JACET Journal of Advances in Computer Engineering and Technology, 3(2) 2017

Recognizing the Emotional State Changes in Human Utterance by a Learning Statistical Method based on Gaussian Mixture Model

Reza Ashrafidoost[1], Saeed Setayeshi[2], Arash Sharifi[3]

Received (2016-11-03)
Accepted (2017-05-03)

Abstract **- Speech is one of the most opulent and instant methods to express emotional characteristics of human beings, which conveys the cognitive and semantic concepts among humans. In this study, a statistical-based method for emotional recognition of speech signals is proposed, and a learning approach is introduced, which is based on the statistical model to classify internal feelings of the utterance. This approach analyzes and tracks the emotional state changes trend of speaker during the speech. The proposed method classifies utterance emotions in six standard classes including, boredom, fear, anger, neutral, disgust and sadness. For this purpose, it is applied the renowned speech corpus database, EmoDB, for training phase of the proposed approach. In this process, once the pre-processing tasks are done, the meaningful speech patterns and attributes are extracted by MFCC method, and meticulously selected by SFS method. Then, a statistical classification approach is called and altered to employ as a part of the method. This approach is entitled as the LGMM, which is used to categorize obtained features. Aftermath, with the help of the classification results, it is illustrated the emotional states changes trend to reveal speaker feelings. The proposed model also has been compared with some recent models of emotional speech classification, in which have been used similar methods and materials. Experimental results show an admissible overall recognition rate and stability in classifying the uttered speech in six emotional states, and also the proposed algorithm outperforms the other similar models in classification accuracy rates.**

Index Terms **- speech processing, emotional states, pattern recognition, mel frequency cepstral coefficient, Gaussian mixture model.**

1- Department of Computer Science, Science and Research Branch, Islamic Azad University, Tehran, Iran. (ashrafidoost@gmail.com)
2- Amirkabir University of Technology, Tehran, Iran.
3- Department of Computer Science, Science and Research Branch, Islamic Azad University, Tehran, Iran

I. INTRODUCTION

The methods of speaking have significant role in human communication, which are the natural method to express the emotions and feelings in the conversation. Besides, tone of voice is a way to express the state of speaker's emotion. When an utterance expresses a word with an emotion that makes his tone of speech change, the meaning of the word is completed.

Up to date, the emotion recognition of speech is one of the challenging fields in designing systems, which are based on intelligent human computer user interface. Accordingly, the automatic emotional recognition of human speech is one of the key areas, which has attracted a lot of attentions among researchers. These systems could recognize the feelings including uttered speech, if equipped with the intelligent emotional recognition methods and algorithms [1]. Using these kinds of systems would describe the attributes of uttered speech including mental, cognitive background, and the emotions of speaker. This approach provides the possibility for intelligent and adaptive system designers to design machines, which represent appropriate automatic reactions in accordance with the natural human needs at different situations.

Scientists, who have been working on voice and speech technology for the past four decades, have now profound understandings of the voice analysis, human speech and speech processing-based systems, which leads to develop various useful applications in this field. With the respect to the capabilities, which are provided by speech signal analysis, researchers in the field of artificial intelligence (AI), robotics and human-computer interaction (HCI) could design systems that would be useful to develop tools and machines, which are in collaboration to human natural behavior.

In this paper, it is proposed an approach that could acquire the characteristics of the utterance and his emotional changes by studying the specifications of speech signal. This information is used to recognize the conceptual attributes of speech, which associates with the voice of humans by an intelligent machine. To this end, some pre-processing tasks are done on the raw speech signal at first, and then preferred features are extracted by Mel Frequency Cepstral Coefficient (MFCC) method. Then, the features of each uttered word is extracted separately, using the Learning Gaussian Mixture Model (LGMM), as the innovative classification approach, which is based on the classic GMM technique. These emotional states, which represent the emotional states of speaker for each word during speech, are labeled and juxtaposed in according to the speech stream. At last, it is delineated the changes diagram of speaker emotional states, during the speech.

