We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

22

Automatic Recognition of Emotional States From Human Speeches

Ling Cen[1], Wee Ser[2], Zhu Liang Yu[3] and Wei Cen[4]
[1]*Institute for Infocomm Research*
Singapore
[2]*School of Electrical and Electronic Engineering,*
Nanyang Technological University
Singapore
[3]*College of Automation Science and Engineering,*
South China University of Technology,
Guangzhou, China
[4]*Elektrotechnik GmbH, Postfach, Kassel, Germany*

1. Introduction

As computer-based applications receive increasing attention in the real world and in our daily life, the Human-Computer Interaction (HCI) technology has also advanced rapidly over the recent decades. One essential enabler of natural interaction between human and computers is the computers' ability to understand the emotional states expressed by the human subjects (so that personalized responses can be delivered accordingly). This is not surprising as it is well known that emotion plays an important role in human-human communications. Emotions are mental and physiological states associated with feelings, thoughts, and behaviors of human subjects. The emotional state expressed by a human subject reflects not only the mood but also the personality of the human subject.

Automatic recognition of emotional states from cues expressed by human subjects, such as face expression or tone of voice, has found increasing applications in security, learning, medicine, entertainment, etc. For example, detecting abnormal emotions, such as stress or nervousness, helps to detect lie or identify suspicious human subjects. Emotion recognition in automatic tutoring systems, such as web-based e-learning, can adjust the tutoring content and delivery speed according to users' responses. Automatically recognizing emotions from patients could also be helpful in clinical studies as well as in psychosis monitoring and diagnosis assistance. Emotion classification has been applied in the customer service sector

This work was supported in part by the National Natural Science Foundation of China under Grant 60802068, and Guangdong Natural Science Foundation under Grant 8451064101000498, Program for New Century Excellent Talents in University under Grant NCET-10-0370 and the Fundamental Research Funds for the Central Universities, SCUT under Grant 2009ZZ0055.

too, where machines at call centers adjust their responses automatically according to the emotions expressed by the customers (e.g. anger, frustration, satisfaction, etc.). In the entertainment sector, interactive games have been developed that can interact adaptively with human players too. All these examples clearly illustrate that the demand for natural human-like machines is the major motivation or driving factor for the increasing research effort invested in automatic identification of the emotional states of human subjects.

Speech conversation is an important way for natural and effective communications between humans and computers. Over the past decades, advancement in robust speech recognition and synthesis techniques has contributed significantly in making such communications natural and effective. Human speech conveys not only the linguistic content but also the emotion of the speaker. Although the emotion does not alter the linguistic content, it carries important information on the speaker's desire and intent (Cowie et al., 2001; Ververidis & Kotropoulos, 2006). As such, it is important that computers understand the emotional states conveyed in human speech for effective human-computer interaction applications.

Modeling and analysis of emotions from human speech span across several fields, including psychology, linguistics, and engineering. As there is a lack of precise definition and models for emotions, automatic recognition of emotions has been a challenging task to researchers. Indeed, research on speech based emotion recognition has been undertaken by many for around 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). In engineering, speech emotion recognition has been formulated as a pattern recognition problem that involves feature extraction and emotion classification. This is the model adopted in this chapter. In particular, this chapter discusses the designs and performances of several popular classification techniques namely, the Probabilistic Neural Network (PNN), the Universal Background Model - Gaussian Mixture Model (UBM-GMM), the Support Vector Machines (SVMs), and the Hidden Markov model (HMM), for emotion classifications. For completeness, a hybrid technique that combines the strengths of multiple classifiers (Ser et al., 2008) will also be discussed. Experimental results using the LDC database (University of Pennsylvania) will be presented and discussed in the chapter too, to provide a feel of the recognition accuracies for the various emotion recognition techniques described above.

The remaining part of this chapter is organized as follows. Some related works are briefly presented in Section 2. The acoustic feature extraction process is discussed in Section 3. In Section 4, several popular classifiers, including a hybrid method, for emotion recognition are discussed. Experimental results and performance comparison are shown in Section 5 and the concluding remarks are given in Section 6.

2. Related Works

Speech emotion recognition can be formulated as a pattern recognition problem. Such pattern recognition machines consist of two major modules, i.e. feature extraction (including speech signal pre-processing) and emotion classification. Fig. 1 shows a typical structure of the speech emotion recognition system.

