotions	30	35	40	45	50	55	60
Anxiety	87	86	84	84	88	87	81
Boredom	82	78	82	77	74	76	79
Cold Anger	62	69	65	62	59	63	59
Contempt	72	66	72	60	63	66	58
Despair	68	68	68	53	61	55	60
Disgust	81	78	81	72	78	78	73
Elation	78	71	71	67	67	67	70
Hot Anger	79	79	76	82	79	75	69
Happiness	62	58	56	62	48	47	45
Interest	55	50	53	50	52	43	38
Neutral	92	82	82	71	69	82	72
Panic	70	65	62	61	61	61	58
Pride	28	33	28	26	28	22	24
Sadness	71	68	61	63	63	64	61
Shame	53	53	44	45	43	45	36
Average	69.33	66.93	65.67	62.33	62.20	62.07	58.87

Table 1. Recognition accuracies (%) achieved with different segment sizes (the overlap size is fixed to be 20)

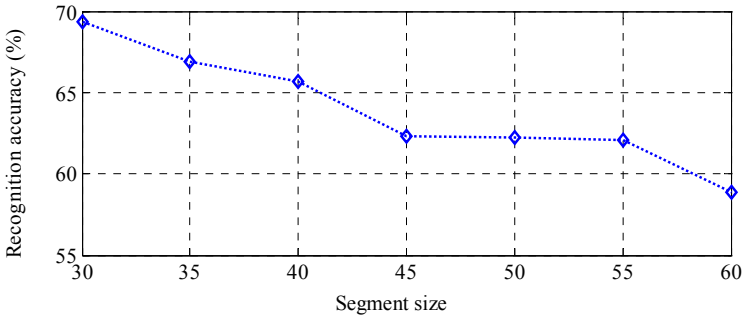


Fig. 6. Comparison of the average accuracies achieved with different segment sizes (ranging from 30 to 60) and a fixed overlap size of 20.

Secondly, the segment size is fixed to 40 and different overlap sizes ranging from 5 to 30 are used in the experiment. The recognition accuracies for all emotions are listed in Table 2. The trend of average accuracy with the increase of the overlap size is shown in Fig. 7, where we can see an increase trend when the overlap size becomes larger.

Emotions \ Δ	5	10	15	20	25	30
Anxiety	81	84	83	84	84	84
Boredom	73	73	77	82	82	79
Cold Anger	68	56	62	65	65	67
Contempt	63	64	71	72	74	76
Despair	60	59	63	68	59	69
Disgust	71	71	72	81	74	76
Elation	72	71	75	71	71	76
Hot Anger	75	76	81	76	78	81
Happiness	48	47	60	56	64	63
Interest	45	44	45	53	58	52
Neutral	69	72	77	82	87	82
Panic	61	63	63	62	63	70
Pride	28	24	29	28	32	36
Sadness	57	60	63	61	61	68
Shame	41	41	43	44	52	57
Average	60.80	60.33	64.27	65.67	66.93	69.07

Table 2. Recognition accuracies (%) achieved with different overlap sizes (the segment size is fixed to be 40)

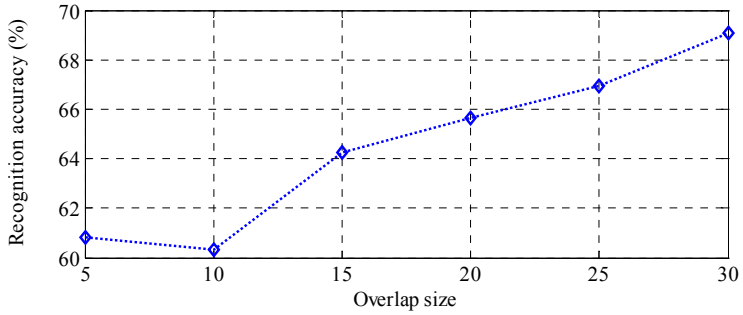


Fig. 7. Comparison of the average accuracies achieved with different overlap sizes (ranging from 5 to 30) and a fixed segment size of 40.

