Neural Comput & Applic (2013) 23:215–227
DOI 10.1007/s00521-012-0814-8

ORIGINAL ARTICLE

Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method

Mansour Sheikhan · Mahdi Bejani · Davood Gharavian

Received: 10 November 2011 / Accepted: 5 January 2012 / Published online: 20 January 2012

Abstract The speech signal consists of linguistic information and also paralinguistic one such as emotion. The modern automatic speech recognition systems have achieved high performance in neutral style speech recognition, but they cannot maintain their high recognition rate for spontaneous speech. So, emotion recognition is an important step toward emotional speech recognition. The accuracy of an emotion recognition system is dependent on different factors such as the type and number of emotional states and selected features, and also the type of classifier. In this paper, a modular neural-support vector machine (SVM) classifier is proposed, and its performance in emotion recognition is compared to Gaussian mixture model, multi-layer perceptron neural network, and C5.0-based classifiers. The most efficient features are also selected by using the analysis of variations method. It is noted that the proposed modular scheme is achieved through a comparative study of different features and characteristics of an individual emotional state with the aim of improving the recognition performance. Empirical results show that even by discarding 22% of features, the average emotion recognition accuracy can be improved by 2.2%. Also, the proposed modular neural-SVM classifier improves the recognition accuracy at least by 8% as compared to the simulated monolithic classifiers.

Keywords Modular · Neural network · SVM · ANOVA · Speech emotion recognition

M. Sheikhan (✉) · M. Bejani
EE Department, Faculty of Engineering, Islamic Azad University, South Tehran Branch, P.O. Box: 11365-4435, Tehran, Iran
e-mail: msheikhn@azad.ac.ir

M. Bejani
e-mail: st_m_bejani@azad.ac.ir

D. Gharavian
EE Department, Shahid Abbaspour University of Technology, Tehran, Iran
e-mail: gharavian@pwut.ac.ir

1 Introduction

The acoustic and prosodic features of speech are affected by emotions and speaking styles as well as speaker characteristics and linguistic features [1]. Although the emotional state does not alter the linguistic content, it is an important factor in human communication, and improving the voice-based man–machine interactions (MMIs) is one of the key goals in developing automatic emotion recognition (AER) systems [2]. The AER system is a key component in many applications such as spoken tutoring systems [3], medical-emergency domain to detect stress and pain [4], interactions with robots [5, 6], computer games [7], call centers [8, 9] and developing man–machine interfaces for helping weak and old people [10].

Most of the existing efforts in this field have focused on the recognition of a subset of basic emotions from speech signals, e.g., "Anger" and "Neutral" [11–13], "Negative" and "Non-negative" [14, 15], "Emotional" and "Neutral" [16]. However, some studies have been conducted on the interpretation of speech signals in terms of certain application-dependent affective states, e.g., "Annoyed" and "Frustrated" detection [17], detecting "Certainty" [18], "Relief" detection [4], "Stress" detection [19, 20], and detection of irritation and resignation [21].

Recognizing emotions from speech by a machine is first investigated around the mid-1980s using statistical properties of certain acoustic features [22]. In the next decade,

more complicated emotion recognition algorithms were implemented and market requirements motivated further research [19, 23, 24]. In recent years, research is focused on finding reliable informative features and combining powerful classifiers that improve the performance of emotion detection systems in real-life applications [25–33]. In this way, developing optimal design methods for combining classifiers has become an active research field. So, we propose a modular neural-support vector machine (SVM) classifier in this study that combines monolithic multi-layer perceptron (MLP)-based and SVM-based expert systems.

By considering various supplementary features, based on the first three formants (F_1, F_2, and F_3) and pitch frequency (F_0), and concatenating them to a popular feature vector, which includes "Mel-frequency cepstral coefficients (MFCCs) shown by c_i; $i = 1, 2, \ldots, 12$," "log energy (LE)," and "their velocity or first derivative (dc_i, dLE) and their acceleration or second derivative (ddc_i, ddLE)," a new rich medium-size feature vector is proposed in this study. A total of 55 features have been extracted over Farsi language sentences. Recognizing emotional states in speech is performed by using Gaussian mixture model (GMM) [34], MLP neural network, and C5.0 decision tree algorithm [35], and the proposed modular neural-SVM classifier. To reduce the number of features, a feature selection method based on the analysis of variations (ANOVA) method [36] is used in this paper.

The rest of paper is organized as follows: the background and related works are reviewed in Sect. 2. The speech corpus is introduced in Sect. 3. The feature selection approach and corresponding results are presented in Sect. 4. The performance comparison of different monolithic classifiers is performed in Sect. 5. Section 6 presents the proposed modular neural-SVM classifier scheme. Experimental results are given in Sect. 7. Finally, the paper is concluded in Sect. 8.

2 Background and related works

Generally, the emotion recognition system has three components: feature extraction unit, feature selection unit, and emotion recognition unit. First, the features of speech are extracted. These features are the basic acoustic or linguistic features, such as pitch- and spectral-related features. In this way, some transformations can also be employed to convert the speech features between different data domains [37]. The list of some recent researches and corresponding extracted features are given in Table 1.

Table 1 Sample components of feature vectors used in emotion recognition systems from speech in recent years

Research group (year)	Components of feature vector
Kao and Lee (2006)	Pitch, log energy, formants, MFCCs [38]
Ververidis and Kotropoulos (2006)	Pitch, energy, formants, MFCCs, vocal tract cross-section areas, speech rate [27]
Neiberg et al. (2006)	Pitch, MFCCs [39]
Pao et al. (2008)	LPCs[a], MFCCs [40]
Sidorova (2009)	Formants, intensity, pitch [41]
Altun et al. (2009)	Pitch, energy, MFCCs, LPCs [29]
Gajšek et al. (2010)	ZCR[b], energy, pitch, HNR[c] [42]
Yang and Lugger (2010)	Harmony features [43]
Bitouk et al. (2010)	Statistics of MFCCs computed over three phoneme types [44]
Yeh et al. (2010)	Jitter, shimmer, LPC, LPCC[d], MFCCs, dMFCCs, ddMFCCs, LFPC[e], PLP[f] [45]
Wu et al. (2011)	MSFs[g] [46]
Polzehl et al. (2011)	Statistics of pitch, loudness and MFCCs as acoustic features and probabilistic and entropy-based models of words and phrases as linguistic features [13]
He et al. (2011)	Average Renyi entropy for the IMF[h] channels achieved by EMD[i] of speech [47]

[a] Linear prediction coefficients
[b] Zero crossing rate
[c] Harmonics-to-noise ratio
[d] Linear prediction cepstral coefficients
[e] Log frequency power coefficients
[f] Perceptual linear prediction
[g] Modulation spectral features
[h] Intrinsic mode function
[i] Empirical model decomposition

As sample recent researches for feature extraction that are mentioned in Table 1: Yang and Lugger [43] have proposed a set of harmony features for emotion recognition. The mentioned features are based on the psychoacoustic harmony perception known from music theory. Bitouk et al. [44] have used statistics of MFCCs computed over three phoneme types (stressed vowels, unstressed vowels, and consonants). Their experiments have shown that spectral features computed from consonant regions contain more information about emotion than either stressed or unstressed vowel features. Wu et al. [46] have used modulation spectral features (MSFs) for emotion recognition. In this way, they have used an auditory filter-bank and a modulation filter-bank for speech analysis. They have shown that MSFs outperform short-term spectral representations such as MFCCs and perceptual linear prediction (PLP) coefficients. Polzehl et al. [13] have exploited both linguistic and acoustic features for anger classification. In this way, the statistics of pitch, loudness, and MFCCs have been used in acoustic modeling and in linguistic modeling, the probabilistic and entropy-based models of words and phrases, e.g., bag-of-words (BOW), term frequency (TF), term frequency-inverse document frequency (TF-IDF) and the self-referential information (SRI) have been used. He et al. [47] have proposed two feature extraction methods for stress and emotion classification. The first method uses the empirical model decomposition (EMD) of speech into intrinsic mode functions (IMF) and calculates the average Renyi entropy for the IMF channels. The second method calculates the average spectral energy in the sub-bands of speech spectrograms.