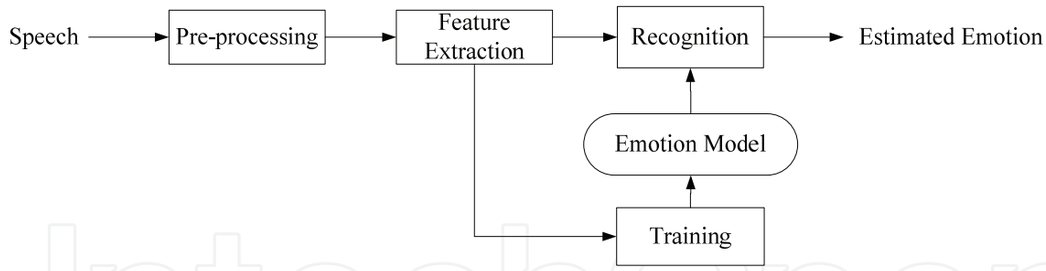


Fig. 1. Typical structure of a speech emotion recognition system

Feature extraction is an important module that provides the acoustic correlates of emotions in human speech for emotion classification. The basic acoustic features extracted directly from the original speech signals, e.g. pitch and intensity related features, are widely used in speech emotion recognition (Ververidis & Kotropoulos, 2006; Lee & Narayanan, 2005; Dellaert et al., 1996; Petrushin, 2000; Amir, 2001). 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. The pitch of speech is the main acoustic correlate of tone and intonation, which represents the highness or lowness of a tone as perceived by the ear. It depends on the number of vibrations per second produced by the vocal cords. As the pitch is related to the tension of the vocal folds and the subglottal air pressure, it can provide the information about the emotions of speakers (Ververidis & Kotropoulos, 2006). The energy related features are also commonly used in emotion recognition. As the pitch and energy are calculated on a frame basis, statistics of the extracted features are usually used, such as mean, median, range, standard deviation, maximum, minimum, and linear regression coefficient (Lee & Narayanan, 2005; Ververidis & Kotropoulos, 2006). Speech rate and ratio of duration of voiced and unvoiced are considered to indicate emotional states of speakers too (Lee & Narayanan, 2005; Ververidis & Kotropoulos, 2006).

The behavior of the acoustic features in different emotional states has been studied in the literature (Davitz, 1964; Huttar, 1968; Fonagy, 1978; Moravek, 1979; Van Bezooijen, 1984; Havrdova & McGilloway et al., 1995, Ververidis & Kotropoulos, 2006). Anger is associated with the highest energy and pitch level among anger, disgust, fear, joy and sadness. Disgust has a lower mean pitch level, a lower intensity level and a slower speech rate than the neutral state. Low levels of the mean intensity and mean pitch are measured with sadness. A high pitch level and a raised intensity level are found in speech expressed with fear. The pitch contour trend separates fear from joy. A downwards slope in the pitch contour can be observed in speech expressed with fear and sadness, while the speech with joy shows a rising slope. Observable variance on speech rate is found in different emotions. More detailed information can be found in (Ververidis & Kotropoulos, 2006). However, even though many researches have been carried out to find acoustic features suitable for emotion recognition, there is still no conclusive evidence to show which set of features can provide the best recognition accuracy (Zhou, 2006).

After the acoustic features are extracted and processed, they are sent to emotion classification module. Many popular classifiers are employed in the literature. 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, happy, 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%.

3. Acoustic Feature Extraction for Emotion Recognition

In this section, the features used in our work and the process involved are briefly described. In the experiments given below, three short time cepstral features are extracted, which are Perceptual Linear Prediction (PLP) Cepstral Coefficients, Mel-Frequency Cepstral Coefficients (MFCC), and Linear Prediction-based Cepstral Coefficients (LPCC). Before extracting the raw features, the speech data are first high-pass filtered by a FIR filter given by

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

Signal frames of length 25 msec are then extracted from the filtered speech signal at an interval of 10 msec. A Hamming window is applied to each signal frame to reduce signal discontinuity. The list below shows the feature set used in this chapter.

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

Fusing the PLP, MFCC and LPCC features, a vector with dimension of R^M is achieved, where M = 132 is the total number of the features extracted for each frame.

4. Classifiers for Emotion Recognition

The features extracted from the speech samples as described in the previous section, are sent to the emotion classification module. The module output is the estimated emotion category of an utterance. Before a classifier can be used to automatically label the emotion categories, a training process has to be carried out. The speech samples in the whole database are divided into two parts. One is used to train the classifiers, and the other is for the test use. In the below sub-sections, we will introduce several popular classification methods used in emotion classification.

4.1 Probabilistic Neural Network (PNN)

The Probabilistic Neural Network (PNN) (Specht, 1988) has been employed as an excellent pattern classifier in many applications due to its excellent characteristics such as simple training, quick convergence and easy implementation. The PNN solves classification problems using Bayesian classifiers. A basic structure of a PNN is shown in Fig. 2. It consists of 4 network layers, i.e. input layer, pattern layer, summation layer and output layer.

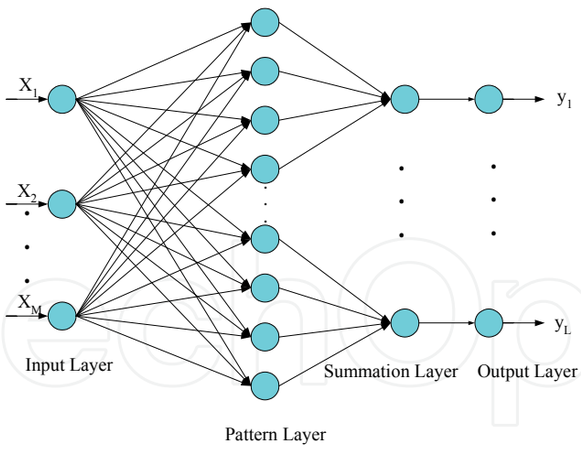


Fig. 2. Structure of a Probabilistic Neural Network

As shown in Fig. 2, the input of the PNN, $\mathbf{x} = [x_1, x_2, ..., x_M]^T$ is the M-dimension feature vector. The distribution function that estimates the likelihood of an input feature vector belonging to a learned category is developed in the pattern layer via supervised training using a given training set. This layer works in the same way as a Bayes classifier, where the class dependent Probability Density Functions (PDF) are approximated using a Parzen