Recognition of emotion in speech and tracking its changes trend to reveal the internal feelings of speaker, is the current topic in the field of artificial intelligence, signal processing, and human-computer interaction in the recent years. To the best of our knowledge, some scientists have been working specifically on localizing emotion transition wrapped in speech signals. Most of them have investigated the acoustical features of the speech signal. In the way that, using troughs and peaks in the profile of fundamental frequency; intensity and boundaries of pauses; and energy of speech signal were the popular clues to design emotion classifiers. For instance, in 2010, S. Wu et al. studied on speech emotion recognition and using modulation spectral features (MSF), in which they extracted speech features from long term spectro-temporal representation using an auditory and modulation filter bank. So they captured acoustic and temporal modulation frequency components, and concluded that the combination of spectral and prosodic features has improved the regression results. Also, the classic discrete emotion classification and continuous emotion estimation are scrutinized in this study. They also have represented some improvements in recognition performance when used augmented prosodic features [2].

Also, In 2012, X.Anguera et al. proposed a method to detect the emotional changes of speaker, which using two consecutive fixed-lengths windows, modeling each by GMM and distance-based approaches, such as Generalized Likelihood Ratio (GLR), Kullback-Leibler (KL) divergence, and Cross Log Likelihood Ratio (CLLR), have been investigated [3].

In 2013, another study done by C.N. Van der Wal and W. Kowalczyk, who designed a system to measure changes in the utterance emotional states by analyzing the speaker voice [4]. They represented the achieved results by visualizing them in 2-D space, and also applied the Random Forest algorithm for classification and

regression problems. Their results showed some improvements in performance and error reduction in compare with similar studies, which focused on predicting changes of intensity measured by Mean Square Error (MSE).

Furthermore, the other relevant studies, which use the extraction methods of emotions from the speech signals, have used some of the methods like SVM [5], Variational Bayes, free energy and factor analysis [6], [7]. However, it seems that these methods require the large databases for testing and training phases to be effective.

The reminder of the paper is organized as follows, Section 2, reviews some important methods and approaches, which have been used as the basis of the proposed method. Section 3, introduces the database, which has been used while the test and train phases are done. In Section 4, the computational and theoretical architecture of proposed approach is presented and described the structure of proposed approach. Section 5 provides experimental results, performance measurements and comparison with the similar recent methods. Finally, Section 6 draws conclusion remarks.

II. PRELIMINARIES

1. Feature Extraction (MFCC)

Cepstrum coefficients of Mel frequency are the representation of the speech signals that extract the non-linear frequency components of the human auditory system. This method converts linear spectrum of speech signal into non-linear frequency scale that is called "Mel" [8].

At the first step of the proposed method, some pre-processing tasks are performed on the raw speech signal including windowing techniques [9]. The windowing method is performed after providing Discrete Fourier Transform (DFT) of each frame to find the spectrum scale of speech signal [10]. Then after, the frequency wrapping is used to convert spectrum of speech to Mel scale, where the triangle filter bank at uniform space is attained [11]. These filters are multiplied by the size of spectra and then obtained MFCCs. Mel-scale frequency conversion equation and the transpose equation of the Mel frequency transformation is showed in equation 1 and 2.

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (1)$$

$$M^{-1}(f) = 700 \left(\exp\left(\frac{m}{1125}\right) - 1\right) \quad (2)$$

2. Gaussian Mixture Model (GMM)

In the field of statistical science, the mixture model is considered as a probabilistic model, which is used to show the existence of the class subsets that belong to the greater population. A Bayesian model like GMM is one of the special instances of these statistical models. It is used as a successful model in various systems, particularly in the field of speech recognition and speaker identification systems. Accordingly, the Gaussian Mixture Modeling first invented by N. Day and then followed by J. Wolfe at the late 1960's [12] known as Expectation-Maximization (EM) algorithm [13]. Therefore, the main reason of using this model, in the wide range in intelligent systems, is the ability of this technique to model the data classes and the distribution form of acoustical observations [14]. To explain mathematically, the GMM likelihood function is represented in equation 3, which has been used in providing the D-dimensional feature vector.

$$F(x|\lambda_k) = \sum_{i=1}^{K} c_i f_i(x) = \sum_{i=1}^{K} c_i \mathcal{N}(x|\Phi_i) = \sum_{i=1}^{K} c_i \mathcal{N}(x|\mu_i . \Sigma_i) \quad (3)$$

In this equation, x is a weighted sum of K multivariate Gaussian components, $f_i(x)$, is D×1 for each mean vector (μi) and D×D covariance matrix (Σ_i). As shown in equation 3, λ_k stands for parameters of GMM and includes K components in order to the confined states, in which the combined weights should be satisfied by the following two conditions; $c_i \geq 0$ for i=1,…,K and $\sum_{i=1}^{K} c_i = 1$. i-th component could be written as shown in equation 4.

