World Applied Sciences Journal 14 (4): 626-638, 2011
ISSN 1818-4952
© IDOSI Publications, 2011

GMM-Based Emotion Recognition in Farsi Language Using Feature Selection Algorithms

[1,2]Davood Gharavian, [1]Mansour Sheikhan and [1]Mansoureh Pejmanpour

[1]EE Department, Islamic Azad University, South Tehran Branch, Iran
[2]EE Department, Shahid Abbaspour University, Tehran, Iran

Abstract: Emotion recognition is the first step toward implementation of an emotional speech recognition system. Emotion has an important role in information transfer from a speaker to a listener. Therefore, emotion recognition of the sentences is very important in real speech recognition systems. The accuracy of an emotion recognition system is dependant on different factors such as the type and number of emotional states, the type of classifier for emotion recognition and the type and number of features. On the other hand, using more features in emotion recognition results in more computational load. In this research, meanwhile implementation of a Gaussian mixture model (GMM) for Farsi language emotion recognition, the most efficient features are selected by using fast correlation-based filter (FCBF) and analysis of variations (ANOVA) approaches for speech emotion recognition. Empirical results show that even by discarding 85% of the features, the average Farsi language emotion recognition accuracy is deteriorated by only about 5%. We also investigate the importance of Mel frequency cepstral coefficients (MFCCs), energy and also the pitch and formants related features on speech emotion recognition accuracy.

Key words: Emotional speech · Emotion recognition · GMM · Feature selection

INTRODUCTION

With the fast growth of telecom services and multimedia devices, contributions in natural communication between machine and human have become necessary [1-4]. Speech is the main tool for human communication. Some factors such as the gender of speaker, dialect, age, language, emotion and stress can influence the speech [5]. All of the mentioned factors give additional information to listener.

Usually it is possible to use different emotional states in a sentence. It is well-known that a sentence without any emotional state can not transfer extra information to speaker and listener, although using emotion in speech leads to some problems for automatic speech recognition [5, 6]. Emotion has an important role in naturalness of man-machine communication, e.g., in speech synthesis [7-10] and automatic speech recognition (ASR) [11-14].

Recognizing the emotions from speech by a machine is first investigated around the mid-1980s using the statistical properties of certain acoustic features [15]. In 1990s, more complicated emotion recognition algorithms were implemented and market requirements motivated further research. For example, ASRs were trained by employing stressed speech instead of neutral in environments such as aircraft cockpits [16]. Iterative algorithms estimated the acoustic features more precisely. In this way, advanced classifiers which used timing information were proposed [17, 18]. Nowadays, the research in this field is focused on finding the reliable informative features and combining powerful classifiers that improve the performance of emotion detection systems in real-life applications [19-23].

The effect of using formants and pitch frequency features on improving the performance of emotion recognition systems is investigated in this paper. So, by generating various supplementary features, based on the first three formant frequencies (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)", "log energy" and "their velocity (dC_i, dLE) and acceleration (ddC_i, ddLE)", a new rich

Corresponding Author: Davood Gharavian, EE Department, Shahid Abbaspour University,
Tehran, Iran. Tel: +98-21-7393-2612.

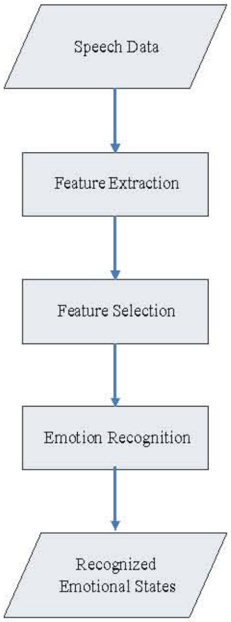

Fig. 1: Framework for emotion recognition from speech.

medium-sized feature vector is proposed in this study. Recognizing the emotional states in speech is performed by using Gaussian mixture model (GMM), as well. To reduce the number of features, two feature selection methods, based on mutual information (MI) and analysis of variations (ANOVA), are used in this research.

The rest of the paper is organized as follows: we introduce the background and related works in section 2. The speech corpus and GMM toolkit is introduced in section 3. The feature selection approaches are reviewed in section 4. The experiment design and empirical results are presented in section 5 and finally in section 6, we conclude the paper.

Background and Related Works: The basics of most existing researches on emotion recognition can be summarized in the diagram shown in Fig. 1. As shown in Fig. 1, the features of speech are extracted at the first stage. These features are the basic acoustic or linguistic features, such as pitch-related or spectral-related features. In addition, some transform functions are often employed to convert the speech features between different data domains [24]. Some of the extracted features used by some research groups in the recent decade are listed in Table 1.

The second stage reduces the size of feature set by selecting the most relevant subset of features and removing the irrelevant ones [25-30].

The third stage in this system is for training and building a classification model, e.g. using machine learning algorithms, to predict the emotional states. In 1990s, most of the emotion recognition models were based on the maximum likelihood Bayes (MLB) [27, 31-33] and linear discriminate classification (LDC) [27]. In the recent years, artificial neural networks (ANNs) [23, 34-38], support vector machines (SVMs) [4, 20, 22, 23, 39-43], K-nearest neighbor (KNN) [43-45], GMMs [46, 47] and hidden Markov models (HMMs) [23, 48-52] have been used for emotion recognition.

To reduce the size of features, the feature selection methods have been used in some researches. For example, considering the features at different levels such as frame-level, syllable-level and word-level and using them in emotion recognition system has been reported in [42]. Some feature selection methods such as sequential floating forward selection (SFFS) [32], wrapper approach with forward selection [38], forward feature selection (FFS) and backward feature selection (BFS) [44],

Table 1: Feature vectors used in emotion recognition from speech

Feature vector	Research group
Pitch, bandwidth, energy, duration, formants [45]	Petrushin (2000)
Pitch, intensity, duration [56]	Amir (2001)
Pitch, energy, duration, formants [57, 58]	Cai *et al.* (2003), Lee *et al.* (2003)
Pitch, energy [48]	Schuller *et al.* (2003)
Pitch, log energy, formants, MFCCs [59, 42]	Kwon *et al.* (2003), Kao *et al.* (2006)
Pitch, energy, formants, MFCCs, vocal tract cross-section areas, speech rate [23]	Ververidis *et al.* (2006)
Pitch, intensity, speech rate [22]	Shami *et al.* (2007)
Pitch, energy, MFCCs, LPCs[a] [20]	Altun *et al.* (2009)
Formants, pitch, energy, spectral features [32]	Ververidis *et al.* (2006)
Formants, intensity, pitch [38]	Sidorova (2009)
LPCs, MFCCs [44]	Pao *et al.* (2008)
Pitch, energy, duration, MFCCs [33]	Haq *et al.* (2008)
Pitch, Energy, duration [43]	Yacoub *et al.* (2003)
Pitch, MFCCs [47]	Neiberg *et al.*(2006)
V/UV[b], energy, pitch, VAD[c] [46]	Luengo *et al.* (2005)