estimator. Each unit in the pattern layer represents an exemplar in the training set. The activated function in this layer can be a Gaussian function given by

$$f(z_j) = \exp\left[\left(z_j - 1\right)/\sigma^2\right], \text{ for } j = 1, 2, ..., N, \tag{2}$$

where σ, which is also called smoothing factor, is the variance of the Gaussians, N is the number of the exemplars in the training set (equal to the number of the units in this layer), and z_j is the weighted input of the j^{th} unit expressed as

$$z_j = \mathbf{x} \cdot \mathbf{w}_j. \tag{3}$$

Here, $\mathbf{w}_j = [w_{j1}, w_{j2}, ..., w_{jM}]^T$ is the M-dimensional weighting vector for the j^{th} exemplar in the training set. Let $\mathbf{x}_j^{c_i}$ for j = 1,…,N_i denote the N_i training exemplars belonging to emotion c_i ($c_i \in C$, where C is the vector of class labels under reorganization). The probability density function, $p\left(\mathbf{x}\middle|c_i\right)$, is expressed as

$$p(\mathbf{x} \mid c_i) = \frac{1}{\sqrt{(2\pi)^M}\sigma^M} \cdot \frac{1}{N_i} \cdot \sum_{j=1}^{N_i} \exp\left(\frac{\left(\mathbf{x} - \mathbf{x}_j^{c_i}\right)^T \left(\mathbf{x} - \mathbf{x}_j^{c_i}\right)}{2\sigma^2}\right). \tag{4}$$

Via training process, the outputs from the units that belong to one emotion category are combined in the summation layer. Each unit in the summation layer is associated to one emotion category. The output layer works in a competitive way, where only one category is generated for any given input vector. The output $\mathbf{y} = \left[y_1, y_2, ..., y_L\right]^T$ is an L-dimension vector, where L is the number of emotion categories. The output of the unit related to the predicted category has the value of "1" and the others have "0".

4.2 Universal Background Model - Gaussian Mixture Model (UBM-GMM)

The Gaussian Mixture Model (GMM) assumes that the observed variables are generated via a probability density distribution that is the weighed linear combination of a set of Gaussian PDF. It is considered as a single-state HMM with a Gaussian mixture observation density and has been shown to be the most successful probability density function in text-independent speaker recognition, where no *prior* knowledge is available on what the speaker will say (Reynolds, 2000).

In the GMM, the distribution of a random variable $\mathbf{x} \in R^M$ is a mixture of G Gaussians given as

$$p(\mathbf{x} \mid \lambda_{c_i}) = \sum_{g=1}^{G} w_g p_g(\mathbf{x}), \tag{5}$$

where λ_{c_i} is the density model related to the class, c_i, G is the number of the Gaussian components, and w_g is the mixture weights satisfying the constraint $\sum_{g=1}^{G} w_g = 1$. The Gaussian densities, p_g, is shown as

$$p_g(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^M |\Sigma_g|}} \cdot \exp\left(-\frac{(\mathbf{x}-\mu_g)^T (\mathbf{x}-\mu_g)}{2\Sigma_g} \right), \tag{6}$$

where μ_g and Σ_g are the M-dimension mean vector and $M \times M$ -dimension covariance matrix, respectively. The density model, λ_{c_i} is denoted as $\lambda_{c_i} = \{w_g, \mu_g, \Sigma_g\}_{g=1}^{G}$, which represents the probability density distribution of the feature vectors (**x**) for the category c_i. The optimum set of parameters of λ_{c_i} can be identified in an iterative manner using the Maximal likelihood Principle (MLP) and Expectation-Maximization (EM) algorithm. The likelihood of $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_i}$ is defined as

$$L(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_i} \mid \lambda_{c_i}) = \prod_{j=1}^{N_i} \sum_{g=1}^{G} w_g p_g(\mathbf{x}_j), \tag{7}$$

where $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{N_i}\}$ are the exemplars in the training set belonging to the category, c_i. The EM algorithm iteratively updates w_g, μ_g, Σ_g for g = 1,…,G, in order to monotonically increase the likelihood in (7).

The EM algorithm consists of two steps: Expectation Step and Maximization Step. In the Expectation Step, we calculate

$$e_{jg} = p(g \mid \mathbf{x}_j) = \frac{w_g p_g(\mathbf{x}_j)}{\sum_{t=1}^{G} w_t p_t(\mathbf{x}_j)}, \text{for g} = 1,...G. \tag{8}$$

Then in the Maximization Step, w_g, μ_g, Σ_g for g = 1,…,G, are updated as

$$\hat{a}_g = \frac{1}{N}\sum_{j=1}^{N_i} e_{jg},$$
$$\hat{\mu}_g = \frac{\sum_{j=1}^{N_i} e_{jg}\mathbf{x}_j}{\sum_{j=1}^{N_i} e_{jg}}, \tag{9}$$
$$\hat{\Sigma}_g = \frac{\sum_{j=1}^{N_i} e_{jg}\left(\mathbf{x}_j - \mu_g\right)^T\left(\mathbf{x}_j - \mu_g\right)}{\sum_{j=1}^{N_i} e_{jg}}.$$

By iteratively performing the above steps, the EM algorithm is able to find an optimum set of parameters for the GMM.

