Neural Comput & Applic
DOI 10.1007/s00521-011-0643-1

ORIGINAL ARTICLE

Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network

Davood Gharavian · Mansour Sheikhan · Alireza Nazerieh · Sahar Garoucy

Received: 15 January 2011 / Accepted: 6 May 2011

Abstract Emotion recognition from speech has noticeable applications in the speech-processing systems. In this paper, the effect of using a rich set of features including formant frequency related, pitch frequency related, energy, and the two first mel-frequency cepstral coefficients (MFCCs) on improving the performance of speech emotion recognition systems is investigated. To do this, the different sets of features are employed, and by using the fast correlation-based filter (FCBF) feature selection method, some efficient feature subsets are determined. Finally, to recognize the emotion from speech, fuzzy ARTMAP neural network (FAMNN) architecture is used. Also, the genetic algorithm (GA) is employed to determine optimum values of the choice parameter (α), the vigilance parameters (ρ_{a}, ρ_{b}, and ρ_{ab}), and the learning rate (β) of FAMNN. Experimental results show the improvement in emotion recognition rate of angry, happiness, and neutral states by using a subset of 25 selected features and the GA-optimized FAMNN-based emotion recognizer.

D. Gharavian (✉) · M. Sheikhan · A. Nazerieh · S. Garoucy
EE Department, Islamic Azad University, South Tehran Branch, Tehran, Iran
e-mail: gharavian@pwut.ac.ir

M. Sheikhan
e-mail: msheikhn@azad.ac.ir

A. Nazerieh
e-mail: st_a_nazerieh@azad.ac.ir

S. Garoucy
e-mail: st_s_garoucy@azad.ac.ir

D. Gharavian
EE Department, Shahid Abbaspour University, Tehran, Iran

Keywords Emotion recognition · Speech · Feature selection · Fuzzy ARTMAP · Genetic algorithm

1 Introduction

Some factors such as the gender of speaker, dialect, age, language, emotion, and stress can influence the speech [1]. All of the mentioned factors give additional information to the listener. In other words, there are two channels in speech communication: the explicit channel carrying the linguistic content and the implicit channel containing the paralinguistic information about the speaker. Automatic speech recognition (ASR) systems have been attracted enormous efforts to extract the linguistic information in the recent four decades. However, much research is needed to reliably decode the implicit channel [2].

In the recent years, the emotion recognition from speech has noticeable applications in the speech-processing systems, such as spoken tutoring systems [3], medical emergency domain to detect stress and pain [4], interactions with robots [5, 6], computer games [7], and call centers [8]. Extracting the emotion from the short utterances of interactive voice response (IVR) systems is another typical application [9]. Also, emotion recognition is the first step toward the implementation of emotional speech recognition systems [10, 11]. Developing the human–computer interfaces for helping weak and old people is another application of the emotion recognition from speech [12].

There are many emotional states such as anger, hate, fear, happiness, sad, calm, and boredom. Each person transfers his emotional states to others through his face, body movements, or sufficient changes in his neutral speech. The facial motion and the tone of speech play major roles in expressing these emotions. The muscles of

Published online: 28 May 2011

face can be changed, and the tone and energy in the production of speech can be intentionally modified to communicate the different feelings.

In many studies, a discrete number of emotions has been considered, e.g., ‘Anger’ and ‘Neutral’ [9, 13], ‘Negative’ and ‘Non-negative’ [14, 15], ‘Neutral’, ‘Annoyed’, and ‘Frustrated’ [16], ‘Emotional’ and ‘Neutral’ [17], ‘Uncertain’, ‘Certain’, ‘Mixed’, and ‘Neutral’ [3], and ‘Anger’, ‘Fear’, ‘Relief’, and ‘Sadness’ [4].

Indeed, certain emotional states are often correlated with particular physiological states, which in turn have quite mechanical and thus predictable effects on speech, especially on the pitch frequency (F_0), timing, and voice quality [9]. On the other hand, using the emotion in speech often leads to reduction in speech recognition rate [1, 18].

The basics of most existing researches on emotion recognition can be summarized in the functional block diagram shown in Fig. 1. As can be seen, the emotion recognition has three stages: feature extraction, feature selection, and emotion recognition.

It is noted that the feature extraction is a critical functional block in the emotion recognition system [19]. Nowadays, most of the researches have been focused on finding reliable informative and independent features and combining powerful classifiers that improve the performance of emotion recognition systems in real-life applications [2, 5, 8, 20–26]. Some factors such as the number and gender of speakers, dialect, age, language, and skills are the effective factors of emotion recognition accuracy.

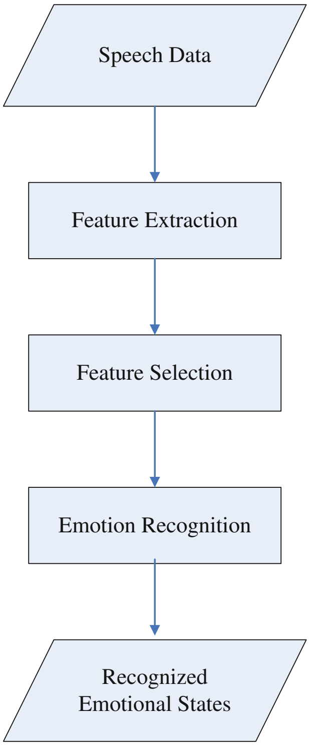

Fig. 1 Functional block diagram of speech emotion recognition system

In several studies, the prosodic features that are derived from parameters like F_0, loudness, energy contours, and speaking rate have been used [27]. Good performance has also been achieved by using short-term acoustic features, such as mel-frequency cepstral coefficients (MFCCs), logarithm frequency power coefficients (LFPCs) [28], and modulation spectral features (MSFs) [26].