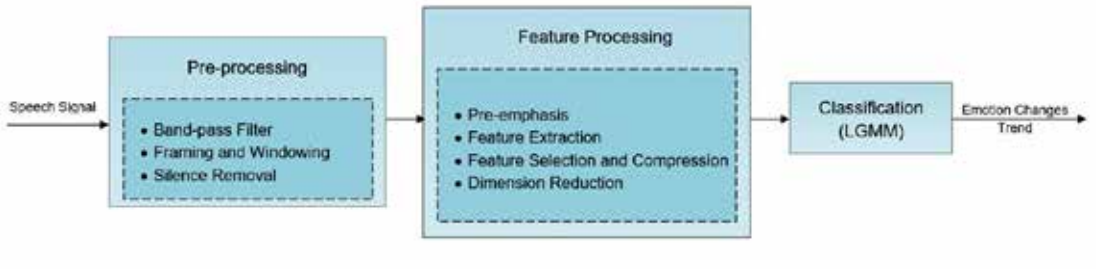

Fig. 1: Overview block diagram of emotional speech recognition routine for each word

$$f_i(x) = \mathcal{N}(x|\Phi_i) = \mathcal{N}(x|\mu_i.\Sigma_i)$$
$$= \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}} \times \exp(-\frac{1}{2}(x-\mu_i)^{Tr} \sum_{i}^{-1}(x-\mu_i)) \quad (4)$$

In equation 4, $\Phi_i = (\mu_i, \Sigma_i)$ represents the parameters for i-th Gaussian density and A^{Tr} is the transpose of matrix A. Generally, GMM could be identified by its associated parameters, the parameters are; $\lambda_k = (c_i, \Phi_i, i=1,\ldots,K)$.

III. DATA SOURCE

The emotional speech database, which is provided by Berlin University, is a standard collection of speech corpus. This database is used widely in the voice sciences and speech processing scientific resources. This database includes the audio recordings of ten actors and actresses (five males and five females), who have pronounced the sentences with six standard expressions of emotions in German. These sextet classes of emotions include anger, disgust, fear, neutral, sadness and boredom.

In this process, each actor has been requested to express one out of ten determined sentences, which has more vowels are expressed the dedicated emotion. Approximately 800 recorded sentences are used to prepare this database, and then 500 samples of them selected to choose precisely with respect to emotion recognition by human factor. This method makes it possible to select the best expressed sentences, which represent the most similar emotions to the actual natural emotions of human speakers with the relevant emotional states. Also, it performs more accurate recognition rate with more than 60% of choices to increase performance and accuracy of this database [15].

In this research, it has been used 454 enounced emotions as the emotional speech input with respect to six standard emotions which exist in the EmoDB.

IV. PROPOSED APPROACH

Various moods and emotional feelings, which are reflected in the speaker voice are represented by the exceptional patterns of acoustical features in speech signals. This means that the meaningful information, which is wrapped with emotional states of the speaker is encoded in the acoustical speech signal of the human voice. This information would be decoded and the embedded emotions, disclosed and then could be retrieved and felt once receiving by audience. Therefore, the first step to design an automatic emotion recognition system (AER) is to find out how to encode emotional states within the speech signal, which is expressed by the speaker. This task is done by extracting the most discriminator features from the speech samples in training phase. So, the classification method resolves this problem and decodes the data due to recognize the class of the particular emotional state [16].

Besides, the other approaches that commonly have been used for speech processing, have been derived from the methods that are known as pattern recognition. Above all, each moment of the speech signal stream, characterizes the encoded data, which leads to that the procedures on speech emotion recognition (SER) are closely similar to the pattern recognition cycle. Therefore, in our proposed approach, the words which are uttered in the input speech signal are

analyzed separately and are performed the routine to recognize contained emotions.

As depicted in Fig.1, the overview block diagram of the proposed model represents the main sections using in this method. The model contains pre-processing, feature processing, and classification as its prominent parts.

In the Pre-processing stage, a band-pass filter eliminates all irrelevant frequencies from the speech signal, framing and windowing procedure, converts the speech stream signal into frames, and finally the silence removal process excludes all useless (silence) frames. In the second stage, feature processing stage includes pre-emphasis, feature extraction, feature selection and compression, and dimension reduction procedures.