The second component in an emotion recognition system reduces the size of feature set by selecting the most relevant subset of features and removing the irrelevant ones [48–51]. As sample researches in this field, Kao and Lee [38] have considered the features at different levels such as frame-level, syllable-level, and word-level. Some feature selection methods such as sequential floating forward selection (SFFS) [52, 53], wrapper approach with forward selection [41], forward feature selection (FFS), and backward feature selection (BFS) [40], principal component analysis (PCA) or linear discriminate analysis (LDA) [54], and fast correlation-based filter (FCBF) [30] have been also used for selecting features in speech emotion recognition systems. In addition, Rong et al. [37] have proposed a decision tree-random forest ensemble hybrid feature selection algorithm that can be applied to a small-size dataset with a high number of features, and the potential benefits of continuous emotion models have been exploited in [55] through a 3-D model.

The third component in this system is a classification model that predicts the emotional states. In 1990s, most of the emotion recognition models were based on the maximum likelihood Bayes (MLB) and linear discriminate classification (LDC) [24]. In the recent years, *K*-nearest neighbor (KNN) [12, 33, 40, 45, 56–58], decision trees [3, 5, 12, 33, 37, 58], Bayesian networks [33, 58], optimum path forest (OPF) classification [58], hidden Markov models (HMMs) [1, 2, 27, 45, 59–61], GMMs [2, 39, 58, 59, 62, 63], variants of support vector machines (SVMs) [12, 27–29, 33, 38, 45, 58, 59, 64–69], artificial neural networks (ANNs) [2, 12, 27, 30, 45, 58, 59, 70–73], and hybrid or ensemble methods [2, 8, 9, 33, 57, 67, 74–82] have been used for emotion recognition.

Similar to the approach that is proposed in this paper, some of the emotion recognition researches have been focused on hybrid or ensemble methods, e.g., multiple classifiers [2, 9, 33, 57, 67, 74–78] and combining different information sources and classifier fusion [8, 79–82].

As example researches for emotion recognition from speech using multiple classifiers: Albornoz et al. [2] have tested standard classifiers based on GMM, HMM, and MLP and proposed a hierarchical method for emotion classification. A class of hierarchical directed graphical models has been proposed in [9] on the task of recognizing affective categories from prosody in both acted and natural speech. In the proposed framework, speech has been structurally modeled as a dynamically evolving hierarchical model in which levels of the hierarchy have been determined by prosodic constituency. Fersini et al. [33] have proposed a multilayer SVM as a hierarchical classification system to recognize emotional states from speech signatures in real courtroom recordings. A multiple KNN classifier has been proposed in [57] using both prosodic and vocal source features. Morrison et al. [67] have used stacked generalization and unweighted vote as classification methods in vocal emotion recognition. Planet et al. [74] have used multiple classifiers with different hierarchical structures. In one of these, they have used a binary classifier to distinguish between 'Neutral' and other emotional classes at the first level and then a multiclass classifier has been used at the next level to separate the rest of classes. Also, Lee et al. [75] have introduced a hierarchical computational structure to recognize emotions. In this way, input speech maps into an emotion class through subsequent layers of binary classifications. Multiple classifier systems have been considered in [76] for the classification of facial expressions, and additionally present a prototype of an audio-visual laughter detection system. Schwenker et al. [77] have used GMM as a universal background model (UBM). Then, from GMM, the mean vectors have been extracted and concatenated to the so-called GMM supervectors which are then applied to a SVM classifier. Finally, Scherer et al. [78] have used multi-classifier and radial basis function (RBF) ensembles for emotion recognition.

As example researches for emotion recognition using classifier fusion: López-Cózar et al. [8] have used two modules that combine different information sources to enhance emotion detection. The first module (Fusion-0) combines emotion predictions generated by a set of classifiers. Using the output of Fusion-0, the second module (Fusion-1) acts as a posteriori processing stage to combine information. Wu and Liang [79] have adopted GMM, SVM, and MLP as the base-level classifiers. Then, a meta decision tree (MDT) has been employed for classifier fusion. In addition, semantic labels (SLs) have been used to automatically extract emotion association rules (EARs). Finally, a weighted product fusion method has been used to integrate the acoustic-prosodic (AP)-based and SL-based recognition results for the final emotion decision. Also, multiple emotion detectors have been fused into a single detection system in [80]. Scherer et al. [81] have proposed multi-classifier systems utilizing three types of features using two-classifier fusion techniques. These features were perceived loudness features, robust relative spectral transform (RASTA)-PLP features, and long-term modulation spectrum-based features. Finally, Pao et al. [82] have focused on combination schemes of multiple classifiers to achieve best possible emotion recognition rate. The investigated classifiers were KNN, weighted KNN (WKNN), weighted discrete nearest neighbor, and SVM. The classifier combination schemes were majority voting, minimum misclassification and maximum accuracy methods.

3 Emotional speech corpus

The proper preparation of an emotional speech database requires recording of emotional manifestations. However, real-life emotion data are hard to collect [59, 83]. The text of sentences of FARSDAT speech corpus [84] is used in our experiments. The FARSDAT is a continuous Farsi neutral speech corpus including 6,000 utterances from 300 speakers with various accents. Using 30 non-professional speakers, the emotional speech corpus has been recorded. The non-professional speakers were graduate students and speech samples were recorded in a quiet room. The speakers were also directed to keep the degree of expressiveness of each emotion almost constant. For this purpose, each speaker uttered 202 sentences in three emotional states: neutral (N), happiness (H), and anger (A).

The speakers have been amateur and uttered each sentence several times from the template corpus. The emotional sentences with better quality have been selected from the recorded sentences.

The base features for classifiers are 12 MFCCs, log energy (LE), the first three formant frequencies and pitch frequency. Each feature vector contains 39 components that are 12 MFCCs (c_1–c_{12}), LE, and their velocity (dc_1–dc_{12} and dLE) and their acceleration (ddc_1–ddc_{12} and ddLE) coefficients of the 13 mentioned features. Also, using three formant frequencies and pitch frequency, 16 supplementary features are calculated. These features contain pitch and first three formant frequencies (F_0, F_1, F_2, F_3), derivative and logarithm of them (dF_0, dF_1, dF_2, dF_3 and LF_0, LF_1, LF_2, LF_3), and their normalized (zero-mean) values (zF_0, zF_1, zF_2, zF_3) at each frame. To compute the zF_i, the mean value of F_i in each sentence is subtracted from the original value at each 25-ms frame.