In the recent decades, most of the emotion recognition models have been based on the maximum likelihood Bayes (MLB) [29–32] and linear discriminate classification (LDC) [29, 33]. However, artificial neural networks (ANNs) [9, 21, 25, 34–38], support vector machines (SVMs) [9, 21, 23–25, 38–43], decision trees [3, 5, 9, 44], K-nearest neighbor (KNN) [9, 38, 45–47], Gaussian mixture models (GMMs) [25, 27, 33, 48], hidden Markov models (HMMs) [21, 25, 28, 38, 49–52], and Bayesian networks [53] have also been used for emotion recognition. Also, in some researches, the hybrid and ensemble methods have been used, e.g., an ensemble of ANNs [45].

Extracting a limited, meaningful, and informative set of features is an important step in automatic recognition of emotions [54]. The irrelevant features reduce the correct classification rates. So, feature selection methods are used to reduce the size of feature set and also the computational load [55]. The filter and wrapper methods are two common techniques for feature selection. The filter method employs intrinsic properties of data, such as mutual information, as the criterion for feature subset evaluation. But the selection of feature subsets in the wrapper method relies on the performance of the classifier.

By considering the features at different levels such as frame level, syllable level and word level, the feature selection methods have been widely used in the emotion recognition researches [2, 24, 38, 42, 44, 54]. Some of the usual feature selection methods that have been used in emotion recognition systems are as follows: sequential floating forward selection (SFFS) [31], wrapper approach with forward selection [56], forward feature selection (FFS) and backward feature selection (BFS) [46], principal component analysis (PCA), or linear discriminate analysis (LDA) [32].

Using the modern emotion classification techniques, accuracies of about 80% have been achieved for acted non-spontaneous speech. But the accuracies drop significantly (below 40%) for non-acted spontaneous emotional speech [57].

The effect of using a rich set of features including formant frequency related, pitch frequency related, energy, and the two first MFCCs on improving the performance of emotion recognition systems is investigated in this paper. To do this, the different sets of features are employed, and by using fast correlation-based filter (FCBF) feature selection method, the efficient feature subsets are determined. Finally, to recognize the emotion of speech, fuzzy ARTMAP neural network (FAMNN) architecture is used. To determine the optimum values of the choice parameter (α), the vigilance parameters (ρ_a, ρ_b, and ρ_{ab}), and learning rate (β) of FAMNN, the genetic algorithm (GA) is employed.

The rest of paper is organized as follows: The candidate feature sets are introduced in Sect. 2. The emotional speech corpus and the emotion recognition results by employing 21 different feature sets, when using FAMNN as the neural classifier, are reported in Sect. 3. In Sect. 4, the FCBF feature selection method and the influence of feature selection on the performance of emotion recognition system are presented. In Sect. 5, the influence of optimizing FAMNN parameters, by using GA, on the emotion recognition rates is investigated. Finally, Sect. 6 concludes the main advantages of the proposed method.

2 Candidate feature sets for emotion recognition

In order to create a rich set of features, the MFCCs, energy, pitch frequency (F_0), and three formant frequencies (F_1–F_3) have been used as the basic features. In Table 1, 52 features are listed, which have been used in this study for emotion recognition. These features consist of the mean, variance, average of derivative, mean of variations of derivative, maximum, minimum, and the difference between maximum and minimum of the basic features. According to the large number of features at the frame level, the averaged values of features over a sentence have been used for training and testing of FAMNN-based classifier.

Selection of the best feature set is highly dependent on the emotional state. To obtain the best set of features, the different combinations of features are examined. On the other hand, a network is trained separately for each set and then by using the test dataset (detailed in Sect. 3), the emotion is recognized. The different combinations of features, based on Table 1, are obtained and demonstrated in Table 2. For selecting the best set of features, the FCBF feature selection method (detailed in Sect. 4) is used in this study.

3 Emotional speech corpus and base emotion recognition results

The development of emotional recognition systems requires recording of emotional manifestations. However, real-life emotion data are hard to collect [58]. The text of sentences of FARSDAT speech corpus [59] is used in our experiments. This is a continuous Farsi 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 number of simulated emotional sentences in the corpus is as follows: 34 emotionally anger sentences, 69 emotionally happiness sentences, and 99

Table 1 List of the candidate features used for emotion recognition

Feature number	Base feature(s)	Supplementary features
1–6	Pitch frequency (F_0)	Variance (VF_0), Mean (MF_0), Mean of derivative (MdF_0), Maximum (MaF_0), Minimum (MiF_0), Difference between maximum and minimum (DMF_0)
7–21	Formant frequencies (F_1–F_3)	Variance (VF_1–VF_3), Mean (MF_1–MF_3), Maximum (MaF_1–MaF_3), Minimum (MiF_1–MiF_3), Difference between maximum and minimum (DMF_1–DMF_3)
22–36	Derivative of formant frequencies (dF_1–dF_3)	Variance (VdF_1–VdF_3), Mean (MdF_1–MdF_3), Maximum ($MadF_1$–$MadF_3$), Minimum ($MidF_1$–$MidF_3$), Difference between maximum and minimum ($DMdF_1$–$DMdF_3$)
37–42	Logarithm of energy (LE)	Variance (VLE), Mean (MLE), Mean of derivative (MdLE), Maximum (MaLE), Minimum (MiLE), Difference between maximum and minimum (DMLE)
43–46	Derivative of logarithm of energy (dLE)	Variance (VdLE), Maximum (MadLE), Minimum (MidLE), Difference between maximum and minimum (DMdLE)
47–52	MFCCs (First and second coefficients: C_1 and C_2)	Average (MC_1, MC_2), Maximum (MaC_1, MaC_2), Minimum (MiC_1, MiC_2)