In the pre-emphasis section, features are categorized. This stage contains a filter 1-αz-1, in which the value of α is constant and is 0.97. This filter is derived from the lip model, and is same as high-pass filter that buttresses the high-frequency components.

Applying this filter at this step, leads the features to be extracted distinctly [16]. At the feature extraction stage, the most appropriate features, which are required for classification section are extracted. Eventually, feature compression and dimension reduction decreases the amount of redundant data and prevents curse of dimensionality problem. This task also significantly reduces the computational time for the upcoming stages. The last and foremost stage is the proposed classification approach, LGMM, as the learning statistical classifier. The details of the model are explained in the rest of this paper.

Based on the procedure of emotion classification, the changes trend of emotional states could determine the prevailed emotional feelings of the speaker during the lecture or conversation. This result is performed by the probabilistic filtering approach to increase classification accuracy. The overall view of the proposed method is illustrated in the Fig. 2.

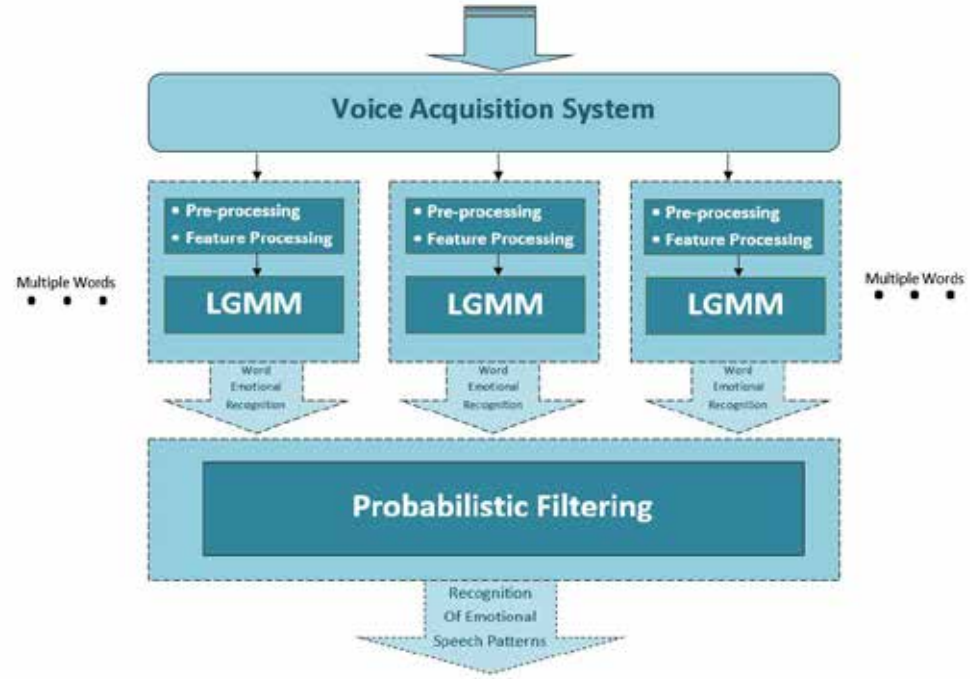

Fig. 2: Block diagram of emotional speech recognition routine

1. Feature Processing

At the first stage of our proposed approach, the pre-processing tasks are performed on the speech input signal using windowing techniques [9]. The windowing procedure is done after providing Discrete Fourier Transform (DFT) of each frame to get the spectrum scale of speech signal, in which it has been used 256 frequency points to calculate the DFT [10]. It is applied the Hamming windowing technique for framing the speech signal into long-term segments (Fig. 3). In this course of action, the 256 ms Hamming window with 64 ms frame shift, is provided to multiply by the speech signal, s(n), and it is also considered that 16 kHz sampling rate is reasonably appropriate for this work.

In this paper, it is also used 20 filter banks, and 12-MFCCs for feature extraction. Moreover, the limited number of training samples versus huge number of extracted speech features leads to many difficulties and onerous burden on classification process. Also, this issue may dramatically reduce the accuracy rate of feature classification. Consequently, applying an efficient feature selection method is inevitable and the vital part of this stage. This is performed to convert obtained coefficients into the required coefficients, which leads to decrease the size of feature vector, and prevents the curse of dimensionality at the classification process. In this way, it is applied principal components analysis (PCA) to eliminate irrelevant and meaningless features. In aftermath, it is achieved higher speed in learning and classifying process. It is also used the method called Sequential Forward Selection (SFS), which is based on an iterative algorithm. In this method, the selected feature subset is augmented before using as the classifier input [18]. Using this procedure to select the appropriate features also caused to reduce feature measurement cost and computational load in addition to reducing the curse of dimensionality.