The training dataset contains 3,880 utterances corresponding to 64% of the corpus and the test dataset includes 2,180 utterances corresponding to 36% of the corpus.

4 Feature selection algorithm and results

Extracting of a limited, meaningful, and informative set of features is an important step in automatic recognition of emotions [85]. The irrelevant features reduce the correct classification rates. So, the feature selection methods are used to reduce the size of feature set and also the computational load [86].

In this study, at the first step, we have chosen some important features using the analysis of variance (ANOVA) ranking scheme. ANOVA is a technique for analyzing experimental data in which one or more response variables are measured under various conditions identified by one or more classification variables. One-way ANOVA is a method for testing null hypotheses on equal means in several populations. Suppose that data are sampled from k different populations, and assume the model as follows:

$$Y_{ij} = \mu_i + \varepsilon_{ij}; \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k \tag{1}$$

where Y_{ij} is the jth observation from the ith population, μ_i is the mean of the ith population, and ε_{ij} denotes the random variation in Y_{ij} away from μ_i. It is assumed that ε_{ij}'s are independent, normally distributed random variables with zero-mean and variance σ^2. If there are two groups, i.e., $k = 2$, a test statistic for the hypothesis of equal variance H_0: $\sigma_1^2 = \sigma_2^2$, is the two-sample F test statistic:

$$F = \frac{s_1^2}{s_2^2} \tag{2}$$

in which s_i^2 is the unbiased estimator of the variance for the ith group. The F statistic has an $F(n_1 - 1, n_2 - 1)$-distribution.

The one-way ANOVA can only tell us whether all the means are equal, or whether there seems to be some difference in the means of different populations. In this section, we discuss about the methods for testing hypotheses

that compare the population means pair-wise. Such tests are known as multiple comparison tests.

Several methods have been suggested for conducting multiple comparison tests of means [87]. Three most usual methods are least significant difference method, Tukey method, and Scheffe method [88]. In this paper, we use Tukey method. It is noted that Scheffe method is a single-step multiple comparison procedure that is applied to the set of estimates of all possible contrasts among the factor level means, not just the pair-wise differences considered by Tukey method. So, if only pair-wise comparisons are to be made, the Tukey method will result in a narrower confidence limit, which is preferable [88].

The first step in Tukey method is to calculate the k sample means $\overline{Y_1}, \overline{Y_2}, \ldots, \overline{Y_k}$ and arrange them in increasing order. Then, we calculate all the pair-wise differences $\overline{Y_i} - \overline{Y_j}$ starting with the difference between the largest sample mean and the smallest one, then the largest and the next smallest, and so on. Finally, the difference between the largest and the second largest is calculated. Then, we calculate the difference between the second largest and the smallest one, the second largest and the next smallest, and so on, until all possible differences have been calculated. In order to test whether any of the differences is significantly different from zero, i.e., whether any pair of means is significantly different, we need a statistical test.

The mentioned test depends on whether the sample sizes of the k samples are equal or not. When all the samples have equal sample sizes, i.e., $n_i = n^*$; $i = 1, \ldots, k$, the test is given by:

$$Q = \frac{\overline{Y_i} - \overline{Y_j}}{\sqrt{s^2/n^*}} \tag{3}$$

in which s^2 is the unbiased estimator for the variance σ^2 and n^* is the common sample size of the populations.

We have used ANOVA and Tukey method in our simulations. In the following, the results for selected features that have different classification ability are shown.

By using the results of ANOVA and Tukey method, the classification ability of the mentioned 55 features in this study are determined. Table 2 shows the result of applying ANOVA and Tukey method for each feature. In Table 2, the value of F in ANOVA method is reported for each feature. Also, the separated groups of emotions using multiple comparison tests for each feature are shown in this table. In Table 2, "A" stands for angry state, "H" stands for happiness state, and "N" stands for neutral state.

For example, Table 2 shows the result of Tukey method for "dc_1" feature as "A-(H, N)", i.e., two separated groups can be observed: one group comprises "happiness" and "neutral" states, and another group comprises "angry" emotional state.

In our experiments, by using the results reported in Table 2, we have selected 43 features that do not result in (H, A, N) as the separated group. In other words, dc_2, dc_5, dLE, ddc_5, ddc_8, ddc_{10}, ddc_{11}, dF_1, dF_2, dF_3, dF_0, and zF_0 features are eliminated.

Also, the selected significant features for individual emotions, as six different feature subsets (FSSs), are listed in Table 3. As can be seen, FSS1 includes some features that can separate happiness/non-happiness (H/NH) emotions. Also, FSS2 and FSS3 can separate angry/non-angry (A/NA) and neutral/non-neutral (N/NN) emotions, respectively. These features are suitable for one-against-all (OAA) classifiers [89]. In the OAA approach, K binary classifiers are constructed, in which a classifier is constructed for each class. Thus, a classifier f_i is trained using the samples of class C_i against all the samples of the other classes [90].

In addition, FSS4, FSS5, and FSS6 can separate two groups of emotions: "happiness and neutral (H/N)," "happiness and angry (H/A)," and "neutral and angry (N/A)," respectively. These features are suitable for one-against-one (OAO) classifiers [89]. In the OAO approach, an independent binary classifier is built for each pair of classes. Thus, a classifier f_{ij} is trained using the samples of classes i and j, and hence this classifier is trained to discriminate between only these two classes [90].

Each emotion is associated with certain characteristics and has a specific set of attributes to distinguish it from others. The recognition performance can be improved by appropriate localization of these attributes.

5 Performance comparison of MLP, GMM, and C5.0 classifiers

To classify the extracted features into different human emotions, we need to select a classifier that can properly model the data and achieve better classification accuracy. Since we do not have any prior knowledge about the characteristics of the features, a comparison of popular classification algorithms in emotion recognition will be helpful. In this work, we compare the performance of GMM, MLP neural network, and C5.0 decision tree models using HTK [34] and Clementine [35] software tools.