Table 2 List of investigated feature sets in this study

Feature set number	Component(s) of feature set
1	All of 52 features (listed in Table 1)
2	VF$_0$, MF$_0$, MdF$_0$, VF$_1$–VF$_3$, MF$_1$–MF$_3$, MdF$_1$–MdF$_3$, MLE, MdLE, MC$_1$, MC$_2$
3	VF$_1$–VF$_3$, MF$_1$–MF$_3$, MaF$_1$–MaF$_3$, MiF$_1$–MiF$_3$, DMF$_1$–DMF$_3$,VdF$_1$–VdF$_3$, MdF$_1$–MdF$_3$, MadF$_1$–MadF$_3$, MidF$_1$–MidF$_3$, DMdF$_1$–DMdF$_3$
4	VF$_0$, MF$_0$, MaF$_0$, MiF$_0$, DMF$_0$, VdF$_0$, MdF$_0$, MadF$_0$, MidF$_0$, DMdF$_0$
5	VF$_0$, MF$_0$, MdF$_0$, MLE, VF$_1$–VF$_3$, MF$_1$–MF$_3$, MdF$_1$–MdF$_3$
6	DMF$_1$–DMF$_3$, DMdF$_1$–DMdF$_3$
7	MaF$_1$–MaF$_3$, MiF$_1$–MiF$_3$
8	VF$_2$, MF$_2$, MdF$_2$
9	VF$_3$, MF$_3$, MdF$_3$
10	MF$_1$
11	MF$_2$
12	MF$_3$
13	MdF$_1$
14	MdF$_2$
15	MdF$_3$
16	VF$_1$
17	VF$_2$
18	MC$_1$, MC$_2$, MaC$_1$, MaC$_2$, MiC$_1$, MiC$_2$, DMC$_1$, DMC$_2$
19	MC$_1$, MC$_2$
20	MC$_1$
21	MC$_2$

neutral sentences. Speech signals were sampled at the rate of 16 kHz and windowed by a 25-ms Hamming window considering 10-ms frame shift.

To provide unambiguous utterances for the emotion classification, a listening test in a two-pass procedure has been performed. First, two listeners were invited to delete the speech data that were very difficult to identify its emotion content. Most of this data suffered from non-professionality of actors, noise, and speaker movement. The second step was to invite five listeners who did not participate in the gathering of speech data. We kept only 4,970 utterances recorded by 30 speakers.

To train and test the classifier, the corpus is divided into two disjoint corpora: one for training (the utterances of 20 speakers including 3,320 utterances corresponding to 67% of the corpus) and the other for testing (1,650 utterances corresponding to 33% of the corpus).

As mentioned before, the fuzzy ARTMAP neural network (FAMNN) is used as the emotion classifier in this study. The FAMNN has been introduced by Carpenter et al. [60]. This network has an architecture for incremental supervised learning of recognition categories and multidimensional maps in response to arbitrary sequences of analog or binary input vectors. It achieves a synthesis of fuzzy logic and adaptive resonance theory (ART) neural networks by exploiting a close formal similarity between the computations of fuzzy method and the ART category choice, resonance, and learning.

The FAMNN has been successfully applied in many tasks such as data mining, remote sensing, and pattern recognition. The FAMNN is considered fast among the members of ARTMAP family due to the computationally cheap mapping between inputs and outputs. Furthermore, as compared to the standard nearest neighbor techniques that are also commonly used, FAMNN requires less memory since it uses a compressed representation of the data and for the same reason, FAMNN requires less classification time.

The ARTMAP networks consist of two ART1 networks, ARTa and ARTb, bridged via inter-ART module. The ART1 module has three layers: input layer (F_0), the comparison layer (F_1), and the recognition layer (F_2). The fuzzy ARTMAP is a natural extension to ARTMAP that uses fuzzy ART instead of ART1 modules. Table 3 shows the specifications of simulated FAMNN in this section.

Figure 2a shows the emotion recognition results when using the feature sets 2, 5, 8, and 9. Comparing the results of feature sets 2 and 5 shows that removing the MFCCs and energy features results in increasing the accuracies of neutral and angry emotion recognition, while it decreases the accuracy of happiness emotion recognition. The features related to F_3 in the feature set 9, in comparison with F_2-related features, are more useful in achieving higher emotion recognition rates.

Figure 2b shows the effects of different feature sets that are related to the first two cepstral coefficients. The effectiveness of using the feature set 18 for happiness emotion recognition, when using the mean of C_1 for neutral state and the mean of C_2 for three emotional states, can be seen in Fig. 2b.

Table 3 Specification of FAMNN in the base experiments

Specification	Value
Learning rate	1
Vigilance parameter	0.99
Number of F_0-layer nodes	104
Number of F_1-layer nodes	104
Number of F_2-layer nodes	3,600
Number of classes	3
Number of training samples	3,320
Number of test samples	1,650