In order to handle mismatches more effectively, the Universal Background Model (UBM) is incorporated into the GMM, and the resultant model is denoted as UBM-GMM (Reynolds, 2000). In the UBM-GMM, not only the hypothesis that an utterance belongs to an emotion category, but also the hypothesis that it does not belong to this category, are tested. Each of the categories is trained with two models. The emotion model, λ_{c_i} is trained using the training samples belonging to the emotion class c_i, and a background model, $\lambda_{\bar{c}_i}$ is meantime trained using those samples that do not belong to c_i. When the emotional state of a new utterance with feature vector **x** is recognized, both λ_{c_i} and $\lambda_{\bar{c}_i}$ are used to generate the PDF of the feature vector, $p\left(\mathbf{x}\middle|\lambda_{c_i}\right)$ and $p\left(\mathbf{x}\middle|\lambda_{\bar{c}_i}\right)$ as illustrated in Fig. 3. The emotion category of the utterance is determined using the likelihood ratio

$$\frac{p\left(\mathbf{x}\middle|\lambda_{c_i}\right)}{p\left(\mathbf{x}\middle|\lambda_{\bar{c}_i}\right)} \geq \eta, \tag{10}$$

where η is a predetermined threshold. In UBM-GMM, the log-likelihood ratio, S_{c_i}, is often used, given as

$$S_{c_i} = \log p\left(\mathbf{x}\middle|\lambda_{c_i}\right) - \log p\left(\mathbf{x}\middle|\lambda_{\bar{c}_i}\right) \geq \eta', \eta' = \log\eta. \tag{11}$$

Highest S_{c_i} determines the emotion class of **x**, which implies higher $p\left(\mathbf{x}\middle|\lambda_{c_i}\right)$ and lower $p\left(\mathbf{x}\middle|\lambda_{\bar{c}_i}\right)$.

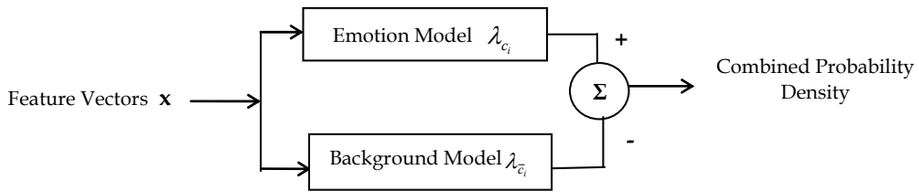


Fig. 3. Output Probability Density of UBM-GMM

4.3 Hidden Markov Model (HMM)

The HMM has a good representation of the temporal behavior of signals being modeled. As such, it is commonly used in temporal based pattern recognition applications (e.g. speech recognition, handwriting recognition, gesture recognition, etc.). In HMM, the system to be modeled is assumed to be a Markov process with a finite set of concatenated states. The states cannot be observed directly, while the observation that is generated in the states according to the associated probability distribution is visible, from which the model is named as Hidden Markov Model.

Assume that there are N states, $S=\{S_1, S_2, \ldots, S_N\}$, and M distinct observation symbols per state, $V=\{v_1, v_2, \ldots, v_M\}$, in the model. Each of these states is associated with a probability distribution over the possible outcome. Transition among the states is controlled by a set of transition probabilities. Let $\Lambda = \{a_{ij}\}$ represent the state transition probabilities given as

$$a_{ij} = p\{q_{t+1} = S_j \,|\, q_t = S_i\}, \text{ for } 1 \le i, j \le N, \tag{12}$$

where q_t is the state at time t and a_{ij} satisfies

$$a_{ij} \ge 0, \ \sum_{j=1}^{N} a_{ij} = 1, \text{ for } 1 \le i, j \le N. \tag{13}$$

The observation symbols correspond to the physical output of the system being modeled. Let $B = \{b_j(k)\}$ denote the observation symbol probability distribution in state j, where

$$b_j(k) = p\{O_t = v_k \,|\, q_t = S_j\}, \text{ for } 1 \le j \le N, \ 1 \le k \le M, \tag{14}$$

and o_t is the observation in time t and v_k is the k^{th} observation.

The distribution $b_j(k)$ satisfies

$$b_j(k) \ge 0, \ \sum_{k=1}^{M} b_j(k) = 1, \text{ for } 1 \le j \le N, \ 1 \le k \le M. \tag{15}$$

Assume that the initial state distribution $\pi = \{\pi_i\}$ is given as

$$\pi_i = p\{q_1 = S_i\}, \text{ for } 1 \le i \le N. \tag{16}$$

An HMM can then be represented as $\lambda = (\Lambda, B, \pi)$. Given a suitable set of the values of *N*, *M* and the initial state distribution, the model can be trained to solve the three fundamental problems (Rabiner, 1989). The process involved is summarized below.