[a] Linear Prediction Coefficients

[b] Voiced/Unvoiced

[c] Voice Activity Detection

Table 2: Pitch and formants frequency features

Features	Abbreviations
Pitch and formants	F_0, F_1, F_2, F_3
Derivative of F_0-F_3	dF_0, dF_1, dF_2, dF_3
Logarithm of F_0-F_3	$\log F_0$, $\log F_1$, $\log F_2$, $\log F_3$
Zero-mean value of F_0-F_3	ZF_0, ZF_1, ZF_2, ZF_3

principal component analysis (PCA), linear discriminate analysis (LDA) [33] and genetic algorithm feature selection (GAFS) [53, 54] have been also used for selecting features in speech emotion recognition systems.

As mentioned earlier, finding the most efficient features for emotion recognition, using ANOVA and MI-based feature selection methods is the main object of this research. The literature surveys show that ANOVA and MI are not used as conventional feature selection methods for speech emotion recognition in spite of the widespread usage in other signal processing systems.

Speech Corpus and Tools: In this study, the utterances of 22 native Farsi speakers have been recorded and formed the emotional speech corpus. Each speaker has uttered 252 sentences in four emotional states: neutral (N), happiness (H), anger (A) and interrogative (I). The numbers of sentences were 34 for anger, 69 for happiness, 50 for interrogative and 99 for neutral states. 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 GMM are 12 MFCCs, logarithm of energy, the first three formant frequencies and the pitch frequency. The training corpus contains sentences of 14 speakers and test corpus includes speech of 8 speakers. The basic model of GMM is trained using 39 features for each frame. Each vector contains MFCC coefficients and logarithm of energy and the velocity and acceleration coefficients of them. To study the effect of formant and pitch frequency features, they are added to the end of basic feature vector.

Using three formant frequencies and pitch frequency, 16 supplementary features are calculated. These features contain formants and pitch, derivative and logarithm of them and their zero-mean values at each frame. To compute the zero-mean value, the mean value of that feature in each sentence is subtracted from the original value at each frame.

Table 2 contains these parameters and their abbreviations. The log operator decreases variations and Z operator eliminates the effect of mean value for each parameter. For each operator, we perform a separate test. These 16 supplementary features and 39 basic features are used in GMM.

Using these feature vectors for emotion recognition, the effect of MFCC coefficients, energy and their velocity and acceleration values can be evaluated. In addition, by using supplementary features the effect of formants and pitch frequency features can be investigated for emotion recognition. At the end, the results of emotion recognition using these features and the results of feature selection methods, can be coupled and used for evaluating the influence of each feature or feature set for emotion recognition in Farsi language.

In the following, the accuracy of emotion recognition system when the mentioned features are used and also the effect of reducing the number of features by the two mentioned feature selection methods are reported.

Feature Selection Algorithms: For dimension reduction and construction of a lower-sized feature space, two open-loop (independent of the classifier) feature selection methods are used in this paper.

The first method, which is MI-based, is fast correlation-based filter (FCBF) algorithm [55]. FCBF selects the features which are individually informative and two-by-two weakly dependant. It is noted that Mutual Information (MI) of two vectors *X* and *Y*, I(X<Y) computes statistical dependency of them in the following way:

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

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

In FCBF method, *Y* is the vector of data labels and X_i is the vector of *i*th feature value for all data. That is, when the number of features is *N*, there are *N+1* vectors. FCBF selects features in the two following steps:

- Removing features (X_i) which are not dependant on the label vector *Y*:

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

- Removing a remained feature (X_i) which its dependency on other remained feature (X_j) is more than $I(X_i,Y)$, while $I(X_i,Y)\leq I(X_j,Y)$: in this way FCBF selects those individually informative features that are also two-by-two weakly dependant.

Another method is the one-way ANOVA in which discrimination is based on the variations between and within classes indicated by an index. This index is called *p-value* that is between 0 and 1. Strong or weak ability of the features in discrimination corresponds to a *p-value* close to 0 and 1, respectively. The *p-value* is computed through *F*-test which is a ratio of "between-group variation" to "within-group variation". Larger *F* means more difference between groups than within groups. It is noted that one-way ANOVA investigates discrimination of groups based on only one feature (by ignoring the interactions with other features).

In this work, the features are sorted based on *p-value* and *F*. Then the features with minimum *p-value* and maximum *F* are selected as the most discriminative features.

Speech Emotion Recognition: As mentioned before, GMM is used for emotion recognition in this research. In this section, the effect of number of mixtures on emotion recognition accuracy using base features (MFCC and energy) is evaluated first. Evaluating the effect of supplementary features on accuracy is the second subject in this section. Finally, using the selected features in emotion recognition system is investigated in this section.

Table 3 shows the accuracy of emotion recognition system for happiness, anger and neutral emotional states using the base model. These results are achieved with 32 and 64 mixtures for GMM, respectively. These results show that, GMM with 64 mixtures absolutely improves average emotion recognition accuracy (AERA) by about 20% as compared to 32 mixtures. However, the training time of GMM with 64 mixtures is noticeably longer than 32 mixtures.

To study the effect of additional formants and pitch frequency features, these 16 features are augmented to the end of feature vectors. Table 4 shows the emotion recognition accuracy using 55 features for GMM.

The results reported in Table 2 and Table 3 show that by using 16 supplementary features and employing GMM with 32 mixtures, the AERA of happiness and anger states is improved by about 11.7% and 1.1%, respectively. However, the AERA of neutral state is deteriorated by about 10.5% in this condition. So, it seems that all of the mentioned 16 supplementary features may not improve emotion recognition of neutral speech. Because of the variety of features, it is needed to study the effect of each feature on emotion recognition accuracy. Selection of effective features can increase the processing speed without noticeable deterioration of accuracy. In the rest of paper, the most effective features for emotion recognition are selected by using MI-based and ANOVA feature selection methods.