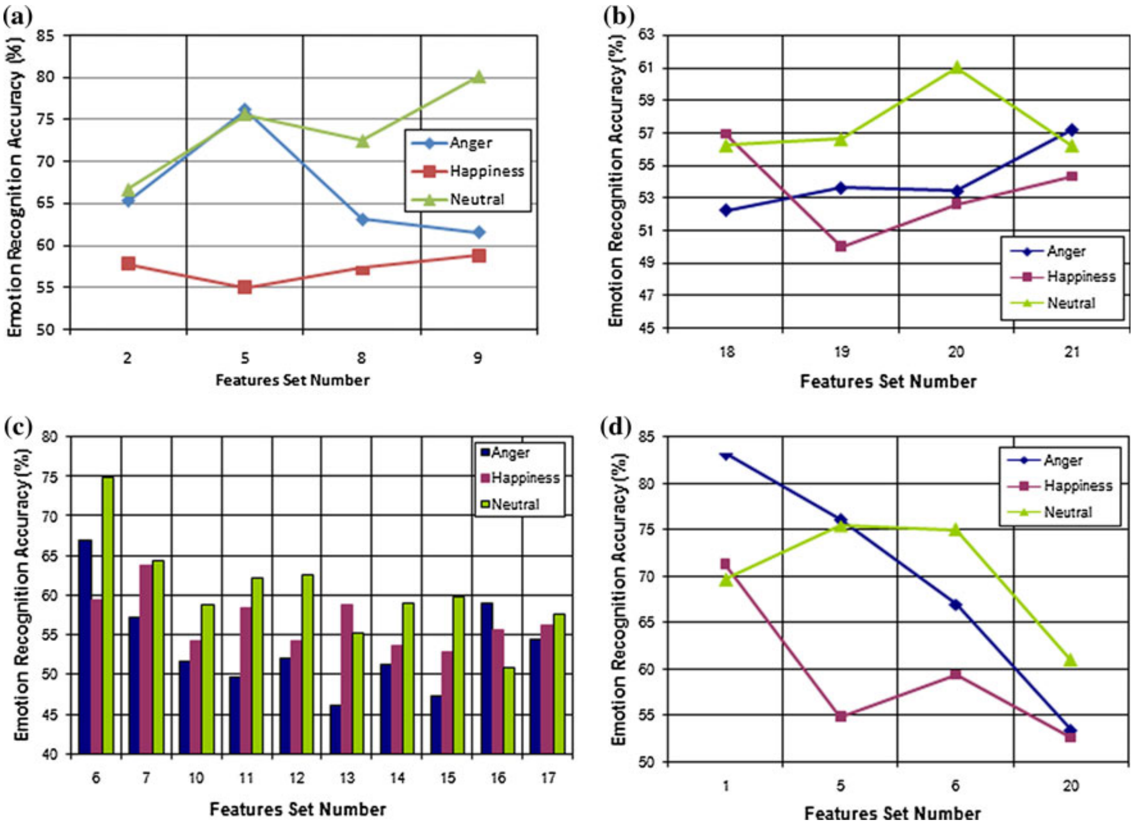

Fig. 2 Emotion recognition accuracy using different feature sets, **a** feature sets 2, 5, 8, and 9, **b** feature sets 18–21, **c** feature sets 6, 7, and 10–17, **d** feature sets 1, 5, 6, and 20

Figure 2c considers the features that are related to formant frequencies. This figure shows the large variations of recognition rate for the different emotions. It can be seen that the features related to the minimum and maximum of F_1–F_3 result in the best performance for happiness emotion recognition. Also, by using the difference between maximum and minimum of F_1–F_3, the recognition rate of neutral and angry emotional states is improved.

Figure 2d shows the emotion recognition results that are obtained for the feature sets 1, 5, 6, and 20 for the three mentioned emotions. As can be seen, some of the formant-related features can be effective in the recognition of a specific emotional state, while they may result in performance degradation in recognition of another emotional state.

Table 4 Average emotion recognition accuracy for different feature sets

Features set number	Emotion recognition accuracy (%)	Features set number	Emotion recognition accuracy (%)
1	74.73	12	56.30
2	63.15	13	53.39
3	79.58	14	54.73
4	75.39	15	53.45
5	68.85	16	55.21
6	67.15	17	66.79
7	61.88	18	55.09
8	64.24	19	53.33
9	66.79	20	55.70
10	54.91	21	55.88
11	56.79		

In Table 4, the emotion recognition accuracies, when averaged over emotional states, are reported for the 21 mentioned feature sets. It is noted that the test dataset includes 1,650 sentences uttered by 10 speakers. These emotion recognition results show that using the feature set 3 with 30 formant frequency-related components results in better performance in comparison with feature set 1 with 52 components. In this way, the emotion recognition results when using feature set 4 with only 10 pitch frequency-related components are noticeable, too. These emotion recognition accuracies show that employing feature selection methods may result in achieving better performance by using a reduced-size feature set.

In Fig. 3, the emotion recognition performance comparison between the feature sets 3 and 4 in different states is depicted. It is noted that after applying the FAMNN, we

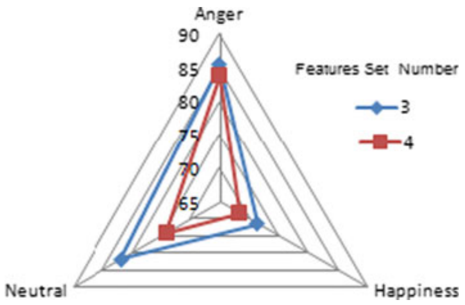

Fig. 3 Performance comparison of emotion recognition for feature sets 3 and 4

obtained three evaluation values, one from each of the three emotion categories. These values are then plotted in the emotion radar chart (Fig. 3), a multi-axes plot that presents the possibility and relativity for each emotion of the test data to each emotion category. Each of the axes stands for one emotion category. An unambiguous emotion should have a peak in one emotion and smaller values in the others. As mentioned earlier, the feature set 3 consists of 30 features that are related to the formant frequencies, and the feature set 4 consists of only 10 features that are related to the pitch frequency. It is noticeable that despite discarding 22 features in feature set 3 as compared to the basic feature set (feature set 1), the emotion recognition accuracy is increased by 4.85%. Also, discarding 42 features from feature set 1, to form the feature set 4, results in 0.66% increment in emotion recognition accuracy. As can be seen, the features that are related to formants are more effective in happiness and neutral emotion recognition than the features that are related to F_0.

4 Influence of feature selection on emotion recognition accuracy

In this study, the fast correlation-based filter (FCBF) method [61] is used for dimension reduction and construction of a lower-size feature space. This method selects the features that are individually informative and two-by-two weakly dependent. It is noted that mutual information (MI) of two vectors **X** and **Y**, $I(\mathbf{X},\mathbf{Y})$, computes the statistical dependency of them in the following way:

$$I(\mathbf{X},\mathbf{Y}) = \sum_{y\in\mathbf{Y}}\sum_{x\in\mathbf{X}} p(\mathbf{X}=x,\mathbf{Y}=y)\log\left(\frac{p(\mathbf{X}=x,\mathbf{Y}=y)}{p(\mathbf{X}=x)p(\mathbf{Y}=y)}\right) \qquad (1)$$

where p is the probability function. Obviously, $I(\mathbf{X},\mathbf{Y})$ is equal to 0, when **X** and **Y** are independent, ($p(\mathbf{X}=x, \mathbf{Y}=y) = p(\mathbf{X}=x)p(\mathbf{Y}=y)$), and is increased when their dependency increases.