4.2 Comparisons among different feature sizes

This experiment aims to find the optimal dimensionality of the feature set. The segment size for calculating feature statistics is fixed with $sf = 40$ and $\Delta = 20$. The full feature set for each segment is a 792-dimensional vector as discussed in Section 2. The PCA is adopted to reduce feature dimensionality. The recognition accuracies achieved with different dimensionalities ranging from 300 to 20, as well as the full feature set with 792 features, are shown in Table 3. The average accuracies are illustrated in Fig. 8.

Feature size / Emotions	Full	300	250	200	150	100	50	20
Anxiety	84	86	88	86	86	81	71	53
Boredom	82	83	78	77	78	76	60	41
Cold Anger	65	68	64	62	63	64	53	32
Contempt	72	71	71	70	66	56	35	29
Despair	68	64	64	64	57	61	44	33
Disgust	81	80	79	75	79	66	60	48
Elation	71	72	72	76	75	70	49	41
Hot Anger	76	78	78	75	78	76	69	59
Happiness	56	58	56	49	53	40	36	19
Interest	53	55	51	50	53	47	36	26
Neutral	82	82	87	79	82	74	41	23
Panic	62	62	66	62	59	56	49	44
Pride	28	29	30	30	30	28	16	09
Sadness	61	64	64	56	60	51	29	28
Shame	44	44	44	40	45	37	23	16
Average	65.67	66.40	66.13	63.40	64.27	58.87	44.73	33.40

Table 3. Recognition accuracies (%) achieved with different feature sizes

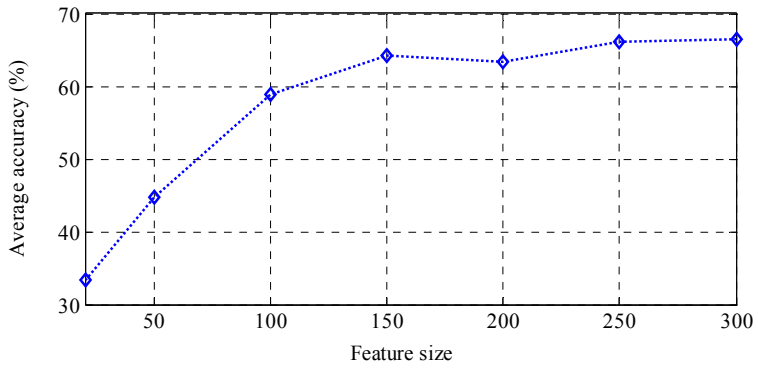


Fig. 8. Comparison of the average accuracies achieved with different feature sizes.

It can be seen from the figure that the average accuracy is not reduced even when the dimensionality of the feature vector is decreased from 792 to 250. The average accuracy is only decreased by 1.40% when the feature size is reduced to 150. This is only 18.94% of the size of the original full feature set. The recognition performance, however, is largely reduced when the feature size is lower than 150. The average accuracy is as low as 33.40% when there are only 20 parameters in a feature vector. It indicates that the classification performance is not deteriorated when the dimensionality of the feature vectors is reduced to

a suitable value. The calculation complexity is also reduced with a decreased dimensionality.

5. Conclusion

The automatic recognition of emotional states from human speech has found a broad range of applications, and as such has drawn considerable attention and interest over the recent decade. Speech emotion recognition can be formulated as a standard pattern recognition problem and solved using machine learning technology. Specifically, feature extraction, processing and dimensionality reduction as well as pattern recognition have been discussed in this chapter. Three short time cepstral features, Linear Prediction-based Cepstral Coefficients (LPCC), Perceptual Linear Prediction (PLP) Cepstral Coefficients, and Mel-Frequency Cepstral Coefficients (MFCC), are used in our work to recognize speech emotions. Feature statistics are extracted based on speech segmentation for capturing longer time characteristics of speech signal. In order to reduce computational cost in classification, Principal Component Analysis (PCA) is employed for reducing feature dimensionality. The Support Vector Machine (SVM) is adopted as a classifier in emotion recognition system. The experiment in the classification of 15 emotional states for the samples extracted from the LDC database has been carried out. The recognition accuracies achieved with different segmentation forms and different feature set sizes are compared for speaker dependent training mode.