Feature Selection Using MI-Based and ANOVA Approaches: In Fig. 2, the block diagram of emotion recognition system with the capability of selection more efficient features is depicted. As shown in Fig. 2, the most effective features are selected and the GMM models are trained. Using the test corpus, emotion recognition

Table 3: Emotion recognition accuracy using base model with 32 and 64 mixtures.

Number of Mixtures	Emotion Recognition Accuracy (%) Happiness	Anger	Neutral	Average Accuracy (%)
32	59.9	71.2	63.2	64.8
64	77.0	91.6	84.1	84.2

Table 4: Emotion recognition accuracy using 55 features

Number of Mixtures	Emotion Recognition Accuracy (%) Happiness	Anger	Neutral	Average Accuracy (%)
32	71.6	73.3	52.7	65.9
64	78.5	93.1	85.2	85.6

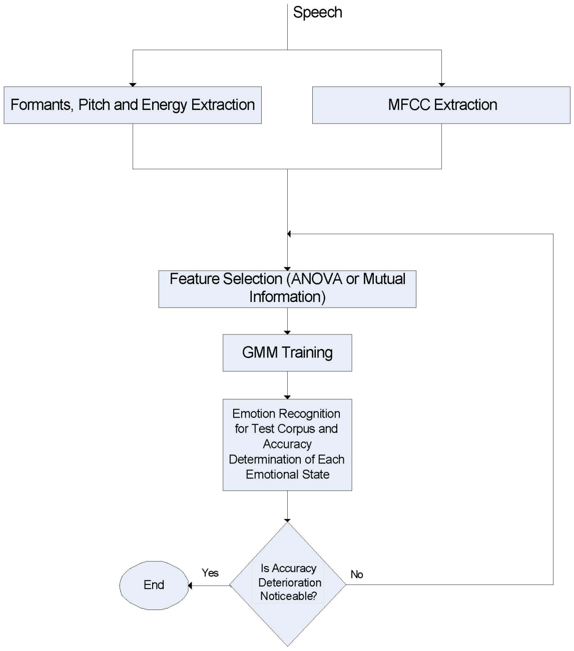
Speech
Formants, Pitch and Energy Extraction
MFCC Extraction
Feature Selection (ANOVA or Mutual Information)
GMM Training
Emotion Recognition for Test Corpus and Accuracy Determination of Each Emotional State
Is Accuracy Deterioration Noticeable?
Yes
No
End

Fig. 2: Block diagram of emotion recognition system with feature selection capability.

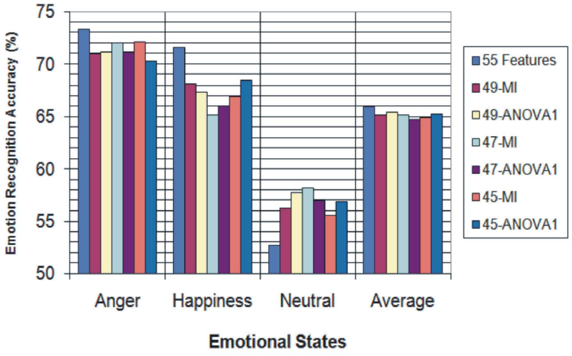

Fig. 3: Emotion recognition accuracy-Supplementary features selected by MI/ANOVA.

accuracy is evaluated. If the accuracy deterioration is negligible, the process is repeated using a selected reduced-size feature set.

To investigate the influence of MFCCs and energy on emotion recognition accuracy, the feature selection algorithms are used in two steps. In the first step, the feature selection approaches are applied to 16 supplementary features and 6, 8, or 10 features are discarded, respectively. In the second step, the feature selection approaches are applied to all of the features and in several trials 6, 8, 10, 21, 40, or 47 features are discarded. The mentioned numbers for discarded features are chosen so that we have sample small- and medium-size feature sets.

Feature Selection for Speech Emotion Recognition: The effects of discarding 6, 8, or 10 features from supplementary features by MI-based and ANOVA methods on the emotion recognition accuracy are shown in Fig. 3. We called this experiment for ANOVA method as ANOVA1 and the selection of 49 features by MI and ANOVA1 are denoted as 49-MI and 49-ANOVA1, respectively. Similar abbreviations are used for selection of 47 or 45 features. As shown in Fig. 3, the emotion recognition accuracies when using all of the features are also depicted for different emotional states.

Based on the experiments of this section, the following statements can be concluded:

- The derivative of formant frequencies, ZF_0 and ZF_3 are common features that are discarded by 49-MI and 49-ANOVA1. The average emotion recognition accuracies for 49-MI and 49-ANOVA1 are 65.1% and 65.4%, respectively.
- The maximum deterioration of average emotion recognition accuracy for 49-MI and 49-ANOVA1 is about 0.8% as compared to the case of no-feature selection (NFS) (using 55 features).
- Based on the average emotion recognition accuracy values, ANOVA performs better as compared to MI in selection of 49 features.
- The maximum deterioration of average emotion recognition accuracy for 47-MI and 47-ANOVA1 is about 1.2% as compared to NFS.
- 45-MI and 45-ANOVA1 algorithms do not offer noticeable deterioration of emotion recognition accuracy and the performance of them is slightly better as compared to 47-MI and 47-ANOVA1.

The detailed results are reported in Table 5. As shown in Table 5, the formant frequencies, their logarithms and also the pitch frequency have the most influence on emotion recognition, respectively. The mean, logarithm and zero-mean parameters of F_1 are more important among three formants frequencies. On the other hand, the derivative of F_1 is the first candidate for discarding by MI and ANOVA, as compared to other formants. So, the selection of features in three mentioned experiments improves neutral speech recognition accuracy.

In 49-ANOVA1, the values of *p* and *F* parameters for energy (LE) feature, as the most discriminative feature are 0 and 5858, respectively. The values of *p* and *F* parameters for the six discarded features are reported in Table 6.

Table 5: Details of empirical results- 49, 47 and 45 features selected by MI/ANOVA.