In the FCBF method, **Y** is the vector of data labels and $\mathbf{X}_i$ is the vector of ith feature value for all data. That is, when the number of features is N, there are $N+1$ vectors. The FCBF selects features in the two following steps:

1. Removing features ($\mathbf{X}_i$) that are not dependent on the label vector **Y**:

$I(\mathbf{X}_i,\mathbf{Y}) > \varepsilon$ where ε is a positive threshold between 0 and 1. In this way, the FCBF selects the features that are individually informative. In this work, ε is set to 0.01.

2. Removing a remained feature ($\mathbf{X}_i$), which is dependent on the other remained feature ($\mathbf{X}_j$), is more than $I(\mathbf{X}_i,\mathbf{Y})$, while $I(\mathbf{X}_i,\mathbf{Y}) \leq I(\mathbf{X}_j,\mathbf{Y})$:

In this way, the FCBF selects those individually informative features that are also two-by-two weakly dependent.

By using the mentioned MI-based feature selection method, 2, 8, 15, 21, 27, 32, 36, 42, and 47 features have been removed, in nine different experiments, from the main feature set including 52 features. The recognition results of these nine experiments for each of the three emotional states are shown in Fig. 4 and are compared with the results of using basic 52-component feature set (feature set 1). As can be seen, the best recognition results are achieved for angry and neutral emotions when 27 features are discarded. On the other hand, when 15 features are discarded, the recognition rate of happiness emotional state is better than the rates in other experiments.

The average emotion recognition rates in the mentioned nine experiments are also depicted in Fig. 5. As can be seen, the best emotion recognition rate is achieved when the number of selected features is set to 25. To investigate this issue in more detail, when the number of selected features is equal to 37 and 25, the list of selected features and also the joint and disjoint features are reported in

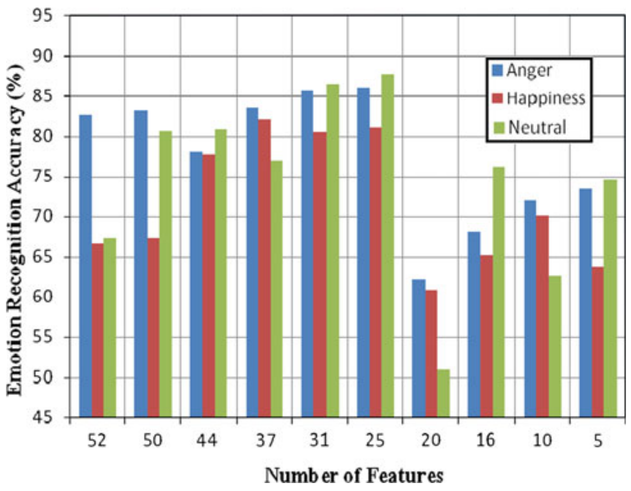

Fig. 4 Influence of feature set-size reduction on emotion recognition accuracy for each emotional state

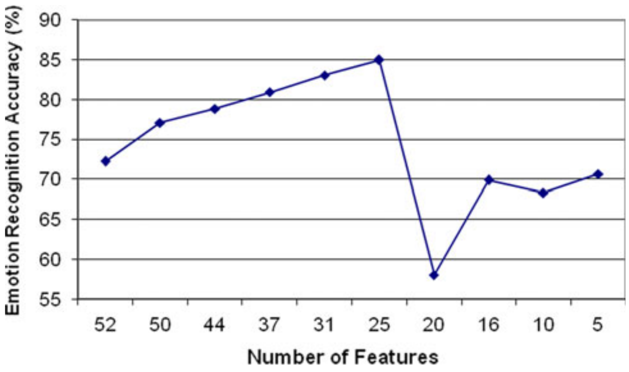

Fig. 5 Influence of feature set-size reduction on average emotion recognition accuracy

Table 5. In this way, we can explore more important features as the joint features in these two feature sets.

As shown in Table 5, most of the joint selected features are related to F_3. Also, most of the supplementary formant features are related to their maximum, minimum, and derivatives. Energy-related features and the mean of cepstral coefficients are also seen in the list of joint selected features. The average slopes of F_2 and LE have the most influence on improving the emotion recognition accuracy as disjoint selected features in Table 5.

Also, due to the significant degradation of emotion recognition rates when 20 features are selected by the FCBF method (Fig. 5), the list of selected features and also the joint and disjoint features for 25, 20, and 16 selected features are reported in Table 6.

As can be seen, most of the joint features in Table 6 are related to derivatives of formants and logarithm of energy. Most of the energy-related features are present in the list of disjoint features in the case of 25 selected features. In the case of 20 selected features, the F_3-related features are discarded in comparison with the case of 25 selected features that show the best performance in these experiments. So, this significant degradation in emotion recognition accuracy is due to discarding such important features.

5 Influence of FAMNN parameters optimization on emotion recognition accuracy

As mentioned before, the GA is used in this study to determine the optimum values for FAMNN parameters.