6. References

Amir, N. (2001), Classifying emotions in speech: A comparison of methods, *Eurospeech*, 2001.

Cen, L., Ser, W. & Yu., Z.L. (2009), Automatic recognition of emotional states from human speeches, *to be published in the book of Pattern Recognition*.

Clavel, C., Vasilescu, I., Devillers, L. & Ehrette, T. (2004), Fiction database for emotion detection in abnormal situations, *Proceedings of International Conference on Spoken Language Proces*s, pp. 2277–2280, 2004, Korea.

Cowie, R. & Douglas-Cowie, E. (1996), Automatic statistical analysis of the signal and prosodic signs of emotion in speech, *Proceedings of International Conference on Spoken Language Processing (ICSLP '96)*, Vol. 3, pp. 1989–1992, 1996.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al (2001), Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine*, Vol. 18, No. 1, (Jan. 2001) pp. 32-80.

Davis, S.B. & Mermelstein, P. (1980), Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 28, No. 4, (1980) pp. 357-365.

Davitz, J.R. (Ed.) (1964), *The Communication of Emotional Meaning*, McGraw-Hill, New York.

Dellaert, F., Polzin, T. & Waibel, A. (1996), Recognizing emotion in speech, *Fourth International Conference on Spoken Language Processing*, Vol. 3, pp. 1970-1973, Oct. 1996.

Fonagy, I. (1978), A new method of investigating the perception of prosodic features. *Language and Speech*, Vol. 21, (1978) pp. 34–49.

Fradkin, D. & Muchnik, I. (2006), Support Vector Machines for Classification, in Abello, J. and Carmode, G. (Eds), Discrete Methods in Epidemiology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 70, (2006) pp. 13–20.

Havrdova, Z. & Moravek, M. (1979), Changes of the voice expression during suggestively influenced states of experiencing, *Activitas Nervosa Superior*, Vol. 21, (1979) pp. 33–35.

Hermansky, H. (1990), Perceptual linear predictive (PLP) analysis of speech, *The Journal of the Acoustical Society of America*, Vol. 87, No. 4, (1990) pp. 1738-1752.

Huttar, G.L. (1968), Relations between prosodic variables and emotions in normal American English utterances, *Journal of Speech Hearing Res.*, Vol. 11, (1968) pp. 481–487.

Lee, C. & Narayanan, S. (2005), Toward detecting emotions in spoken dialogs, *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 2, (March 2005) pp. 293-303.

McGilloway, S., Cowie, R. & Douglas-Cowie, E. (1995), Prosodic signs of emotion in speech: preliminary results from a new technique for automatic statistical analysis, *Proceedings of Int. Congr. Phonetic Sciences*, Vol. 1, pp. 250–253, 1995, Stockholm, Sweden.

Morrison, D., Wang, R. & Liyanage C. De Silva (2007), Ensemble methods for spoken emotion recognition in call-centres, *Speech Communication*, Vol. 49, No. 2, (Feb. 2007) pp. 98-112.

Nguyen, T. & Bass, I. (2005), Investigation of combining SVM and Decision Tree for emotion classification, *Proceedings of 7th IEEE International Symposium on Multimedia*, pp. 540-544, Dec. 2005.

Nicholson, J., Takahashi, K. & Nakatsu, R. (1999), Emotion recognition in speech using neural networks, *6th International Conference on Neural Information Processing*, Vol. 2, pp. 495–501, 1999.

Oudeyer, P.Y. (2003), The production and recognition of emotions in speech: features and algorithms, *International Jounal of Human-Computer Studies*, Vol. 59, (2003) pp. 157-183.

Picone, J.W. (1993), Signal modeling techniques in speech recognition, *Proceedings of the IEEE*, Vol. 81, No. 9, (1993) pp. 1215-1245.