Feature Selection Algorithm	Number of Features	Discarded Features	Effects on Emotion Recognition[a]
MI	49	dF_1-dF_3, ZF_0, ZF_2, ZF_3	AERA=65.1% Discarding ZF_2 and dF_1: Accuracy is increased in N state as compared to 49-ANOVA1 Discarding dF_1: Accuracy is increased in N state as compared to 49-ANOVA1
ANOVA	49	dF_0-dF_3, ZF_0, ZF_3	AERA=65.4% Accuracy as compared to 49-MI: 0.8% decrement in H state and 1.4% increment in N. Discarding dF_0: Accuracy is decreased in H state as compared to 49-MI
ANOVA	49	dC_8, dC_{11}, ddC_5, ddC_8, ddC_{10}, dF_1	AERA=66.1% AERA as compared to 49-ANOVA2: 0.7% increment by discarding MFCCs MFCC discarded features: Velocity and acceleration of high order MFCCs
MI	47	dF_1-dF_3, ZF_0-ZF_3, log F_0	AERA=65.1%
ANOVA	47	dF_0-dF_3, ZF_0, ZF_2, ZF_3, log F_0	AERA=64.7% 47-ANOVA1 discard dF_0 instead of ZF_1 as compared to 47-MI.
ANOVA	47	dC_5, dC_8, dC_{11}, ddC_5, ddC_8, ddC_{10}, ddC_{12}, dF_1	AERA=66.2% Discarded supplementary feature: dF_1 Discarding dC_5 and ddC_{12}: Accuracy is increased in N state and accuracy is decreased in A and H states as compared to 49-ANOVA2
MI	45	dF_1-dF_3, ZF_0-ZF_3, F_0, log F_0, log F_2	AERA=64.9% Discarding log F_2: Accuracy is increased in N state as compared to 45-ANOVA
ANOVA	45	dF_0-dF_3, ZF_0, ZF_2, ZF_3, F_0, log F_0, log F_3	AERA=65.2% Accuracy as compared to 45-MI: Increment in N and H states Discarding log F_3: Accuracy is decreased in A state as compared to 45-MI
ANOVA	45	dC_4, dC_5, dC_8, dC_{11}, ddC_1, ddC_5, ddC_8, ddC_{10}, ddC_{12}, dF_1	AERA=66.2% Discarded supplementary feature: dF_1 AERA as compared to NFS: 0.3 increment Discarding dC_4 and ddC_1: Accuracy is decreased in A and N states and accuracy is increased in H state as compared to 47-ANOVA2.

[a] N, H and A stand for neutral, happiness and angry states.

Table 6: Values of p and F parameters for 6 discarded features for 49-ANOVA1

Features	dF_1	dF_0	dF_2	ZF_0	ZF_3	dF_3
p	0.17	$2.9e^{-11}$	0	0	0	0
F	2	17	54	82	104	110

In another experiment, ANOVA is used to discard 6, 8, or 10 features from the total feature set. We call this setup as ANOVA2. To study the influence of MFCCs and energy on emotion recognition accuracy, the results of ANOVA1 are also depicted in Fig. 4. The results of these three feature selection experiments are shown as 49-ANOVA2, 47-ANOVA2 and 45-ANOVA2 in Fig. 4. Based on the experiments of this section, the following statements can be concluded:

- The results of 49-ANOVA2 experiment show that the h igh-order MFCCs are in priority for feature discarding. Discarding MFCCs improves the emotion recognition accuracy for anger and happiness states and negligible deterioration is experimented for neutral state as compared to 49-ANOVA1.
- The notable point is that by using 49-ANOVA1 and 49-ANOVA2, the average emotion recognition

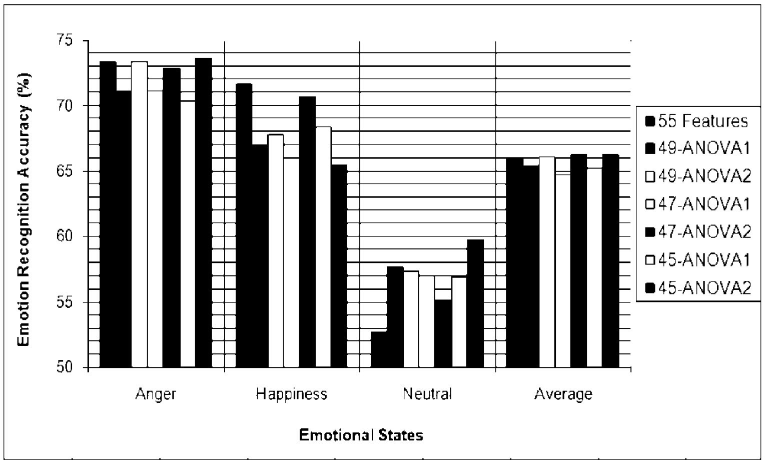

Fig. 4: Emotion recognition accuracy-ANOVA feature selection.

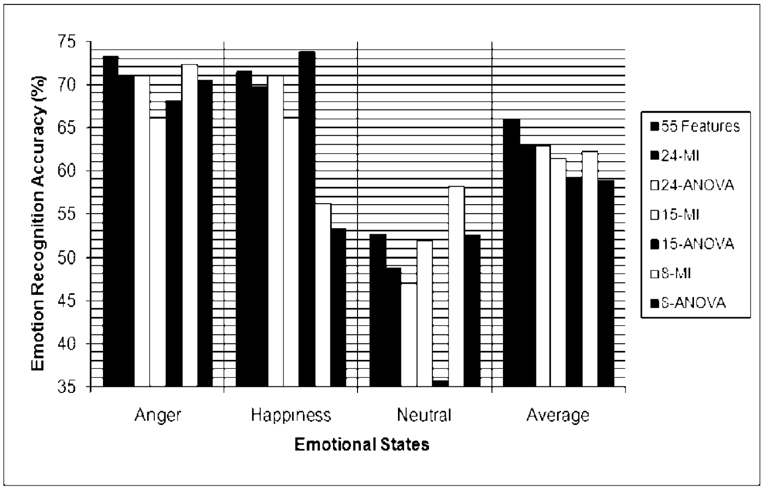

Fig. 5: Emotion recognition accuracy-24, 15 and 8 features selected by MI/ANOVA.

accuracy is improved as compared to NFS. On the other hand, noticeable improvement is achieved for neutral speech.

- 47-ANOVA2 experiment shows better accuracy as compared to 47-ANOVA1 and also NFS.
- The average emotion recognition accuracy using the selected features by 45-ANOVA2, is about 1.0% higher than 45-ANOVA1 experiment and about 0.3% higher than NFS.

In other experiments, 31, 40 and 47 features are discarded from 55 features by MI and ANOVA methods. We call these experiments as 24-MI/24-ANOVA, 15-MI/15-ANOVA and 8-MI/8-ANOVA, respectively. The AERAs of these experiments are depicted in Fig. 5.