Table 5 List of total, joint, and disjoint features for 37 and 25 selected features

Number of selected features	Features	Joint features	Disjoint features
25	VF$_0$, MiF$_0$, DMF$_0$, VF$_3$, MF$_3$, MaF$_3$, DMF$_1$, VdF$_2$, VdF$_3$, MdF$_1$–MdF$_3$, MadF$_3$, MidF$_2$, DMdF$_1$, MLE, MdLE, MiLE, DMLE, MadLE, MidLE, DMdLE, MC$_1$, MC$_2$, MaC$_1$	VF$_0$, MiF$_0$, DMF$_0$, VF$_3$, MF$_3$, MaF$_3$, DMF$_1$, VdF$_2$, VdF$_3$, MdF$_1$, MdF$_3$, MadF$_3$, MidF$_2$, DMdF$_1$, MLE, MiLE, DMLE, MadLE, MidLE, MC$_1$, MC$_2$, MaC$_1$	MdF$_2$, MdLE, DMdLE
37	VF$_0$, MF$_0$, MaF$_0$, MiF$_0$, DMF$_0$, VF$_3$, MF$_2$, MF$_3$, MaF$_1$–MaF$_3$, MiF$_1$, MiF$_2$, DMF$_1$, VdF$_2$, VdF$_3$, MdF$_1$, MdF$_3$, MadF$_1$, MadF$_3$, MidF$_1$–MidF$_3$, DMdF$_1$, VLE, MLE, MaLE, MiLE, DMLE, VdLE, MadLE, MidLE, MC$_1$, MC$_2$, MaC$_1$, MaC$_2$, MiC$_1$		MF$_0$, MaF$_0$, MF$_2$, MaF$_1$, MaF$_2$, MiF$_1$, MiF$_2$, MadF$_1$, MidF$_1$, MidF$_3$, VLE, MaLE, VdLE, MaC$_2$, MiC$_1$

Table 6 List of total, joint, and disjoint features for 25, 20, and 16 selected features

Number of selected features	Features	Joint features	Disjoint features
25	VF$_0$, MiF$_0$, DMF$_0$, VF$_3$, MF$_3$, MaF$_3$, DMF$_1$, VdF$_2$, VdF$_3$, MdF$_1$–MdF$_3$, MadF$_3$, MidF$_2$, DMdF$_1$, MLE, MdLE, MiLE, DMLE, MadLE, MidLE, DMdLE, MC$_1$, MC$_2$, MaC$_1$	MaF$_3$, VdF$_2$, VdF$_3$, MdF$_1$, MdF$_2$, MidLE, DMdLE	VF$_0$, MiF$_0$, DMF$_0$, VF$_3$, MF$_3$, DMF$_1$, MdF$_3$, MadF$_3$, MidF$_2$, DMdF$_1$, MLE, MdLE, MiLE, DMLE, MadLE, MC$_1$, MC$_2$, MaC$_1$
20	MaF$_0$, MiF$_0$, MaF$_1$–MaF$_3$, MiF$_1$, VdF$_2$, VdF$_3$, MdF$_1$, MdF$_2$, DMLE, VdLE, MadLE, MidLE, DMdLE, MC$_1$, MC$_2$, MaC$_1$, MaC$_2$, MiC$_1$		MaF$_0$, MiF$_0$, MaF$_1$, MaF$_2$, MiF$_1$, DMLE, VdLE, MadLE, MC$_1$, MC$_2$, MaC$_1$, MaC$_2$, MiC$_1$
16	DMF$_0$, MF$_3$, MaF$_1$–MaF$_3$, MiF$_1$, VdF$_2$, VdF$_3$, MdF$_1$, MdF$_2$, MadF$_2$, MadF$_3$, VdLE, MidLE, DMdLE, MiC$_1$		DMF$_0$, MF$_3$, MaF$_1$, MaF$_2$, MiF$_1$, MadF$_2$, MadF$_3$, VdLE, MiC$_1$

The genetic algorithm is a method for solving optimization problems that is based on natural selection, the process that drives biological evolution [62]. The genetic algorithm repeatedly modifies a population of individual solutions. At each step, the genetic algorithm selects individuals randomly from the current population to be parents and uses them to produce the children for the next generation. For each individual, a pair of parents is selected. There are several methods for selecting parents such as stochastic uniform selection, remainder selection, uniform selection, roulette selection, and tournament selection. For creating the next generation from the current population, GA performs the following procedures: (1) using crossover rules that combine two parents to form children for the next generation, (2) using mutation rules that apply random changes to the individual parent to form children, (3) selecting the best individual of fitness function as elite child. Finally, the following conditions are considered to terminate the algorithm: (1) if the specified number of generations or amount of time (in seconds), which is determined by the user, is reached, (2) if the value of fitness function (or objective function) for the best point is less than or equal to the value, which is determined by the user in fitness limit variable, (3) if there is no improvement in objective function for a sequence of generations or during an interval of time, which is determined by the user. Over successive generations, the population "evolves" toward an optimal solution. Figure 6 shows the flowchart of GA algorithm.

The operation of fuzzy ARTMAP is affected by two network parameters: the choice parameter, α, and the baseline vigilance parameter, ρ (ρ_a, ρ_b, and ρ_{ab}). The choice parameter takes values in the interval (0,1), while the baseline vigilance parameter assumes values in the interval [0,1]. Both of these parameters affect the number of nodes that are created in the category representation layer of FAMNN. Another important parameter is the learning rate, β, which has a value between 0 and 1.

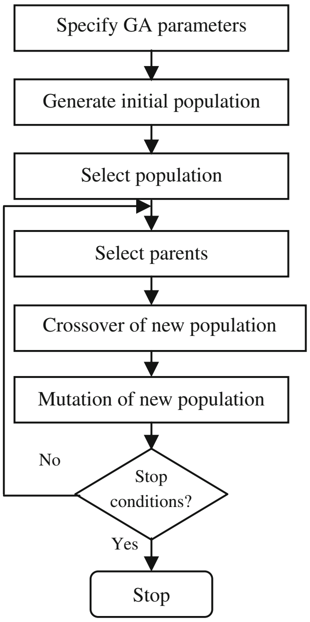
Specify GA parameters
Generate initial population
Select population
Select parents
Crossover of new population
Mutation of new population
No
Stop conditions?
Yes
Stop