It is noted that C5.0 algorithm works by splitting the sample based on the field that provides the maximum information gain. Each subsample defined by the first split is then split again, usually based on a different field, and the process repeats until the subsamples cannot be split any further. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned. Table 4 shows the accuracy of emotion recognition system for happiness, anger, and

Table 2 Result of ANOVA and Tukey-based multiple comparisons for each feature

Feature	Value of F in ANOVA method	Separated groups (Tukey method)[a]	Feature	Value of F in ANOVA method	Separated groups (Tukey method)[a]
c_1	14,982.77	N-A-H	ddc_3	261.55	A-(H, N)
c_2	5,159.59	N-A-H	ddc_4	68.07	N-A-H
c_3	9,270.47	N-A-H	ddc_5	41.01	(H, A, N)
c_4	962.93	N-A-H	ddc_6	36.65	H-(N, A)
c_5	40.76	N-(A, H)	ddc_7	44.66	A-(H, N)
c_6	4,956.55	N-A-H	ddc_8	50.04	(H, A, N)
c_7	357.89	N-A-H	ddc_9	6.34	(N, H)-(N, A)
c_8	187.45	A-(H, N)	ddc_{10}	1.08	(H, A, N)
c_9	644.73	N-A-H	ddc_{11}	6.95	(H, A, N)
c_{10}	418.43	N-A-H	ddc_{12}	13.23	A-(H, N)
c_{11}	2,524.23	N-A-H	ddLE	180.35	N-A-H
c_{12}	121.31	H-(N, A)	F_1	1,783.90	N-A-H
LE	17,110.54	N-A-H	F_2	948.70	N-A-H
dc_1	106.50	A-(H, N)	F_3	2,065.78	N-A-H
dc_2	2.48	(H, A, N)	LF_1	2,863.50	N-A-H
dc_3	14.89	A-(H, N)	LF_2	920.58	N-A-H
dc_4	4.28	(A, H)-(N, A)	LF_3	2,065.98	N-A-H
dc_5	5.59	(H, A, N)	dF_1	1.02	(H, A, N)
dc_6	63.33	N-A-H	dF_2	2.51	(H, A, N)
dc_7	91.75	N-A-H	dF_3	1.78	(H, A, N)
dc_8	11.35	(A, H)-(N, H)	zF_1	13.69	H-(N, A)
dc_9	7.90	(A, H)-(N, A)	zF_2	28.99	N-A-H
dc_{10}	4.67	(A, N)-(H, N)	zF_3	36.40	N-A-H
dc_{11}	22.68	A-(H, N)	F_0	70.71	A-(H, N)
dc_{12}	6.94	A-(H, N)	LF_0	56.55	N-A-H
dLE	0.14	(H, A, N)	dF_0	0.35	(H, A, N)
ddc_1	156.54	N-A-H	zF_0	0.42	(H, A, N)
ddc_2	195.48	N-A-H			

[a] Different emotion groups are separated by "-" and the same groups are shown inside parenthesis

Table 3 Selected feature subsets for individual OAA and OAO classifiers

Feature subset	Group of emotions	Selected features for individual OAA or OAO classifiers	Type of classifier
FSS1	H/NH	c_1, c_2, c_3, c_4, c_6, c_7, c_9, c_{10}, c_{11}, c_{12}, LE, dc_6, dc_7, ddc_1, ddc_2, ddc_4, ddc_6, ddLE, F_1, F_2, F_3, LF_1, LF_2, LF_3, LF_0, zF_1, zF_2, zF_3	OAA
FSS2	A/NA	c_1, c_2, c_3, c_4, c_6, c_7, c_8, c_9, c_{10}, c_{11}, LE, dc_1, dc_3, dc_6, dc_7, dc_{11}, dc_{12}, ddc_1, ddc_2, ddc_3, ddc_4, ddc_7, ddc_{12}, ddLE, F_1, F_2, F_3, F_0, LF_1, LF_2, LF_3, LF_0, zF_2, zF_3	OAA
FSS3	N/NN	c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_9, c_{10}, c_{11}, LE, dc_6, dc_7, ddc_1, ddc_2, ddc_4, ddLE, F_1, F_2, F_3, LF_1, LF_2, LF_3, LF_0, zF_2, zF_3	OAA
FSS4	H/N	c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_9, c_{10}, c_{11}, c_{12}, LE, dc_6, dc_7, dc_9, ddc_1, ddc_2, ddc_4, ddc_6, ddLE, F_1, F_2, F_3, LF_1, LF_2, LF_3, LF_0, zF_1, zF_2, zF_3	OAO
FSS5	H/A	c_1, c_2, c_3, c_4, c_6, c_7, c_8, c_9, c_{10}, c_{11}, c_{12}, LE, dc_1, dc_3, dc_4, dc_6, dc_7, dc_{10}, dc_{11}, dc_{12}, ddc_1, ddc_2, ddc_3, ddc_4, ddc_6, ddc_7, ddc_9, ddc_{12}, ddLE, F_1, F_2, F_3, F_0, LF_1, LF_2, LF_3, LF_0, zF_1, zF_2, zF_3	OAO
FSS6	N/A	c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_9, c_{10}, c_{11}, LE, dc_6, dc_7, ddc_1, ddc_2, ddc_4, ddLE, F_1, F_2, F_3, LF_1, LF_2, LF_3, LF_0, zF_2, zF_3	OAO

Table 4 Emotion recognition accuracy using different classifiers and 55 input features

Classifier	Emotion recognition accuracy (%)			Average accuracy (%)
	Happiness	Anger	Neutral	
MLP	69.3	85.0	50.7	68.3
GMM	71.6	73.3	52.7	65.9
C5.0	56.5	65.5	46.9	56.3

Table 5 Emotion recognition accuracy using MLP with 43 selected input features by ANOVA-Tukey method

Emotion recognition accuracy (%)			Average accuracy (%)
Happiness	Anger	Neutral	
83.4	65.4	62.8	70.5

neutral emotional states using 55 features and employing three mentioned algorithms.

The experimental results show that MLP performs better, so in the next experiment we use MLP. It is noted that the topology of double-hidden layer MLP in this experiment is considered as (3, 50, 52, 55). Then, the selected features by ANOVA method are used to train MLP neural net. Table 5 shows the emotion recognition accuracy using 43 selected features, as mentioned before as the discarded features. As can be seen, the recognition rate is improved about 2.2%. It is noted that the topology of double-hidden layer MLP in this experiment is considered as (3, 37, 39, 43).

Combining the classifiers is an approach to improve the performance of classification particularly for complex problems such as those involving a considerable amount of noise, limited number of training patterns, high-dimension feature sets, and highly overlapped classes.

6 Proposed modular neural-SVM model

From a computational complexity viewpoint, according to the principle of "divide and conquer," a complex computational task can be solved by dividing into a number of simple tasks and then combining the solutions of those tasks. In supervised learning, computational simplicity is achieved by distributing the learning task among a number of experts, which in turn divides the input space into a set of subspaces. In this work, the emotion recognition task is decomposed into a set of simpler emotion recognition subtasks.

The key goal of a multi-classifier system is to obtain a better composite global model, with more accurate and reliable estimates or decisions. One approach in the design of a multi-classifier system is to combine the outputs of individual classifiers, where each classifier solves the same classification problem. Each classifier may use different subsets of the training data and may use different feature extractors. The outputs of individual classifiers are combined through certain rules such as voting, averaging, and product rule [91].

Most of the methods that designed to solve multi-class pattern classification problem are based on splitting the *K*-class classification problem into a number of smaller two-class sub-problems [90]. For each sub-problem, an independent binary classifier is built. Then, the results of binary classifiers are combined to get the classification result. Several techniques have been proposed for decomposing the multi-class problem, including the two popular approaches: OAA and OAO.