1) Given an observation sequence with the following *T* observations

$$O = O_1 O_2 ... O_T,\ O_t \in V,\ t = 1, 2, ..., T, \tag{17}$$

and the model $\lambda = (\Lambda, B, \pi)$, evaluate $p(O|\lambda)$, i.e. the probability of the observation sequence. As the calculation of enumerating every possible state sequence of length *T* for calculating $p(O|\lambda)$ involves on the order of $2T \cdot N^T$ calculations, the problem can be solved efficiently using the forward algorithm that requires on only the order of $N^2 T$ calculations. Consider forward variable $\alpha_t(i)$ that is defined as the probability of the partial observation sequence (until time *t*), $O_1 O_2 ... O_t$, and the state S_i at time *t*, given the model λ, expressed as

$$\alpha_t(i) = P(O_1 O_2 ... O_t, q_t = S_i \mid \lambda). \tag{18}$$

This can be solved inductively, as follows:

Initialization

$$\alpha_1(i) = \pi_i b_i(O_1),\ 1 \le i \le N. \tag{19}$$

Induction

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}),\ 1 \le t \le T-1,\ 1 \le j \le N. \tag{20}$$

Termination

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i). \tag{21}$$

2) Given the observation sequence $O = O_1 O_2 ... O_T$ and the model λ, determine the optimum sequence of the model states. There is no exact and unique solution for this problem. Optimality criteria are applied to find optimal state sequence. The Viterbi algorithm (Viterbi, 1967; Forney, 1973) that finds the single best state sequence based on dynamic programming methods is commonly used to solve this problem.

3) Estimate the model parameters $\lambda = (\Lambda, B, \pi)$ that best match the observed signal, i.e. maximizing the probability of the observation sequence given the model, $p(O|\lambda)$. As there is no analytical approach to solve the problem, λ is usually chosen such that $p(O|\lambda)$ is locally maximized using an iterative procedure such as the Baum-Welch method.

In the above, we have briefly introduced the algorithm for solving the first problem as speech emotion recognition belongs to this problem and can be solved efficiently using HMMs. The observation sequence is the feature vector of an utterance. Through training process, one HMM is established for each emotion category. If an individual training is carried out for each speaker, *S* HMMs are trained for each emotion, where S is the number of speakers. With the trained HMMs, estimating the emotion category of an utterance is equivalent to calculating the probability of $p(O|\lambda)$ for the given observation sequence. It can be solved using the forward algorithm described above. The HMM with the highest probability determines the emotion category of the utterance.

4.4 Support Vector Machines (SVMs)

SVMs that developed by Vladimir 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 dataset, 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

$$\{\mathbf{x}_i, c_i\},\ i = 1,...,N,\ \mathbf{x}_i \in \mathbf{R}^M, c_i \in \{-1,1\} \tag{22}$$

where $\mathbf{x}_i = [x_1, x_2, ..., x_M]$ is the *M*-dimension feature vector of the *i*th samples, *N* is the number of samples and c_i is the category to which $\mathbf{x}_i$ belongs. Suppose there is a hyperplane that the separates feature vectors $\phi(\mathbf{x}_i)$ with positive category from the negative one, here $\phi(\bullet)$ is a nonlinear mapping of the input space into higher dimensional feature space. The set of points $\phi(\mathbf{x})$ that lie on the hyperplane is expressed as

$$\mathbf{w} \cdot \phi(\mathbf{x}) + b = 0, \tag{23}$$

where $\mathbf{w}$ and *b* are the two parameters. For the training data that are linearly seperable, two hyperplanes are selected to yield maximum margin. Suppose $\mathbf{x}_i,\ i = 1,...,N$ satisfies

$$\begin{aligned} \phi(\mathbf{x}_i)\cdot\mathbf{w}+b \geq 1, \text{ for } c_i = 1, \\ \phi(\mathbf{x}_i)\cdot\mathbf{w}+b \leq 1, \text{ for } c_i = -1. \end{aligned} \tag{24}$$

It can be re-written as

$$c_i\left(\phi(\mathbf{x}_i)\cdot\mathbf{w}+b\right)-1 \geq 0, \ \forall i = 1,2,...,N. \tag{25}$$

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

$$\begin{aligned} &\text{Minimize } \|\mathbf{w}\|^2 \\ &\text{subject } c_i\left(\phi(\mathbf{x}_i)\cdot\mathbf{w}+b\right) \geq 1, \forall i = 1,2,...,N. \end{aligned} \tag{26}$$

where $\|\mathbf{w}\|$ 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 (26) is given as

$$\begin{aligned} &\text{Minimize } W(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} c_i c_j \alpha_i \alpha_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j), \\ &\text{subject } \sum_{i=1}^{N} c_i\alpha_i = 0, \ \alpha_i \geq 0, \ \forall i = 1,2,...,N. \end{aligned} \tag{27}$$

where α_i is the Lagrangian variable.