Fig. 6 GA flowchart

In this study, the optimum values of mentioned FAMNN parameters, which are corresponding to the minimum of fitness function, are determined by GA. The fitness function is defined over the genetic representation and measures the quality of solution. The fitness function is problem dependent. In this work, the following fitness function is used:

$$F = \frac{1}{pc} \tag{2}$$

where pc is the percentage of correct classification. At the start of genetic algorithm, the values that are created by fitness function are not suitable for the selection process. So, the fitness scaling that maps those raw values to a suitable range for the selection process is necessary. The range of scaled values affects the performance of genetic algorithm. In our simulations, rank fitness scaling function is selected for scaling and removing the effect of the raw scores spread. This method first ranks the population, and then every chromosome receives fitness from this ranking. There are several methods for selecting parents in the GA. By performing several experiments in this study, the stochastic uniform selection function is used, which lays out a line in which each parent corresponds to a section of the line the length of which is proportional to its scaled value. To create the next generation, GA uses elite children who are individuals with the best fitness values in the current generation and guaranteed to survive to the next generation, crossover children who are fractions of the individuals, and mutation children who are individuals in the current generation that random changes are applied to them. In our simulations, the population size is set to 100. Also, we use two elite children, 78 crossover children with scattered function (the fraction of individuals who are used in crossover process is 0.8), and 20 mutation children. The Gaussian mutation function is used and a random number, chosen from a Gaussian distribution, is assigned to each entry of the parent vector. The amount of mutation is decreased at each new generation (proportional to the standard deviation of the Gaussian distribution). The rate of decreasing the average value of mutation is specified by shrink parameter. The standard deviation of the Gaussian distribution is decreased linearly until its final value reaches to 1-shrink times of its initial value at the first generation. The value of shrink parameter is set to 1 in our simulations.

Table 7 Optimized FAMNN parameters when using two feature sets

Feature set used for emotion recognition	Optimized FAMNN parameters				
	α	β	ρ_a	ρ_b	ρ_{ab}
Feature set 3 (Table 2)	0.9751	0.7199	0.6802	0.7914	0.6430
Twenty-five selected features (Table 5)	0.9238	0.7186	0.8354	0.8150	0.6642

Table 8 Emotion recognition accuracy using optimized parameters for FAMNN

Feature set used for emotion recognition	FAMNN parameters	Recognition accuracy of emotional state (%)			Average emotion recognition accuracy (%)
		Anger	Happiness	Neutral	
Feature set 3 (Table 2)	Non-optimized	85.50	71.48	81.70	79.58
	Optimized	85.91	76.10	83.98	82.00
25 selected features (Table 5)	Non-optimized	86.14	81.12	87.63	84.97
	Optimized	91.62	83.95	87.02	87.52

Table 9 Performance comparison of the proposed method and some related researches

Classifier	Number of emotional states	Number of utterances in emotional dataset	Number of features	Average emotion recognition accuracy (%)
MLP [12]	5	4,800	20	85.9
MLP [9]	2	2,433	8	77.0
KNN [9]	4	2,433	37	65.0
SVM [9]	4	2,433	37	73.0
MLP [27]	7	679	43	70.0
KNN [27]	5	679	8	55.0
HMM [19]	4	16,000	59	70.1
MLP [34]	8	10,000 words	15	50.0
GMM [8]	3	7,923	11	71.5
GMM [8]	3	7,923	39	73.2
UBM[a]-GMM [48]	5	9,002	39	43.5[b]
MLP [63]	3	535	46	85.2[c]
HMM [63]	4	535	46	77.8[c]
Hierarchical [63]	7	535	85	71.2[c]
Multi-class SVM [24]	4	338	29[d]	79.5
Multi-class SVM [24]	4	338	28[e]	80.4
Multi-class SVM [24]	4	338	26[f]	78.1
SVM [64]	6	548	24	68.1
SVM [64]	6	548	44	68.6
SVM [64]	6	548	78	67.0
SVM [64]	6	548	237	75.9
SVM [64]	6	548	261	78.2
Fuzzy ARTMAP (Proposed classifier without parameter optimization)	3	4,970	25	84.97

Table 9 continued

Classifier	Number of emotional states	Number of utterances in emotional dataset	Number of features	Average emotion recognition accuracy (%)
Fuzzy ARTMAP (Proposed classifier with parameter optimization)	3	4,970	25	87.52

[a] Universal background model

[b] Drop in the accuracy is due to using non-acted spontaneous emotionally colored data

[c] Best results over 15 different feature sets

[d] Using least square bound feature selection method

[e] Using sequential forward feature selection method

[f] Using mutual information-based feature selection method

The optimized parameters of the FAMNN network when using the feature set 3 and also the 25 selected features (listed in Table 5) are reported in Table 7.

The accuracies of emotion recognition when using optimized FAMNN parameters are reported in Table 8 for the mentioned feature sets. In this way, the accuracies are also reported when the parameters of FAMNN are set by the user to $\alpha = 0.01$, $\rho_a = 0.99$, and $\beta = \rho_b = \rho_{ab} = 1$ as typical values. This case is referred in Table 8 as "non-optimized". It is noted that the values of accuracy are averaged over the test database that includes the utterances of totally 1,650 sentences by 10 speakers.

As can be seen, by using the optimized FAMNN parameters, the average emotion recognition accuracies of mentioned feature sets are improved at least by 2.4%. In this way, the average emotion recognition accuracy of 87.52% is achievable.

The performance of proposed method is compared with some related works in Table 9. As can be seen, the proposed method, when using FCBF feature selection algorithm and optimized FAMN classifier, offers better emotion recognition accuracy with a medium-size feature set.

6 Conclusion

In this paper, by using the fuzzy ARTMAP neural network, an emotion recognition system has been implemented. Accomplishing this, the base features such as pitch frequency, formant frequencies, energy, MFCCs, and their related features have been used.

To investigate the effectiveness of candidate features, different kinds of feature sets have been utilized. For this purpose, the FCBF feature selection method has been used to select the most important features. Experimental results show that by employing the 25 selected features by the FCBF method, the rate of 84.97% for emotion recognition is achievable. To optimize the choice, vigilance, and learning rate parameters of FAMNN with the aim of improving the performance of the system, the genetic algorithm has been used. In this way, by using the optimized FAMNN, the average emotion recognition accuracy is increased and becomes over 87.52%.