Based on the experiments of this section, the following statements can be concluded:

- All MFCC derivatives are discarded by MI-based and ANOVA feature selection algorithms in the mentioned experiments.
- The emotion recognition accuracy for neutral state is improved when using 24-MI as compared to 24-ANOVA. Although, this accuracy is deteriorated for happiness state. However, the average emotion recognition accuracy for two experiments is almost equal.
- The results of feature selection in 15-MI and 15-ANOVA show that in these two experiments, 7 common features are selected from 16 supplementary features. The 15-MI has better performance as compared to 15-ANOVA. However, the average emotion recognition accuracy in 15-MI is about 4.5% lower than the baseline results. This is noted that in this experiment, the deterioration of accuracy for neutral speech is less than 1%.

Table 7: Details of empirical results-24, 15 and 8 features selected by MI/ANOVA

Feature Selection Algorithm	Number of Features	Selected Features	Effects on Emotion Recognition[a]
MI	24	C_1-C_4, C_6, C_8, C_{11}, E, dE, ddC_1, ddLE, F_0-F_3, dF_0, dF_3, log F_0-log F_3, ZF_1-ZF_3	AERA=63.1% MFCC discarded features: Velocity and acceleration features except for ddC_1 and high order MFCC (C_5, C_7, C_9, C_{10} and C_{12}) features Energy discarded features: Velocity and acceleration features
ANOVA	24	C_1-C_{12}, LE, ddC_2, F_0-F_3, log F_0-log F_3, ZF_1-ZF_2	AERA=62.9% MFCC discarded features: Velocity and acceleration features except for ddC_2 MFCC selected features: C_1-C_{12}
MI	15	C_1-C_3, C_6, C_{11}, LE, F_0-F_3, dF_0, log F_1-log F_3, ZF_1	AERA=61.4% MFCC discarded features: Velocity and acceleration features Selected supplementary features: F_0-F_3, log F_0-log F_3, dF_0, ZF_1
ANOVA	15	C_1-C_6, C_{11}, LE, F_1-F_3, log F_1-log F_3, ZF_1	AERA=59.2% MFCC discarded features: Velocity and acceleration features Accuracy as compared to 15-MI: Increment in H state and decrement in N state
MI	8	C_1-C_4, C_8, LE, ddC_1, F_1	AERA=62.2% MFCC discarded features: Velocity and acceleration features Selected supplementary feature: F_1
ANOVA	8	C_1-C_3, C_6, C_{11}, LE, F_1, log F_1	AERA=62.2% MFCC discarded features: Velocity and acceleration features Selected supplementary features: F_1, log F_1

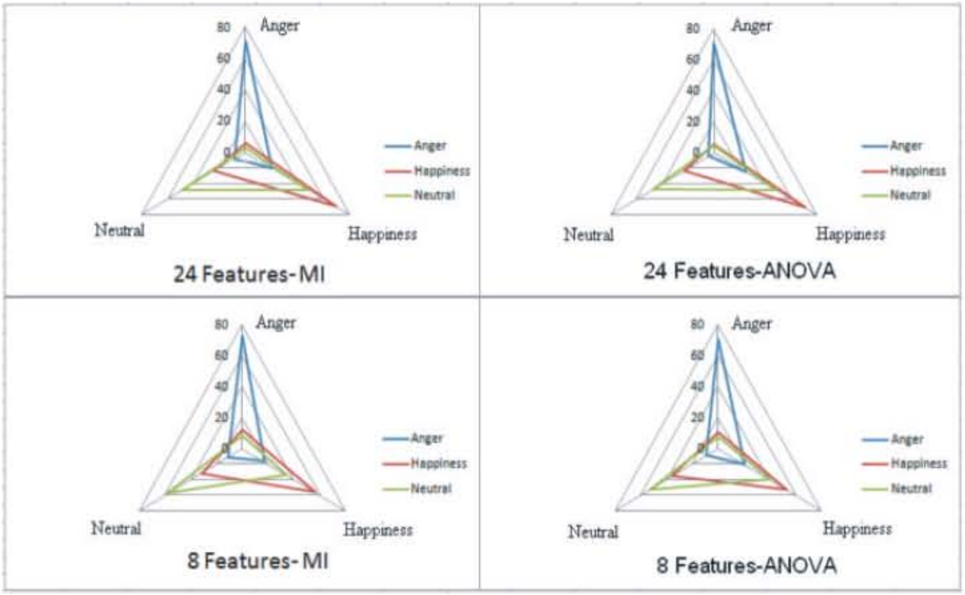

Fig. 6: Radar graph of emotion recognition-24 and 8 features selected by MI/ANOVA.

- The 8-MI has better performance as compared to 8-ANOVA and this is valid for all of the mentioned emotional states.

So, we can conclude that MI-based algorithm performs better in constructing small-size feature sets as compared to ANOVA algorithm and also ANOVA is better for large-size feature sets. In this way, the detailed results are reported in Table 7. The confusion matrix of emotion recognition is shown in Table 7, when using GMM with 32 mixtures. To evaluate the effect of feature selection on confusion matrix, Fig. 6 shows the radar graph of emotion recognition accuracy and confusion results when using 24 and 8 selected features by MI or ANOVA algorithms.

Recognition of Interrogative Sentences: Interrogative sentences are not usually considered as emotional sentences. However, in this study the GMM is trained for these sentences. Our investigations in Farsi language

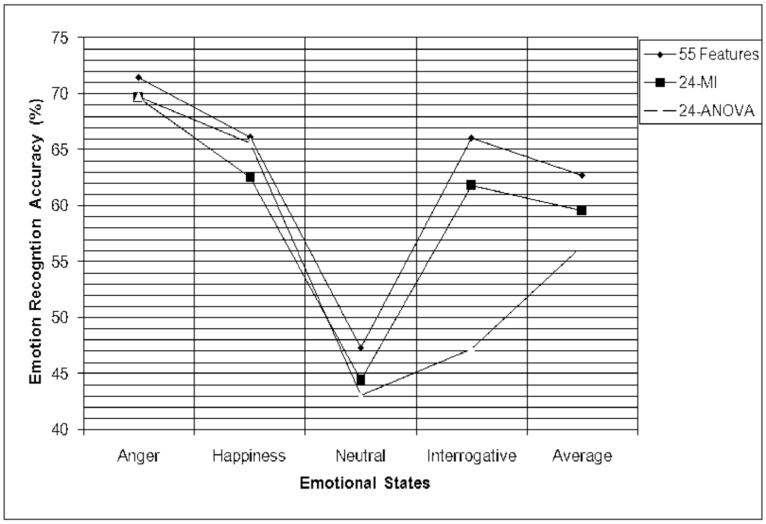

Fig. 7: Emotion recognition accuracy of interrogative sentences as compared to other emotional states using 24 selected features.