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 (27) that the training points are appeared as their inner products in the dual formulation. According to Mercer's theorem, any symmetric positive semi-definite function $k\left(\mathbf{x}_i, \mathbf{x}_j\right)$ implicitly defines a mapping into a feature space

$$\phi : \mathbf{x} \rightarrow \phi\left(\mathbf{x}\right) \tag{28}$$

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

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \phi\left(\mathbf{x}_i\right)\cdot\phi\left(\mathbf{x}_j\right) \tag{29}$$

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

$$\begin{aligned} &\text{Minimize } W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} c_i c_j \alpha_i \alpha_j k\left(\mathbf{x}_i, \mathbf{x}_j\right), \\ &\text{subject } \sum_{i=1}^{N} c_i \alpha_i = 0,\ \alpha_i \geq 0,\ \forall i = 1, 2, ..., N. \end{aligned} \tag{30}$$

By replacing the inner product in (27) 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 in below:

Polynomial (homogeneous) kernel: $k\left(\mathbf{x}, \mathbf{x}'\right) = \left(\mathbf{x} \cdot \mathbf{x}'\right)^d$,

Polynomial (inhomogeneous) kernel: $k\left(\mathbf{x}, \mathbf{x}'\right) = \left(\mathbf{x} \cdot \mathbf{x}' + 1\right)^d$,

Radial basis kernel: $k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\gamma \left\|\mathbf{x} - \mathbf{x}'\right\|^2\right)$, for $\gamma > 0$,

Gaussian radial basis kernel: $k\left(\mathbf{x}, \mathbf{x}'\right) = \exp\left(-\frac{\left\|\mathbf{x} - \mathbf{x}'\right\|^2}{2\sigma^2}\right)$.

A single SVM itself is a classification method for 2-category data. In speech emotion recognition, there are usually multiple emotion categories. Two approaches are commonly used to solve the problem, namely one-versus-all and one-versus-one (Fradkin and Muchnik, 2006). In the first approach, one SVM is built for each emotion. In the second approach, 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 way, 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 way, 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.

4.5. A Hybrid Classification Method

In the previous studies, most emotion recognition schemes use a single classifier, and very few have considered hybrid classification methods (Morrison, 2007). Intuitively, if the individual schemes can be suitably combined, an improvement in accuracy can be expected. This section describes a recently reported hybrid scheme that combines the strengths of multiple classifiers (Ser et al., 2008).

The structure of the hybrid scheme is shown in Fig. 4, which consists of two basic classifiers, i.e. the PNN classifier and the UBM-GMM classifier. The outcomes of these two classifiers are fused together to generate the final result by the Fusion Look-Up Table (LUT) approach.

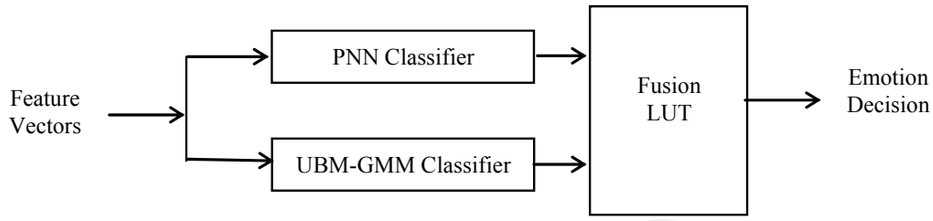


Fig. 4. Structure of Proposed Hybrid Scheme (LUT: Look Up Table)

In the training stage, the PNN and the UBM-GMM classifiers are firstly trained individually. Then the confusion matrices for the two base classifiers and a LUT that is denoted as *F*, is formed using a different part of training data from that used to train the base classifiers. Let *L* be the total number of emotional states to be classified. The *L*-by-*L* confusion matrix for each of the base classifier takes the following form:

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

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

The Fusion LUT, *F*, a matrix with dimension of $L^3 \times 4$, records all possible emotional states estimated by the two classifiers, the actual emotional states, and the conditional probability of the actual emotion being one of the emotional states. This is elaborated below.
Let c_{PNN} and c_{GMM} be the emotional states estimated by PNN and UBM-GMM respectively, and c_r be the actual emotional state. A typical row of *F* takes the form,

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

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

$$p(c_r) = prob(c = c_r \mid c_{PNN}, c_{GMM}). \tag{31}$$

It can be approximated as

$$p(c_r) \approx N_{c_{PNN}-c_{GMM}} / N_{c_r}, \tag{32}$$

where $N_{c_r - c_{PNN} - c_{GMM}}$ represents the number of utterances whose emotions being c_r given that the estimated emotion from the PNN and GMM classifiers being c_{PNN} and c_{GMM}, respectively, and N_{c_r} denotes the number of utterances expressed in the emotion c_r. Note that $N_{c_r - c_{PNN} - c_{GMM}}$ and N_{c_r} count only the utterances used to calculate the LUT. In the ideal situation when the recognition accuracy of every single classifier is 100%, $p(c_r)$ becomes

$$p(c_r) = \begin{cases} 1, & \text{for } c_{PNN} = c_{GMM} = c_r \\ 0, & \text{otherwise} \end{cases}. \tag{33}$$

In the testing stage, the emotion of a speech sample is determined by either the Fusion LUT or the confusion matrices.