Acknowledgment This work is supported by Islamic Azad University-South Tehran Branch under a research project entitled as "Emotion Modelling to Improve Speech Recognition Accuracy in Farsi Language".

References

1. Wang C, Seneff S (2000) Robust pitch tracking for prosodic modeling in telephone speech. In: The proceedings of international conference on acoustics, speech, and signal processing, vol 3, pp 1343–1346
2. Yang B, Lugger M (2010) Emotion recognition from speech signals using new harmony features. Signal Process 90:1415–1423
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 C-C, 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 (Article in Press, doi:10.1016/j.specom.2011.01.006)
9. Yacoub S, Simske S, Lin X, Burns J (2003) Recognition of emotions in interactive voice response systems. In: The proceedings of European conference on speech communication and technology, pp 729–732
10. Gharavian D, Ahadi SM (2005) The effect of emotion on Farsi speech parameters: a statistical evaluation. In: The proceedings of

10th international conference on speech and computer, pp 463–466

11. Gharavian D, Ahadi SM (2008) Stressed speech recognition using a warped frequency scale. IEICE Electron Express 5:187–191
12. Oudeyer P-Y (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum Comput Interact Stud 59:157–183
13. 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 ICSLP, pp 665–668
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. 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 ICSLP, pp 2037–2039
17. Batliner A, Fischer K, Huber R, Spilker J, Nöth E (2003) How to find trouble in communication. Speech Commun 40:117–143
18. Sheikhan M, Gharavian D, Ashoftedel F (2011) Using DTW-neural based MFCC warping to improve emotional speech recognition. Neural Comput Appl (Article in Press, doi:10.1007/s00521-011-0620-8)
19. Kwon OW, Chan K, Hao J, Lee TW (2003) Emotion recognition by speech signals. In: The proceedings of European conference on speech communication and technology, pp 125–128
20. Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13:293–303
21. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Commun 48: 1162–1181
22. Altun H, Polat G (2007) New frameworks to boost feature selection algorithms in emotion detection for improved human-computer interaction. Brain vision and artificial intelligent. Lect Notes Comput Sci 4729:533–541
23. 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
24. 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
25. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44:572–587
26. Wu S, Falk TH, Chan W-P (2011) Automatic speech emotion recognition using modulation spectral features. Speech Commun 53:768–785
27. Luengo I, Navas E, Hernáez I, Sanchez J (2005) Automatic emotion recognition using prosodic parameters. In: The proceedings of Interspeech, pp 493–496
28. Nwe TL, Foo SV, De Silva LC (2003) Speech emotion recognition using hidden Markov models. Speech Commun 41:603–623
29. Dellaert F, Polzin T, Waibel A (1996) Recognizing emotion in speech. In: The proceedings of the international conference on spoken language processing, vol 3, pp 1970–1973
30. Han J, Kamber M (2000) Data mining concepts and techniques. Morgan Kaufman
31. 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
32. 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
33. 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
34. 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
35. 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
36. 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
37. Park CH, Sim KB (2003) Emotion recognition and acoustic analysis from speech signal. In: The proceedings of the international joint conference on neural networks, vol 4, pp 2594–2598
38. Yeh J-H, Pao T-L, Lin C-Y, Tsai Y-W, Chen Y-T (2010) Segment-based emotion recognition from continuous Mandarin Chinese speech. Comput Hum Behav (Article in Press, doi: 10.1016/j.chb.2010.10.027)
39. 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
40. 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
41. 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
42. Kao Y, Lee L (2006) Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In: The proceedings of the international conference on spoken language processing, pp 1814–1817
43. Morrison D, Wang R, de Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centers. Speech Commun 49:98–112
44. Rong J, Li G, Phoebe Chen Y-P (2009) Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manage 45:315–328
45. 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
46. 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 Inf Control 4:1695–1709
47. 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
48. 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)
49. 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

50. Bosch L (2003) Emotions, speech and the ASR framework. Speech Commun 40:213–225
51. Song M, Bu J, Chen C, Li N (2004) Audio-visual based emotion recognition-a new approach. In: The proceedings of IEEE conference on computer vision and pattern recognition, vol 2, pp 1020–1025
52. Song M, Chen C, You M (2004) Audio-visual based emotion recognition using tripled hidden Markov model. In: The proceedings of the international conference on acoustics, speech, and signal processing, vol 5, pp 877–880
53. Barra-Chicote R, Fernández F, Lufti S, Lucas-Cuesta JM, Macías-Guarasa J, Montero JM, San-Segundo R, Pardo JM (2009) Acoustic emotion recognition using dynamic Bayesian networks and multi-space distributions. In: The proceedings of Interspeech, pp 336–339
54. 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)
55. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
56. Sidorova J (2009) Speech emotion recognition with TGI + .2 classifier. In: The proceedings of the EACL, student research workshop, pp 54–60
57. Schuller B, Steidl S, Batliner A (2009) The INTERSPEECH 2009 emotion challenge. In: The proceedings of Interspeech, pp 1–4
58. Clavel C, Vasilescu I, Devillers L (2011) Fiction support for realistic portrayals of fear-type emotional manifestations. Comput Speech Lang 25:63–83
59. Bijankhan M, Sheikhzadegan J, Roohani MR, Samareh Y, Lucas C, Tebiani M (1994) The speech database of Farsi spoken language. In: The proceedings of Australian international conference on speech science and technology, pp 826–831
60. Carpenter GA, Grossberg S, Markuzon N, Reynolds JH, Rosen DB (1992) Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Trans Neural Netw 3:698–713
61. Fleuret F (2004) Fast binary feature selection with conditional mutual information. J Mach Learn Res 5:1531–1555
62. Goldberg DE (1989) Genetic algorithms in search optimization and learning. Addison Wesley
63. Albornoz EM, Milone DH, Rufiner HL (2011) Spoken emotion recognition using hierarchical classifiers. Comput Speech Lang 25:556–570
64. Bitouk D, Verma R, Nenkova A (2010) Class-level spectral features for emotion recognition. Speech Commun 52:613–625