Also, classifiers can be combined in series and/or in parallel. Empirical methods can become extremely time-consuming,

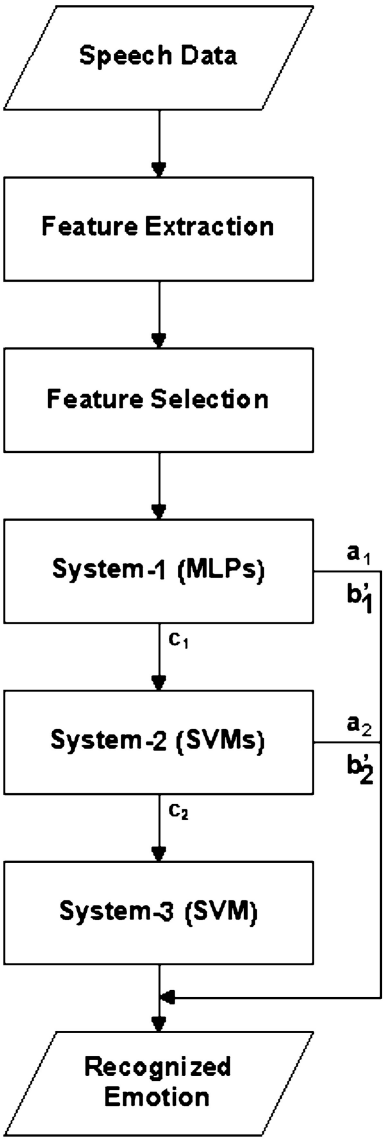

Fig. 1 Proposed modular neural-SVM multi-classifier scheme for emotion recognition

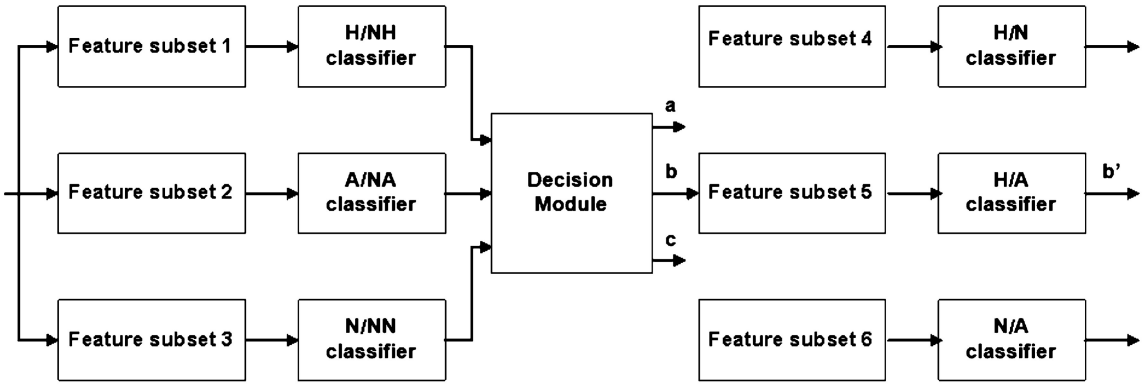

Fig. 2 Detailed structure of System-1 (or System-2)

given the very large number of combination possibilities. We have developed a method of systematically achieving the better architecture for combination of classifiers that can include both parallel and serial methods.

In this paper, a multi-classifier scheme is proposed involving the analysis of individual class and combination of different classes. This multi-classifier includes both parallel and serial methods and also uses the benefits of OAA and OAO classifiers.

In the proposed model, three systems are used in serial mode. Different learning machines are used in this system. In System-1, we use MLPs as experts, and in System-2, we use SVMs as experts. System-3 is a global emotion recognition system using SVM classifier that recognizes three emotional states. It is noted that the SVM is a robust classification and regression technique that maximizes the prediction accuracy of a model without overfitting the training data. SVM is particularly suited to analyzing data with very large numbers (for example, thousands) of predictor fields. The architecture of proposed multi-classifier scheme is shown in Fig. 1. The details of the System-1 (or System-2) are shown in Fig. 2.

Using different feature sets or rule sets is one of the methods for constructing diverse classifiers. This method is useful particularly when more than one source of information is available [92]. In the proposed model, we use six different feature subsets, introduced in Table 3.

As MLP has shown its effectiveness in the previous experiments, the individual classifiers in System-1 are based on MLP. We built three OAA classifiers first, which are represented as "H/NH, A/NA, and N/NN" separately in Fig. 2. The OAA classifiers are designed specifically for individual emotions that each of them performs a 2-class pattern recognition problem. In the training phase, for each OAA classifier, we label all the samples that do not belong to the corresponding emotion as one class. The output of these OAA classifiers is the probability of belonging to the corresponding emotion. For example, in the OAA classifier for "A/NA," all the samples of anger emotion are labeled as "A" while all the other samples are labeled as "NA." The output of each OAA classifier is taken as the input to a decision module for further classification.

Table 6 Functional details of decision module shown in Fig. 2

Input of decision module	Output of decision module	Recognized emotion
H NA NN	a	H
NH A NN	a	A
NH NA NN	a	N
H A NN	b	H/A
NH A N	b	N/A
H NA N	b	H/N
NH NA NN	c	Determined by the next system
H A N	c	Determined by the next system

The decision module works based on the following "a," "b," and "c" rules (as shown at the output of decision module in Fig. 2 and detailed in Table 6):

Table 7 Topology of MLPs in modular neural-SVM model

MLP model in the proposed modular scheme	Number of input layer nodes	Number of first hidden layer nodes	Number of second hidden layer nodes	Number of output layer nodes
H/NH classifier	28	39	17	1
A/NA classifier	34	39	26	1
N/NN classifier	26	28	26	1
H/N classifier	30	7	–	1
H/A classifier	40	12	–	1
N/A classifier	26	28	–	1

Table 8 Confusion matrix of emotion recognition using proposed modular model

Actual	Predicted		
	Anger	Happiness	Neutral
Anger	88.2	6.5	5.3
Happiness	7.2	78.2	14.6
Neutral	6.3	31.1	62.6

Table 9 Confusion matrix of emotion recognition using MLP with 55 input features

Actual	Predicted		
	Anger	Happiness	Neutral
Anger	85.0	9.5	5.5
Happiness	11.1	69.3	19.6
Neutral	8.3	41.0	50.7

(a) If one of the outputs of these OAA classifiers is assigned to a special class (for example, H, NA, and NN), we label the sample as the corresponding class (in this example, happiness emotion or H).
(b) If two of the outputs of these OAA classifiers are assigned to special classes (for example, H, A, and NN), we use one of OAO classifiers for this sample (in this example, H/A).
(c) If none of the outputs of these OAA classifiers or all of them are classified correctly (for example, H, A, and N), the sample will be classified by the next system (as depicted in Fig. 1).

7 Experimental results of proposed model

In our simulations, several MLPs with different topologies, in terms of hidden layer(s) structure, are trained and the model with the lowest root mean square (RMS) error is selected as the final model. The best topology for six MLPs depicted in Fig. 2 is reported in Table 7.