To recap and explain computationally, SFS begins with a void subset of features, then sequentially adds features from the entire input feature space till the subset reaches a desired size. Then, the whole feature subset is evaluated at each step of iteration, except the features in the new subset. Once, the evaluation is done by the criterion function, it assesses and discovers the particular feature that leads to the highest performance enhancement of the feature subset, if it is included.

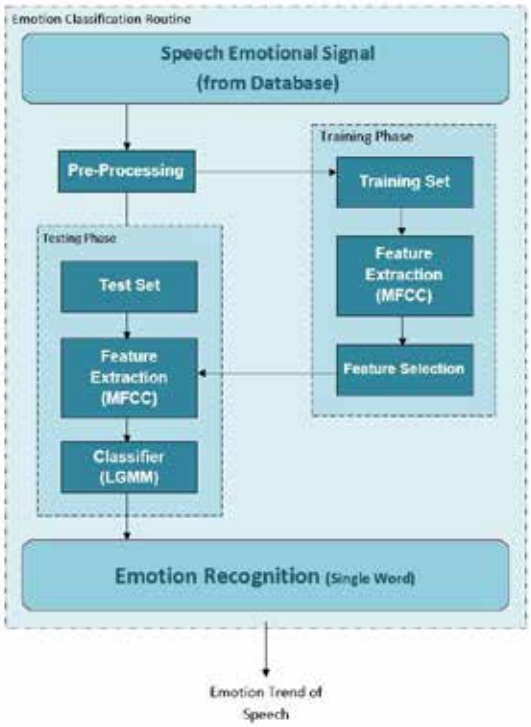

Fig. 4: Block diagram of emotion recognition procedure for single word

2. Emotion Classification (LGMM)

By using the proposed method, the emotion, which is laid in the uttered single word, is determined. The first level of emotion recognition cycle is represented in the Fig. 4. As mentioned in the preceding stage, the pre-processing tasks, including windowing, have been performed and also silent frames have been eliminated from the input speech signal. Moreover, the required features of speech signal have been extracted using MFCC method for each single uttered word. Next, feature selection procedure is performed and provided obtained features are used as an input vector to the classifier. Finally, it is used a type of Gaussian Mixture Model, which has modified to perform learning as the learning-based GMM, and we have entitled this method as LGMM [19].

In this paper, it has been proposed an enhanced derivation of Gaussian Mixture Model to provide emotion classes using combination of Gaussian densities. To give mathematical explanation, a Gaussian Mixture Model is generally a weighted sum of several Gaussian components. In other words, Gaussian Mixture Model is a linear

combination of M Gaussian densities, which is represented in the equation 5.

$$P(\vec{\mathbf{x}}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{\mathbf{x}}) \quad (5)$$

According to the latter equation, $\vec{\mathbf{x}}$ is a D-dimensional stochastic vector, $b_i(\vec{x})$ contains the density components for $i=1,...,M$ and p_i includes the combined weights for $i=1,..,M$. Each component of Gaussian function is D-dimensional and in the form of the equation 6.

$$b_i(\vec{\mathbf{x}}) = \frac{1}{(2\pi)^2|\Sigma^i|^2}\exp\left\{-\frac{1}{2}(\vec{\mathbf{x}}-\vec{\boldsymbol{\mu}}_\mathbf{i})^T\sum_{i}^{-1}(\vec{\mathbf{x}}-\vec{\boldsymbol{\mu}}_\mathbf{i})\right\} \quad (6)$$

In the equation 6, the $\vec{\boldsymbol{\mu}}_\mathbf{i}$ represents the mean vector and Σi determines the covariance matrices. Also combined weights of general probability rule, put emphasis on the concept that sum of probabilities is equal to 1 and satisfies the main statistical rule which is $\sum_{i=1}^{M} p_i = 1$.

The mathematically flexibility is the other prominent benefit of using this method of speech modeling. Intuitively, the density of complete Gaussian components can only be shown by mean vectors and covariance matrices. These components are obtained from combination of weights of all density components. Also, probability density functions of destructed features, which are affected by differences, exist in emotional specifications of those functions.