Petrushin, V.A. (1999), Emotion in speech: recognition and application to call centers, *Proceedings of Artificial Neural Networks in Engineering*, (Nov. 1999) pp. 7-10.

Petrushin, V.A. (2000), Emotion recognition in speech signal: experimental study, development, and application, *Proceedings of the 6th International Conference on Spoken Language Processing*, 2000, Beijing, China.

Psutka, J. Muller, L., & Psutka. J.V. (2001), Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task, Proc. Eurospeech, 2001.

Reynolds, D.A., Quatieri, T.F. & Dunn, R.B. (2000), Speaker verification using adapted Gaussian mixture model, *Digital Signal Processing*, Vol. 10, No. 1, (Jan. 2000) pp. 19-41.

Rong J., Chen, Y-P. P., Chowdhury, M. & Li, G. (2007), Acoustic features extraction for emotion recognition, *IEEE/ACIS International Conference on Computer and Information Science*, Vol. 11, No. 13, pp. 419-424, Jul. 2007.

Scherer, K, A. (2000), Cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology, *Proceedings of ICSLP*, pp. 379–382, Oct. 2000, Beijing, China.

Ser, W., Cen, L. & Yu. Z.L. (2008), A hybrid PNN-GMM classification scheme for speech emotion recognition, *Proceedings of the 19th International Conference on Pattern Recognition (ICPR)*, December, 2008, Florida, USA.

Specht, D. F. (1988), Probabilistic neural networks for classification, mapping or associative memory, *Proceedings of IEEE International Conference on Neural Network*, Vol. 1, pp. 525-532, Jun. 1988.

Steinwart, I. & Christmann, A. (2008), *Support Vector Machines*, Springer-Verlag, New York, 2008, ISBN 978-0-387-77241-7.

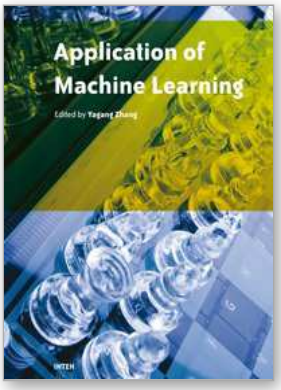
Van Bezooijen, R. (1984), *Characteristics and recognizability of vocal expressions of emotions*, Foris, Dordrecht, The Netherlands, 1984.

Vapnik, V. (1995), *The nature of statistical learning theory*, Springer-Verlag, 1995, ISBN 0-387-98780-0.

Ververidis, D. & Kotropoulos, C. (2006), Emotional speech recognition: resources, features, and methods, *Speech Communication*, Vol. 48, No.9, (Sep. 2006) pp. 1163-1181.

Yu, F., Chang, E., Xu, Y.Q. & Shum, H.Y. (2001), Emotion detection from speech to enrich multimedia content, *Proceedings of Second IEEE Pacific-Rim Conference on Multimedia*, October, 2001, Beijing, China.

Zhou, J., Wang, G.Y., Yang,Y. & Chen, P.J. (2006), Speech emotion recognition based on rough set and SVM, *Proceedings of 5th IEEE International Conference on Cognitive Informatics*, Vol. 1, pp. 53-61, Jul. 2006, Beijing, China.



Application of Machine Learning
Edited by Yagang Zhang

ISBN 978-953-307-035-3
Hard cover, 280 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

The goal of this book is to present the latest applications of machine learning, which mainly include: speech recognition, traffic and fault classification, surface quality prediction in laser machining, network security and bioinformatics, enterprise credit risk evaluation, and so on. This book will be of interest to industrial engineers and scientists as well as academics who wish to pursue machine learning. The book is intended for both graduate and postgraduate students in fields such as computer science, cybernetics, system sciences, engineering, statistics, and social sciences, and as a reference for software professionals and practitioners. The wide scope of the book provides them with a good introduction to many application researches of machine learning, and it is also the source of useful bibliographical information.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821