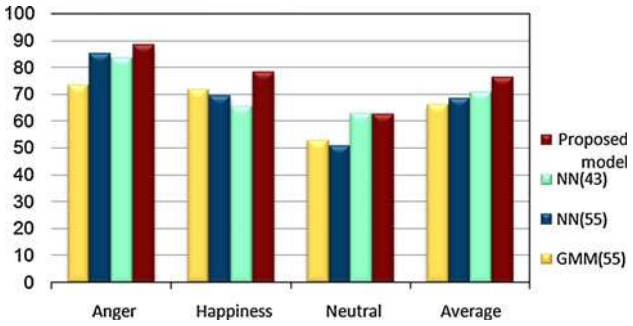

Fig. 3 Emotion recognition rate comparison of the proposed modular model with monolithic NN- and GMM-based simulated classifiers (The number in parenthesis shows the number of input features to the classifier)

It is noted that Clementine software is used in our simulations of MLP, SVM, and C5.0 algorithm [35]. Momentum rate is set to 0.9 in simulation of MLPs in this work. In simulation of SVMs, radial basis function (RBF) kernel type is used and regularization parameter (C) is set to 10.

Table 8 shows the confusion matrix of the proposed model. The average accuracy of this classifier is 76.33%. As can be seen, by performing individual class-based analysis, the recognition rate improves significantly.

Table 9 shows the confusion matrix of the emotion recognition using MLP neural net with 55 input features. As can be seen, some pairs of emotions are usually confused more. For example, neutral emotion is misclassified as happiness state (by 41.0%) and vice versa (by 19.6%). By using the proposed model, this problem is somewhat solved and confusion between some classes becomes less. For example, neutral emotion is misclassified as happiness state (by 31.1%) and vice versa (by 14.6%).

The recognition results using different feature sets and the proposed multi-classifier are depicted in Fig. 3. In Fig. 3, GMM(55) and NN(55) stand for GMM and MLP classifiers using 55 input features, respectively. NN(43) represents MLP using 43 selected input features, and the proposed model represents the multi-classifier neural-SVM system.

Experimental results show that the proposed method improves recognition rate up to 8% when using the MLP-based system and up to 10.4% when using the GMM with 32 mixtures.

Table 10 Performance comparison of proposed modular emotion recognition system and some similar researches

Emotional states	Selected features	Feature selection method(s)	Classifier(s)	Recognition rate (%)
Anger, disgust, fear, happiness, neutral, sadness, surprise [54]	Pitch, energy, duration, MFCCs	PCA, LDA	MLB	53[a]
Happiness, anger, sadness, neutral [93]	Pitch and its first derivative, formants, MFCCs	No	SVM, ANN	71, 42
Happiness, anger, tiredness, sadness, neutral [94]	Pitch, log energy, formants, MFCCs and their first and second derivatives	No	GSVM[b]	41
Happiness, anger, anxiety, fear, tiredness, disgust, neutral [95]	MFCCs, energy, dc_i, dE, ddc_i, ddE	No	GMVAR[c], ANN, HMM	76, 55, 71
Anger, happiness, neutral, sadness, surprise [52]	Formants, pitch, energy, spectral features	SFFS	MLB	53.7 (DES Database)
				57.2 (SUSAS Database)
Happiness, anger, sadness, fear, neutral [56]	Pitch, speaking rate, formants and their bandwidths	Instance-base learning	KNN	70
Happiness, anger, neutral, interrogative [32]	MFCCs, LE and their first and second derivatives, formant- and pitch-related features	FCBF[d]	GMM (32 mixtures)	65.1
Anger, fear, surprise, disgust, joy, sadness [62]	V/UV[e], energy, pitch, VAD[f]	No	GMM (512 mixtures)	92.3
Neutral, emphatic, negative [39]	Pitch, MFCCs	No	GMM (512 mixtures)	93
Happiness, anger, neutral (simulated in this study)	MFCCs, LE and their first and second derivatives, formant- and pitch-related features	No	GMM (32 mixtures)	65.9
Happiness, anger, neutral (simulated in this study)	MFCCs, LE and their first and second derivatives, formant- and pitch-related features	No	C5.0	56.3
Happiness, anger, neutral (simulated in this study)	MFCCs, LE and their first and second derivatives, formant- and pitch-related features	No	MLP	68.3
Happiness, anger, neutral (simulated in this study)	MFCCs, LE and their first and second derivatives, formant- and pitch-related features	ANOVA-Tukey	MLP	70.5
Happiness, anger, neutral (proposed in this study)	MFCCs, LE and their first and second derivatives, formant- and pitch-related features	ANOVA-Tukey	Modular neural-SVM model	76.3

[a] The maximum emotion recognition rate is reported

[b] Gaussian SVM

[c] Gaussian mixture vector autoregressive model

[d] Fast correlation-based filter

[e] Voiced/unvoiced

[f] Voice activity detection

This multi-classifier scheme takes advantage of the analysis of significant features in an individual class and uses such to distinguish combinations of classes. It helps to obtain more detailed insight into individual emotion and the way to separate specific emotions. This "divide and conquer" method partitions classification into finer analysis and is a popular practice in pattern recognition problems that is proposed in the modular neural-SVM scheme in this paper.

The proposed multi-classifier scheme produces noticeable improvement in individual class recognition accuracy and achieves the best overall recognition rate of 76.33% among the simulated monolithic classifiers. The performance of the proposed system has been compared with some other emotion recognition systems (Table 10).

The accuracy of the proposed modular neural-SVM system is reported in the last row of Table 10. Because of the different target emotional states and also feature sets in each research, selection of the most effective approach is impossible. However, as can be seen the performance of proposed model is superior to the most of reported systems except the systems reported in [39, 62] that have very high computational load in the training phase as one of their disadvantages.

8 Conclusion

Emotion recognition is an important step toward implementing an emotional speech recognition system. The type

and number of emotional states, extracted features, feature selection algorithm, and type of the classifier are important factors in the accuracy of emotion recognition systems. In this paper, a modular neural-SVM classifier has been proposed for recognition of three emotional states. The MFCCs, log energy and their velocity and acceleration coefficients have been used as the base features. Also, 16 supplementary formant- and pitch-related features have been considered. The combination of ANOVA and Tukey methods has been used for feature selection and constructing a modular multi-classifier scheme. The performance of proposed model has been compared to GMM, MLP, and C5.0 monolithic classifiers. Experimental results have shown that the proposed neural-SVM classifier can improve the recognition accuracy at least by 8% as compared to the simulated monolithic classifiers.