The training-testing process can be summarized into the following 6 steps.
Step 1 Train each of the two classifiers, PNN and UBM-GMM, independently using the training data.
Step 2 Use the trained classifiers to recognize the emotions of the utterances extracted from another speech data training set**.**
Step 3 Calculate the confusion matrices for both the base classifiers.
Step 4 Calculate the fusion LUT, *F*, according to the process described before.
Step 5 Apply the two base classifiers to the test data separately, and obtain the estimated emotional states, c_{PNN} and c_{GMM}, respectively.
Step 6 Compare the values of $p(c_r)$ in the fusion LUT where the first 2 indices are c_{PNN} and c_{GMM}. Determine the fusion output as $c_{fus} = c_r$ where the value $p(c_r)$ is the highest. In the case when $p(c_r) = 0$ (which can happen when the training sample size is too small), compare the values of $p_{i,i}$ in the two confusion matrices, where i corresponds to c_{PNN} and c_{GMM}, for the respective confusion matrices. The final decision of the emotional state, c_{fus} is then taken to be the c_r corresponding to the highest $p_{i,i}$.

5. Experiments

5.1 Database

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.

5.2 Experiment Description

The PNN, UBM-GMM, HMM, SVM and the hybrid classification method are employed to automatically recognize emotional states from speech samples. In our experiment, we consider the different characteristics of speech among the speakers. The speech data are trained in speaker dependent training mode, in which an individual training process is carried out for each speaker. In the experiment, the database is divided into two parts, i.e. training dataset and testing dataset. For the PNN, GMM, HMM and SVM classifiers used individually, three quarters of the data are employed to train the classifiers; for the hybrid classification method, half of the data are employed to train the base classifiers, a quarter of the data are used to calculate the Fusion LUT and the confusion matrices. The rest quarter of data are used for testing purpose in our methods.

5.3 Results and Discussion

Numerical results obtained by the PNN, UBM-GMM, HMM, SVM, and the hybrid scheme are shown in Tables 1. The average accuracies achieved by PNN, UBM-GMM, HMM, SVM and the hybrid scheme are 68.60%, 72.73%, 69.60%, 62.67%, and 75.13%, respectively.

	PNN	UBM-GMM	HMM	SVM	Hybrid
Anxiety	79	77	82	76	80
Boredom	71	79	76	81	76
Cold Anger	64	69	58	59	71
Contempt	73	80	79	58	82
Despair	65	79	76	71	79
Disgust	78	89	81	65	84
Elation	59	81	68	73	72
Hot Anger	75	85	85	74	77
Happiness	61	76	49	60	68
Interest	64	70	63	51	73
Neutral	80	54	82	61	81
Panic	62	75	70	52	75
Pride	72	53	53	54	72
Sadness	74	63	51	46	74
Shame	52	61	71	59	63
Average	68.60	72.73	69.60	62.67	75.13

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

The accuracies for individual emotion recognition achieved by the PNN, UBM-GMM, HMM and SVM are plotted in Fig. 5. It is shown from the experiment results that among the 4 classifiers, the UBM-GMM has achieved highest average accuracy. The recognition performance of the HMM in this experiment is similar to that of the UBM-GMM. For each of these emotion categories, the highest recognition accuracy is achieved by different classification method, e.g. for anxiety, the highest accuracy of 82% is achieved by the HMM; for despair, the accuracy of 79% obtained by the UBM-GMM is the highest; for Neutral, the highest accuracy of 82% is achieved by the HMM; and for Pride, the PNN gives the highest accuracy of 72%. It indicates that one cannon simply make a conclusion that one classifier is better than another classifier.

It can be seen from Table 1, the hybrid scheme is able to improve the recognition accuracy compared to the classification methods used individually. The average accuracy is 75.13%, which is 6.53% and 2.40% higher than the PNN and UBM-GMM individually used, respectively. In the literature, usually only 2-6 different emotional states are classified. Considering the difficulties encountered due to the facts that the number of emotional states is as large as 15, the results are rather satisfying.