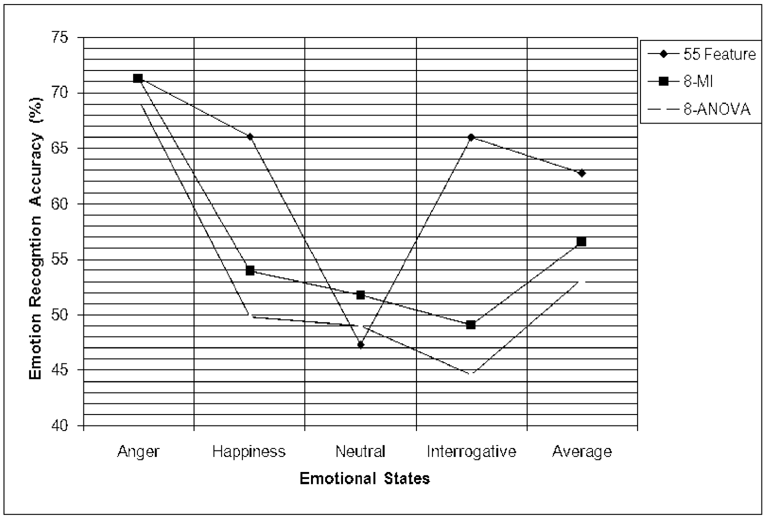

Fig. 8: Emotion recognition accuracy of interrogative sentences as compared to emotional states using 8 selected features.

show that a speaker usually changes a neutral sentence to an interrogative one by applying modifications at the end of it. Therefore, after silence deletion for each sentence, only the last 25% of it is used in training of GMM models. We performed all the mentioned experiments for this case, as well. As sample results, Fig. 7 and Fig. 8 depict the emotion recognition accuracy for 55 features (NFS) and also for 24 and 8 selected features, respectively. As shown in these figures, by decreasing the number of selected features, the recognition accuracy of interrogative sentences has experienced the most deterioration as compared to other emotional states.

CONCLUSION

In this paper, the effect of MFCCs, energy, formant and pitch related features on improving the performance of emotion recognition systems have been investigated. To decrease the computational load, MI-based and ANOVA feature selection methods have been employed. In this way, various combinations of the features have been selected by feature selection algorithms. The performance of the proposed system has been compared with some other emotion recognition systems (Table 9).

Table 8: Confusion matrix of emotion recognition using base model with 32 mixtures

	Predicted		
Actual	Anger	Happiness	Neutral
Anger	71.2	17.1	11.7
Happiness	8.1	59.9	31.9
Neutral	3.5	33.2	63.2

Table 9: Performance of typical systems for emotion recognition in the recent decade

Emotional States	Selected Features	Classifier(s)	Recognition Rate (%)	Feature Selection Methods
Happiness, anger, sadness, neutral [60]	Pitch and its slope, formants, MFCCs	SVM, ANN	71, 42	No
Happiness, anger, tiredness, sadness, neutral [59]	Pitch, log energy, formants, MFCCs and their Δ and ΔΔ	GSVM[a]	41	No
Happiness, anger, anxiety, fear, tiredness, disgust, neutral [61]	MFCCs, energy, dC_i, dE, ddC_i, ddE	GMVAR[b], ANN, HMM	76, 55, 71	No
Happiness, anger, tiredness, sadness, disgust, fear, neutral [62]	MFCCs, log energy, dC_i, dE, ddC_i, ddE	HMM	81	No
Happiness, anger, sadness, neutral [20]	Pitch, sub-band energies, MFCCs, LPC	Multi-class SVM	80	No
Happiness, anger, sadness, fear, neutral [24]	Pitch, intensity, zero crossing rate, spectral features	K-NN	66	No
Neutral, anger, fear, happiness, sadness [42]	Pitch, log energy, formants, MFCCs	SVM	90	Dividing features to different levels
Anger, happiness, neutral, sadness, surprise [32]	Formants, pitch, energy, spectral features	MLB	53.7 (DES Database) 57.2 (SUSAS Database)	SFFS
Fear, disgust, happiness, boredom, neutral, sadness, anger [38]	Formants, intensity, pitch	ANN	78.6	Wrapper approach with forward selection
Anger, happiness, sadness, boredom, neutral [44]	LPC, MFCCs	KNN	79.55[c]	FFS, BFS
Anger, disgust, fear, happiness, neutral, sadness, surprise[33]	Pitch, energy, duration, MFCCs	MLB	53[c]	PCA, LDA
Happiness, anger, sadness, fear, neutral [45]	Pitch, speaking rate, formants, bandwidth	KNN	70	Instance-base learning
Anger, fear, surprise, disgust, joy, sadness[46]	V/UV, energy, pitch, VAD	GMM (512 mixtures)	92.3	No
Neutral, emphatic, negative [47]	Pitch, MFCCs	GMM (512 mixtures)	93	No
Happiness, anger, neutral and interrogative (proposed model)	MFCCs, log energy and their Δ and ΔΔ, formant and pitch-related features	GMM (64 mixtures)	84.2	No
Happiness, anger, neutral and interrogative (proposed model)	MFCCs, log energy and their Δ and ΔΔ, formant and pitch-related features	GMM (32 mixtures)	65.1, 66.3	MI, ANOVA

[a] *Gaussian SVM*

[b] *Gaussian Mixture Vector Autoregressive Model*

[c] *Maximum Emotion Recognition Rate*

The accuracy of the proposed system is reported in the last row of Table 9. Because of the different target emotional states and also feature sets in each research, selection of the most effective approach is impossible. However, the proposed medium-size feature vector in this research and the performance improvement by adding formant and pitch-related parameters along with using feature selection methods, show the effectiveness of proposed approach in developing customized emotion recognition and emotion spotting systems.

It is noted that the proposed model with 64-mixtures for GMM achieved an AERA close to the recognition rate reported in [46] and [47] with 512 mixtures. So, by using GMM with more mixtures and employing MI-based feature selection algorithm, achieving emotion recognition system with small-size feature set and competitive AERA is expectable. As the future research, the authors would like to apply the achievements of the current study to their recent researches in emotion recognition systems [63-65], emotion spotting systems [64], ad emotional speech recognition systems [66].