References

1. Bosch L (2003) Emotions, speech and the ASR framework. Speech Commun 40:213–225
2. Albornoz EM, Milone DH, Rufiner HL (2011) Spoken emotion recognition using hierarchical classifiers. Comput Speech Lang 25:556–570
3. Ai H, Litman DJ, Forbes-Riley K, Rotaru M, Tetreault J, Purandare A (2006) Using system and user performance features to improve emotion detection in spoken tutoring systems. In: The proceedings of Interspeech, pp 797–800
4. Devillers L, Vidrascu L (2006) Real-life emotions detection with lexical and paralinguistic cues on human–human call center dialogs. In: The proceedings of Interspeech, pp 801–804
5. Lee CC, Mower E, Busso C, Lee S, Narayanan S (2009) Emotion recognition using a hierarchical binary decision tree approach. In: The proceedings of Interspeech, pp 320–323
6. Polzehl T, Sundaram S, Ketabdar H, Wagner M, Metze F (2009) Emotion classification in children's speech using fusion of acoustic and linguistic features. In: The proceedings of Interspeech, pp 340–343
7. Klein J, Moon Y, Picard RW (2002) This computer responds to user frustration: theory, design and results. Interact Comput 14:119–140
8. López-Cózar R, Silovsky J, Kroul M (2011) Enhancement of emotion detection in spoken dialogue systems by combining several information sources. Speech Commun 53:1210–1228
9. Fernandez R, Picard R (2011) Recognizing affect from speech prosody using hierarchical graphical models. Speech Commun 53:1088–1103
10. Oudeyer PY (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum Comput Interact Stud 59:157–183
11. Huber R, Batliner A, Buckow J, Nöth E, Warnke V, Niemann H (2000) Recognition of emotion in a realistic dialogue scenario. In: The proceedings of international conference on spoken language processing, pp 665–668
12. Yacoub S, Simske S, Lin X, Burns J (2003) Recognition of emotions in interactive voice response systems. In: The proceeding of European conference on speech communication and technology, pp 729–732
13. Polzehl T, Schmitt A, Metze F, Wagner M (2011) Anger recognition in speech using acoustic and linguistic cues. Speech Commun 53:1198–1209
14. Lee CM, Narayanan S (2003) Emotion recognition using a data-driven fuzzy inference system. In: The proceedings of Eurospeech, pp 157–160
15. Litman DJ, Forbes-Riley K (2006) Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Commun 48:559–590
16. Batliner A, Fischer K, Huber R, Spilker J, Nöth E (2003) How to find trouble in communication. Speech Commun 40:117–143
17. Ang J, Dhillon R, Krupski A, Shriberg E, Stolcke A (2002) Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: The proceedings of international conference on spoken language processing, pp 2037–2039
18. Liscombe J, Hirschberg J, Venditti JJ (2005) Detecting certainness in spoken tutorial dialogues. In: The proceeding of European conference on speech communication and technology, pp 1837–1840
19. Womack BD, Hansen JHL (1996) Classification of speech under stress using target driven features. Speech Commun 20:131–150
20. Gharavian D, Ahadi SM (2008) Stressed speech recognition using a warped frequency scale. IEICE Electron Express 5:187–191
21. Laukka P, Neiberg D, Forsell M, Karlsson I, Elenius K (2011) Expression of affect in spontaneous speech: acoustic correlates and automatic detection of irritation and resignation. Comput Speech Lang 25:84–104
22. Tolkmitt FJ, Scherer KR (1986) Effect of experimentally induced stress on vocal parameters. J Exp Psychol Hum Percept Perform 12:302–313
23. Cairns D, Hansen JHL (1994) Nonlinear analysis and detection of speech under stressed conditions. J Acoust Soc Am 96:3392–3400
24. Dellaert F, Polzin T, Waibel A (1996) Recognizing emotion in speech. In: The proceedings of international conference on spoken language processing, vol 3, pp 1970–1973
25. Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13:293–303
26. Gharavian D, Ahadi SM (2005) The effect of emotion on Farsi speech parameters: a statistical evaluation. In: The proceedings of international conference on speech and computer, pp 463–466
27. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Commun 48:1162–1181
28. Shami M, Verhelst W (2007) An evaluation of the robustness of existing supervised machine learning approaches to the classifications of emotions in speech. Speech Commun 49:201–212
29. Altun H, Polat G (2009) Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection. Expert Syst Appl 36:8197–8203
30. Gharavian D, Sheikhan M, Nazerieh AR, Garoucy S (2011) Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput Appl (published online 27 May 2011). doi:10.1007/s00521-011-0643-1
31. Sheikhan M, Safdarkhani MK, Gharavian D (2011) Emotion recognition of speech using small-size selected feature set and ANN-based classifiers: a comparative study. World Appl Sci J 14:616–625
32. Gharavian D, Sheikhan M, Pezhmanpour M (2011) GMM-based emotion recognition in Farsi language using feature selection algorithms. World Appl Sci J 14:626–638
33. Fersini E, Messina E, Archetti F (2012) Emotional states in judicial courtrooms: an experimental investigation. Speech Commun 54:11–22
34. Young SJ, Evermann G, Kershaw D, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P (2002) The HTK book (Ver. 3.2). Cambridge University Press, Cambridge