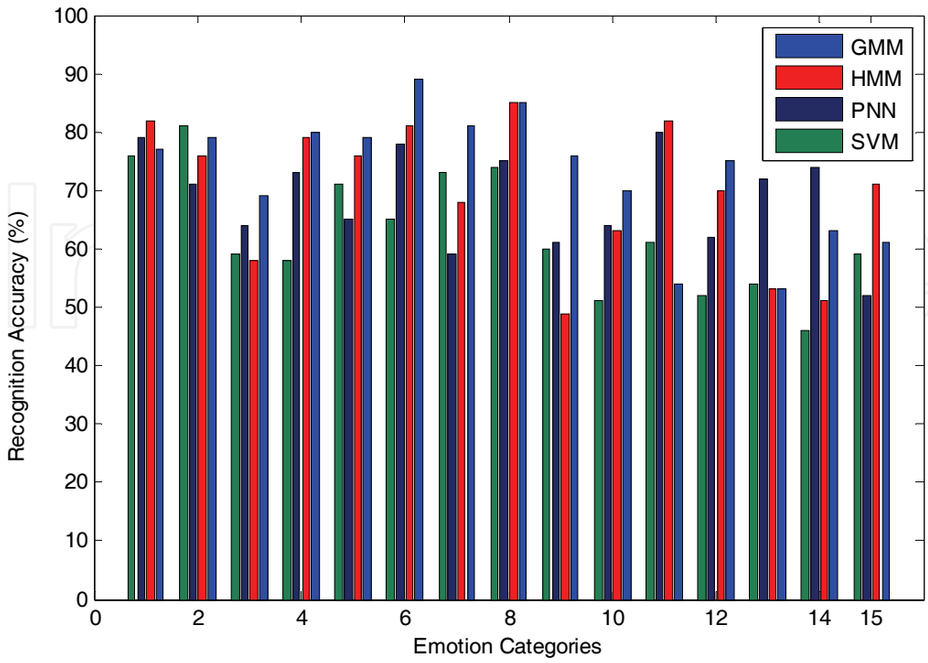


Fig. 5. Recognition accuracies of PNN, UBM-GMM, HMM and SVM

6. Conclusions

Over the recent decades, automatic recognition of emotional states has attracted increasing interest among the researchers. This chapter addresses the problem of emotion recognition from human speech cues. The processes involved and some popular methods for feature extraction and emotion classification have been discussed in the chapter. In particular, acoustic features such as the short time cepstral features, i.e. Perceptual Linear Prediction (PLP) Cepstral Coefficients, the Mel-Frequency Cepstral Coefficients (MFCC), and the Linear Prediction-based Cepstral Coefficients (LPCC), have been discussed in the chapter. Several popular classification methods, including the Probabilistic Neural Network (PNN), the Universal Background Model -Gaussian Mixture Model (UBM-GMM), the Hidden Markov model (HMM), the Support Vector Machines (SVMs), and a recently proposed hybrid method have been discussed too. Experimental results, in terms of recognition accuracies, obtained by using the LDC database (University of Pennsylvania) have been included and discussed in the chapter too.

7. References

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

Clavel, C., Vasilescu, I., Devillers, L.& Ehrette, T. (2004), Fiction database for emotion detection in abnormal situations, Proceedings of International Conference on Spoken Language Process, 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*, 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.

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.

Forney, G.D. (1973), The Viterbi algorithm, *Proc. IEEE*, Vol. 61, (Mar. 1973), pp. 268-278.

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.

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.

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.

Rabiner, L.R. (1989), A tutorial on Hidden Markov Model and selected applications in speech recognition, *Proceeding of the IEEE*, Vol. 77, No. 2, (Feb. 1989) pp. 257-285.

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.

Scherer, K, A. (2000), Cross-cultural inverstigation of emotion inferences from voice and speech: Implications for speech technology, *Proceedings of International Conference on Spoken Language Processing*, 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.

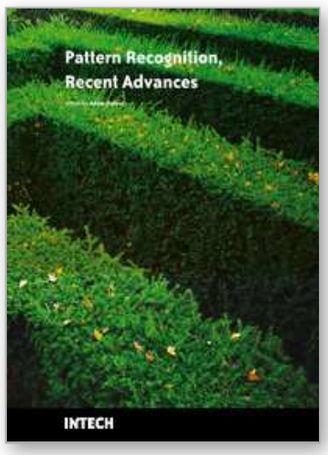
[27] Ververidis, D. & Kotropoulos, C. (2006), Emotional speech recognition: resources, features, and methods, *Speech Communication*, Vol. 48, No.9, (Sep. 2006) pp. 1163-1181.

[28] Viterbi, A.J. (1967), Error bounds for convolutional codes and an asymptotically optimal decoding algorithm, *IEEE Trans. Informat. Theory*, Vol. IT-3, (Apr. 1967) pp. 260-269.

[29] 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.

[30] 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.

IntechOpen



Pattern Recognition Recent Advances
Edited by Adam Herout

ISBN 978-953-7619-90-9
Hard cover, 524 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

Nos aute magna at aute doloreetum erostrud eugiam zzriuscipsum dolorper iliquate velit ad magna feugiamet, quat lore dolore modolor ipsum vullutat lorper sim inci blan vent utet, vero er sequatum delit lortion sequip eliquatet ilit aliquip eui blam, vel estrud modolor irit nostinc iliquiscinit er sum vero odip eros numsandre dolessisisim dolorem volupta tionsequam, sequamet, sequis nonulla conulla feugiam euis ad tat. Igna feugiam et ametuercil enim dolore commy numsandiam, sed te con hendit iuscidunt wis nonse volenis molorer suscip er illan essit ea feugue do dunt utetum vercili quamcon ver sequat utem zzriure modiat. Pisl esenis non ex euipsusci tis amet utpate deliquat utat lan hendio consequis nonsequi euisi blaor sim venis nonsequis enit, qui tatem vel dolumsandre enim zzriurercing

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Ling Cen, Wee Ser, Zhu Liang Yu and Wei Cen (2010). Automatic Recognition of Emotional States From Human Speeches, Pattern Recognition Recent Advances, Adam Herout (Ed.), ISBN: 978-953-7619-90-9, InTech, Available from: http://www.intechopen.com/books/pattern-recognition-recent-advances/automatic-recognition-of-emotional-states-from-human-speeches

INTECH
open science | open minds

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

IntechOpen