ACKNOWLEDGEMENT

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. Clavel, C., I. Vasilescu, L. Devillers, G. Richard and T. Ehrette, 2008. Fear-Type Emotion Recognition for Future Audio-Based Surveillance Systems. Speech Communication, 50: 487-503.
2. Inanoglu, Z. and S. Young, 2009. Data-Driven Emotion Conversion in Spoken English. Speech Communication, 51: 268-283.
3. Leon, E., G. Clarke, V. Callaghan and F. Sepulveda, 2007. A User-Independent Real-Time Emotion Recognition System for Software Agents in Domestic Environments. Engineering Applications of Artificial Intelligence, 20: 337-345.
4. Morrison, D., R. Wang and L.C. de Silva, 2007. Ensemble Methods for Spoken Emotion Recognition in Call-Centers. Speech Communication, 49: 98-112.
5. Wang, C. and S. Seneff, 2000. Robust Pitch Tracking for Prosodic Modeling in Telephone Speech. In the Proceedings of International Conference on Acoustics, Speech and Signal Processing, 3: 1343-1346.
6. Huang, X., A. Acero and H.W. Hon, 2005. Spoken Language Processing. A Guide to Theory, Algorithm and System Development, Prentice Hall.
7. Sheikhan, M., M. Nasirzadeh and A. Daftarian, 2005. Design and Implementation of Farsi Text to Speech System. Journal of Engineering Faculty, Ferdowsi University of Meshed, 17: 31-48. (in Farsi)
8. Sheikhan, M., 2007. Automatic Prosody Generation by Neural-Statistical Hybrid Model for Unit Selection Speech Synthesis. Journal of Biomedical Engineering, 1(new): 227-240. (in Farsi)
9. Sheikhan, M., 2003. Prosody Generation in Farsi Language. In the Proceedings of International Symposium on Telecommunications, pp: 250-253.
10. Sheikhan, M., M. Nasirzadeh and A. Daftarian, 2006. Text to Speech for Iraninan Dialect of Farsi Language. In the Proceedings of Second Workshop on Farsi Computer Speech, University of Tehran, pp: 39-53.
11. Sheikhan, M., M. Tebyani and M. Lotfizad, 1997. Continuous Speech Recognition and Syntactic Processing in Iranian Farsi Language. Intl. J. Speech Technol., 1: 135-141.
12. Gharavian, D. and S.M. Ahadi, 2005. The Effect of Emotion on Farsi Speech Parameters: A Statistical Evaluation. In the Proceedings of the International Conference on Speech and Computer, pp: 463-466.
13. Gharavian, D. and S.M. Ahadi, 2006. Recognition of Emotional Speech and Speech Emotion in Farsi. In the Proceedings of International Symposium on Chinese Spoken Language Processing, 2: 299-308.
14. Gharavian, D., 2004. Prosody in Farsi Language and Its Use in Recognition of Intonation and Speech. Ph.D. Dissertation, Electrical Engineering Department, Amirkabir University of Technology, Tehran, (in Farsi).
15. Tolkmitt, F.J. and K.R. Scherer, 1986. Effect of Experimentally Induced Stress on Vocal Parameters. Journal of Experimental Psychology: Human Perception and Performance, 12: 302-313.
16. Hansen, J.H.L. and D.A. Carins, 1995. Icarus: Source Generator Based Real-Time Recognition of Speech in Noisy Stressful and Lombard Effect Environments. Speech Commun., 16: 391-422.
17. Cairns, D. and J.H.L. Hansen, 1994. Nonlinear Analysis and Detection of Speech Under Stressed Conditions. Journal of Acoustic Society of America, 96: 3392-3400.
18. Womack, B.D. and J.H.L. Hansen, 1996. Classification of Speech under Stress Using Target Driven Features. Speech Communication, 20: 131-150.
19. Altun, H. and G. Polat, 2007. New Frameworks to Boost Feature Selection Algorithms in Emotion Detection for Improved Human-Computer Interaction. Brain Vision and Artificial Intelligent. Lecture Notes in Computer Science, 4729: 533-541.
20. Altun, H. and G. Polat, 2009. Boosting Selection of Speech Related Features to Improve Performance of Multi-Class SVMs in Emotion Detection. Expert Systems with Applications, 36: 8197-8203.
21. Lee, C.M. and S.S. Narayanan, 2005. Toward Detecting Emotions in Spoken Dialogs. IEEE Transactions on Speech and Audio Processing, 13: 293-303.
22. Shami, M. and W. Verhelst, 2007. An Evaluation of the Robustness of Existing Supervised Machine Learning Approaches to the Classifications of Emotions in Speech. Speech Communication, 49: 201-212.
23. Ververidis, D. and C. Kotropoulos, 2006. Emotional Speech Recognition: Resources, Features and Methods. Speech Communication, 48: 1162-1181.