35. SPSS Inc. (2007) Clementine® 12.0 algorithms guide. Integral Solutions Limited, Chicago
36. Freedman DA (2005) Statistical models: theory and practice. Cambridge University Press, Cambridge
37. Rong J, Li G, Chen YP (2009) Acoustic feature selection for automatic emotion recognition from speech. Info Process Manage 45:315–328
38. Kao Y, Lee L (2006) Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In: The proceedings of international conference on spoken language processing, pp 1814–1817
39. Neiberg D, Elenius K, Laskowski K (2006) Emotion recognition in spontaneous speech using GMMs. In: The proceedings of international conference on spoken language processing, pp 809–812
40. Pao T, Chen Y, Yeh J, Chang Y (2008) Emotion recognition and evaluation of Mandarin speech using weighted D-KNN classification. Int J Innov Comput Info Control 4:1695–1709
41. Sidorova J (2009) Speech emotion recognition with TGI+.2 classifier. In: The proceedings of the EACL student research workshop, pp 54–60
42. Gajšek R, Štruc V, Mihelič F (2010) Multi-modal emotion recognition using canonical correlations and acoustic features. In: The proceedings of international conference on pattern recognition, pp 4133–4136
43. Yang B, Lugger M (2010) Emotion recognition from speech signals using new harmony features. Signal Process 90:1415–1423
44. Bitouk D, Verma R, Nenkova A (2010) Class-level spectral features for emotion recognition. Speech Commun 52:613–625
45. Yeh J, Pao T, Lin C, Tsai Y, Chen Y (2010) Segment-based emotion recognition from continuous Mandarin Chinese speech. Comput Hum Behav 27:1545–1552
46. Wu S, Falk TH, Chan WY (2011) Automatic speech emotion recognition using modulation spectral features. Speech Commun 53:768–785
47. He L, Lech M, Maddage NC, Allen NB (2011) Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. Biomed Signal Process Control 6:139–146
48. Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97:273–324
49. Hyvarinen A (1999) Survey of independent component analysis. Neural Comput Surv 2:94–128
50. Talavera L (1999) Feature selection as a preprocessing step for hierarchical clustering. In: The proceedings of international conference on machine learning, pp 389–397
51. Liu H, Motoda H, Yu L (2002) Feature selection with selective sampling. In: The proceedings of international conference on machine learning, pp 395–402
52. Ververidis D, Kotropoulos C (2006) Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections. In: The proceedings of European signal processing conference, pp 1–5
53. Batliner A, Steidl S, Schuller B, Seppi D, Vogt T, Wagner J, Devillers L, Vidrascu L, Aharonson V, Kessous L, Amir N (2011) Whodunnit-Searching for the most important feature types signalling emotion-related user states in speech. Comput Speech Lang 25:4–28
54. Haq S, Jackson PJB, Edge J (2008) Audio-visual feature selection and reduction for emotion classification. In: The proceedings of international conference on auditory-visual speech processing, pp 185–190
55. Pérez-Espinosa H, Reyes-García CA, Villaseñor-Pineda L (2011) Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model. Biomed Signal Process Control (published online 3 April 2011). doi:10.1016/j.bspc.2011.02.008
56. Petrushin VA (2000) Emotion recognition in speech signal: experimental study, development, and application. In: The proceedings of the international conference on spoken language processing, pp 222–225
57. Väyrynen E, Toivanen J, Seppänen T (2011) Classification of emotion in spoken Finnish using vowel-length segments: increasing reliability with a fusion technique. Speech Commun 53:269–282
58. Iliev AI, Scordilis MS, Papa JP, Falcão AX (2010) Spoken emotion recognition through optimum-path forest classification using glottal features. Comput Speech Lang 24:445–460
59. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit 44:572–587
60. Nwe TL, Foo SV, De Silva LC (2003) Speech emotion recognition using hidden Markov models. Speech Commun 41:603–623
61. Schuller B, Rigoll G, Lang M (2003) Hidden Markov model-based speech emotion recognition. In: The proceedings of the international conference on acoustics, speech, and signal processing, vol 2, pp 1–4
62. Luengo I, Navas E, Hernáez I, Sanchez J (2005) Automatic emotion recognition using prosodic parameters. In: The proceeding of Interspeech, pp 493–496
63. Kockmann M, Burget L, Černocky JH (2011) Application of speaker- and language identification state-of-the-art techniques for emotion recognition. Speech Commun (article in press). doi:10.1016/j.specom.2011.01.007
64. Schuller B, Rigoll G, Lang M (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: The proceedings of the international conference on acoustics, speech, and signal processing, vol 1, pp 577–580
65. Chuang ZJ, Wu CH (2004) Emotion recognition using acoustic features and textual content. In: The proceedings of the international conference on multimedia and expo, vol 1, pp 53–56
66. Hoch S, Althoff F, McGlaun G, Rigooll G (2005) Bimodal fusion of emotional data in an automotive environment. In: The proceedings of the international conference on acoustics, speech, and signal processing, vol 2, pp 1085–1088
67. Morrison D, Wang R, de Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centers. Speech Commun 49:98–112
68. Chandaka S, Chatterjee A, Munshi S (2009) Support vector machines employing cross-correlation for emotional speech recognition. Measurement 42:611–618
69. Wang F, Verhelst W, Sahli H (2011) Relevance vector machine based speech emotion recognition. Lecture Notes in Computer Science. Affect Comput Intell Interact 6975:111–120
70. Nicholson J, Takahashi K, Nakatsu R (1999) Emotion recognition in speech using neural networks. In: The proceedings of the international conference on neural information processing, vol 2, pp 495–501
71. Lee CM, Narayanan S, Pieraccini R (2002) Combining acoustic and language information for emotion recognition. In: The proceedings of the international conference on spoken language processing, pp 873-876
72. Park CH, Lee DW, Sim KB (2002) Emotion recognition of speech based on RNN. In: The proceedings of the international conference on machine learning and cybernetics, vol 4, pp 2210–2213
73. Caridakis G, Karpouzis K, Kollias S (2008) User and context adaptive neural networks for emotion recognition. Neurocomputing 71:2553–2562

74. Planet S, Iriondo I, Socor′o J, Monzo C, Adell J (2009) GTMURL contribution to the INTERSPEECH 2009 emotion challenge. In: The proceedings of 10th annual of the international speech communication association (Interspeech'09), pp 316–319
75. Lee CC, Mower E, Busso C, Lee S, Narayanan S (2011) Emotion recognition using a hierarchical binary decision tree approach. Speech Commun 53:1162–1171
76. Schwenker F, Scherer S, Schmidt M, Schels M, Glodek M (2010) Multiple classifier systems for the recognition of human emotions. Lecture Notes in Computer Science. Multiple Classif Syst 5997:315–324
77. Schwenker F, Scherer S, Magdi YM, Palm G (2009) The GMM-SVM supervector approach for the recognition of the emotional status from speech. Lecture Notes in Computer Science. Artif Neural Netw 5768:894–903
78. Scherer S, Schwenker F, Palm G (2008) Emotion recognition from speech using multi-classifier systems and RBF-ensembles. Studies in Computational Intelligence. Speech, audio, image and biomedical signal processing using neural networks, vol 83, pp 49–70
79. Wu CH, Liang WB (2011) Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans Affect Comput 2:10–21
80. Lefter I, Rothkrantz LJM, Wiggers P, van Leeuwen DA (2010) Emotion recognition from speech by combining databases and fusion of classifiers. Lecture Notes in Computer Science. Text Speech Dialogue 6231:353–360
81. Scherer S, Schwenker F, Palm G (2009) Classifier fusion for emotion recognition from speech. In: Advanced intelligent environments, pp 95–117
82. Pao TL, Chien CS, Chen YT, Yeh JH, Cheng YM, Liao WY (2007) Combination of multiple classifiers for improving emotion recognition in Mandarin speech. In: The proceedings of the international conference on intelligent information hiding and multimedia signal processing, vol 1, pp 35–38
83. Clavel C, Vasilescu I, Devillers L (2011) Fiction support for realistic portrayals of fear-type emotional manifestations. Comput Speech Lang 25:63–83
84. Bijankhan M, Sheikhzadegan J, Roohani MR, Samareh Y, Lucas C, Tebiani M (1994) The speech database of Farsi spoken language. In: The proceedings of the international conference on speech science and technology, pp 826–831
85. Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun (article in press). doi:10.1016/j.specom.2011.01.011
86. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
87. Hochberg Y, Tamhane AC (1987) Multiple comparison procedures. Wiley, New York
88. NIST/SEMATECH (2011) e-Handbook of statistical methods. (http://www.itl.nist.gov/div898/handbook/)
89. Hastie T, Tibshirani R (1998) Classification by pairwise coupling. Ann Stat 26:451–471
90. Ghanem AS, Venkatesh S, West G (2010) Multi-class pattern classification in imbalanced data. In: The proceedings of the international conference on pattern recognition, pp 2881–2884
91. Wang Y, Guan L (2005) Recognizing human emotion from audiovisual information. In: The proceedings of the international conference on acoustics, speech, and signal processing, pp 1125–1128
92. Kittler J, Hojjatoleslami A, Windeatt T (1997) Weighting factors in multiple expert fusion. In: The proceedings of the British machine vision conference, pp 42–50
93. Yu F, Chang E, Xu Y, Shum H (2001) Emotion detection from speech to enrich multimedia content. In: The proceedings of the IEEE Pacific Rim conference on multimedia: advances in multimedia information processing, pp 550–557
94. Kwon OW, Chan K, Hao J, Lee TW (2003) Emotion recognition by speech signal. In: The proceedings of the European conference on speech communication and technology, pp 125–128
95. Ayadi M, Kamel S, Karray F (2007) Speech emotion recognition using Gaussian mixture vector autoregressive models. In: The proceedings of the international conference on acoustics, speech, and signal processing, vol 5, pp 957–960