24. Rong, J., G. Li and Y.P. Chen, 1997. Acoustic Feature Selection for Automatic Emotion Recognition from Speech. Information Processing and Management, 45: 315-328.
25. Kohavi, R. and G.H. John, 1997. Wrappers for Feature Subset Selection. Artificial Intelligence, 97: 273-324.
26. Tenenbaum, J.B., V. de Silva and J.C. Langford, 2000. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science, 290: 2319-2323.
27. Dellaert, F., T. Polzin and A. Waibel, 1996. Recognizing Emotion in Speech, In the Proceedings of the International Conference on Spoken Language Processing, 3: 1970-1973.
28. Hyvarinen, A., 1999. Survey of Independent Component Analysis. Neural Computing Surveys, 2: 94-128.
29. Liu, H., H. Motoda and L. Yu, 2002. Feature Selection with Selective Sampling. In the Proceedings of the International Conference on Machine Learning, pp: 395-402.
30. Talavera, L., 1999. Feature Selection as a Preprocessing Step for Hierarchical Clustering. In the Proceedings of the International Conference on Machine Learning, pp: 389-397.
31. Han, J. and M. Kamber, 2000. Data Mining Concepts and Techniques. Morgan Kaufman.
32. Ververidis, D. and C. Kotropoulos, 2006. Fast Sequential Floating Forward Selection Applied to Emotional Speech Features Estimated on DES and SUSAS Data Collections. In the Proceeding of European Signal Processing Conference, pp: 1-5.
33. Haq, S., P.J.B. Jackson and J. Edge, 2008. Audio-Visual Feature Selection and Reduction for Emotion Classification. In the Proceeding of International Conference on Auditory-Visual Speech Processing, pp: 185-190.
34. Lee, C.M., S. Narayanan and R. Pieraccini, 2002. Combining Acoustic and Language Information for Emotion Recognition. In the Proceedings of the International Conference on Spoken Language Processing, pp: 873-876.
35. Nicholson, J., K. Takahashi and R. Nakatsu, 1999. Emotion Recognition in Speech Using Neural Networks. In the Proceedings of the International Conference on Neural Information Processing, 2: 495-501.
36. Park, C.H., D.W. Lee and K.B. Sim, 2002. Emotion Recognition of Speech Based on RNN. In the Proceedings of the International Conference on Machine Learning and Cybernetics, 4: 2210-2213.
37. Park, C.H. and K.B. Sim, 2003. Emotion Recognition and Acoustic Analysis from Speech Signal. In the Proceedings of the International Joint Conference on Neural Networks, 4: 2594-2598.
38. Sidorova, J., 2009. Speech Emotion Recognition with TGI+.2 Classifier. In the Proceedings of the EACL, Student Research Workshop, pp: 54-60.
39. Chuang, Z.J. and C.H. Wu, 2004. Emotion Recognition Using Acoustic Features and Textual Content. In the Proceedings of the International Conference on Multimedia and Expo, 1: 53-56.
40. Hoch, S., F. Althoff, G. McGlaun and G. Rigooll, 2005. Bimodal Fusion of Emotional Data in an Automotive Environment. In the Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2: 1085-1088.
41. Schuller, B., G. Rigoll and M. Lang, 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, 1: 577-580.
42. Kao, Y. and L. Lee, 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. Yacoub, S., S. Simske, X. Lin and J. Burns, 2003. Recognition of Emotions in Interactive Voice Response Systems. HP Labs, HPL-2003-136.
44. Pao, T., Y. Chen, J. Yeh and Y. Chang, 2008. Emotion Recognition and Evaluation of Mandarin Speech Using Weighted D-KNN Classification. Intl. J. Innovative Computing, Information and Control, 4: 1695-1709.
45. Petrushin, V.A., 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. Luengo, I., E. Navas, I. Hernaez and J. Sanchez, 2005. Automatic Emotion Recognition Using Prosodic Parameters. In the Proceedings of the European Conference on Speech Communication and Technology, pp: 493-496.
47. Neiberg, D., K. Elenius and K. Laskowski, 2006. Emotion Recognition in Spontaneous Speech Using GMMs. In the Proceedings of the International Conference on Spoken Language Processing, pp: 809-812.

48. Schuller, B., G. Rigoll and M. Lang, 2003. Hidden Markov Model-Based Speech Emotion Recognition. In the Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 2: 1-4.
49. Bosch, L., 2003. Emotions, Speech and the ASR Framework. Speech Communication, 40: 213-225.
50. Nwe, T.L., S.W. Foo and L.C. de Silva, 2003. Speech Emotion Recognition Using Hidden Markov Models. Speech Communication, 41: 603-623.
51. Song, M., J. Bu, C. Chen and N. Li, 2004. Audio-Visual Based Emotion Recognition- A New Approach. In the Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2: 1020-1025.
52. Song, M., C. Chen and M. You, 2004. Audio-Visual Based Emotion Recognition Using Tripled Hidden Markov Model. In the Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 5: 877-880.
53. Sim, K.B., I.H. Jang and C.H. Park, 2006. The Novel Feature Selection Method Based on Emotion Recognition System. Computational Intelligence and Bioinformatics, 4115: 731-740.
54. Gu, Y., S.L. Tan, K.J. Wong, M.H.R. Ho and L. Qu, 2008. Using GA-Based Feature Selection for Emotion Recognition from Physiological Signals. In the Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems, pp: 1-4.
55. Fleuret, F., 2004. Fast Binary Feature Selection with Conditional Mutual Information. J. Machine Learning Res., 5: 1531-1555.
56. Amir, N., 2001. Classifying Emotions in Speech: A Comparison of Methods. In the Proceedings of the European Conference on Speech Communication and Technology, pp: 127-130.
57. Cai, L., C. Jiang, Z. Wang, L. Zhao and C. Zou, 2003. A Method Combining the Global and Time Series Structure Features for Emotion Recognition in Speech. In the Proceedings of the International Conference on Neural Networks and Signal Processing, 2: 904-907.
58. Lee, C.M. and S. Narayanan, 2003. Emotion Recognition Using a Data-Driven Fuzzy Inference System. In the Proceedings of the European Conference on Speech Communication and Technology, pp: 157-160.
59. Kwon, O.W., K. Chan, J. Hao and T.W. Lee, 2003. Emotion Recognition by Speech Signal. In the Proceedings of the European Conference on Speech Communication and Technology, pp: 125-128.
60. Yu, F., E. Chang, Y. Xu and H. Shum, 2001. Emotion Detection from Speech to Enrich Multimedia Content. In Proceedings of the IEEE Pacific Rim Conference on Multimedia: Advances in Multimedia Information Processing, pp: 550-557.
61. Ayadi, M., S. Kamel and F. Karray, 2007. Speech Emotion Recognition Using Gaussian Mixture Vector Autoregressive Models. In the Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 5: 957- 960.
62. Vlasenko, B. and A. Wendemuth, 2007. Tuning Hidden Markov Model for Speech Emotion Recognition. In Proceedings of the 33rd German Annual Conference on Acoustics, pp: 317-320.
63. Gharavian, D., M. Sheikhan and M. Janipour, 2010. Pitch in Emotional Speech and Emotional Speech Recognition Using Pitch Frequency. Majlesi Journal of Electrical Engineering, 4(1): 19-24.
64. Gharavian, D. and M. Sheikhan, 2010. Emotion Recognition and Emotion Spotting Improvement Using Formant-Related Features. Majlesi Journal of Electrical Engineering, 4(4): 1-8.
65. Gharavian, D., M. Sheikhan, A.R. Nazerieh and S. Garoucy, 2011. Speech Emotion Recognition Using FCBF Feature Selection Method and GA-Optimized Fuzzy ARTMAP Neural Network. Neural Computing and Applications (Article in Press, DOI:10.1007/s00521-011-0643-1).
66. Sheikhan, M., D. Gharavian and F. Ashoftedel, 2011. Using DTW-Neural Based MFCC Warping to Improve Emotional Speech Recognition. Neural Computing and Applications (Article in Press, DOI:10.1007/s00521-011-0620-8).