As a result, it is used a set of GMMs to calculate probability of particular emotion, which are prevailed by the utterance. This method also concludes maximum likelihood estimation, which its class-condition probability density function should be determined by providing a Bayesian classifier. For instance, the selection of the initial model could be done by using the test data, but parameter configuration of this model needs some measures of optimality such as the degree of accuracy, when the data distribution is fitted to the observed data. Accordingly, the value of data likelihood is an optimality measure. As an assumption, there is a set of independent samples such as $X=\{x_1, x_2,\ldots,x_N\}$, which is derived from a data distribution, and represented by probability density function like $p(x;\theta)$. In this function the θ stands for the set of parameters of PDF. The likelihood is represented in the equation 7.

$$L(X;\theta) = \prod_{x=1}^{N} P(x_N;\theta) \quad (7)$$

This equation denotes the likelihood of data distribution of X, or in a nutshell, it shows the data distribution of parameter θ. The main purpose of this equation is to find that $\hat{\theta}$, which would maximize value of the likelihood. It is also represented in the equation 8.

$$\hat{\theta} = \arg max_\theta = L(X;\theta) \quad (8)$$

This function most often does not reach to its maximum value, but the algorithm mentioned in the equation 9 analytically and mathematically is palpable and lucid. This equation also called likelihood function.

$$L(X;\theta) = \ln L(X;\theta) = \sum_{n=1}^{N} \ln p(x_N;\theta) \quad (9)$$

Due to uniformity of the logarithm function, a solution that mentioned in the equation 8 has similar usage as *L(X;θ)*. According to these definitions, the implementation of LGMM classifier is described. At the first step, the parameters are initialized, and then mathematical expectation is taken based on preceding probabilities for i=1,…,n and then k=1,…,K are calculated.

$$P_{i,k} = \frac{a_k^{(r)} \emptyset(x_i \mu_k^{(r)}, \Sigma_k^{(r)})}{\Sigma_{k=1}^{(r)} a_k^{(r)} \emptyset(x_i \mu_k^{(r)}, \Sigma_k^{(r)})} \quad (10)$$

Then maximization likelihood value is provided.

$$a_k^{(r+1)} = \frac{\sum_{i=1}^{n} P_{i,k}}{n} \quad (11)$$

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n} P_{i,k} X}{\sum_{i=1}^{n} P_{i,k}} \quad (12)$$

$$\mu_k^{(r+1)} = \frac{\sum_{k}^{(r+1)} P_{i,k}(x_i \mu_k^{(r+1)})(x_i \mu_k^{(r+1)})t}{\sum_{i=1}^{n} P_{i,k}} \quad (13)$$

So, as long as the data converged, steps of performing mathematical expectations and maximizations are repeated iteratively. Besides, at the first stage, this data distribution is unknown, so in the next step, the features are obtained by applying MFCC and SFS methods. These features are in the form of 12-dimensional space. Additionally, the mode of this data and the number of peaks in its distribution are unidentified. So, in this way it is initiated using a Gaussian component for each emotional class and then calculate parameters.

This phase of the proposed approach is the training phase, which the learning tasks are took place. Next, each component is divided into two chunks and retrained over and over again for each part, same as the classical "divide and conquer" method. Divisions and trainings continue repeatedly until they reach to the final number of required components. Another issue, which it is emerged using GMM, is that there is not any possible solution to train a Gaussian mixture model with C components (calculation of parameters, $\Sigma.\vec{\mathbf{x}}_\mathbf{i}.p_i$) as a Closed-form equation.

The procedure, which could provide an idea to implement our favorite training phase is an iterative algorithm to find the best parameters that is required.

This algorithm is the Expectation Maximization (EM), which is the extended version of Baum-Welch algorithm [20]. The EM algorithm have been employed to model the Probability Density Function (PDF) of the emotional speech prosody features in [21], [22]. By using this method, optimal Gaussian components are obtained, and at last the LGMM perform the training tasks successfully in the repeated iterations.

Since it has not been existed an adequate amount of data to calculate all parameters of complete covariance matrix, training of GMMs is performed using the diagonal covariance matrices. Furthermore, the samples in each class of mentioned sextet classes are randomly partitioned into 10 subsets approximately equal in size. Thereafter, each validation trial takes nine subsets from every class of training, with the remaining subset, which is kept invisible from the training phase and is used just for testing phase. It is also worth noting that the training phase is

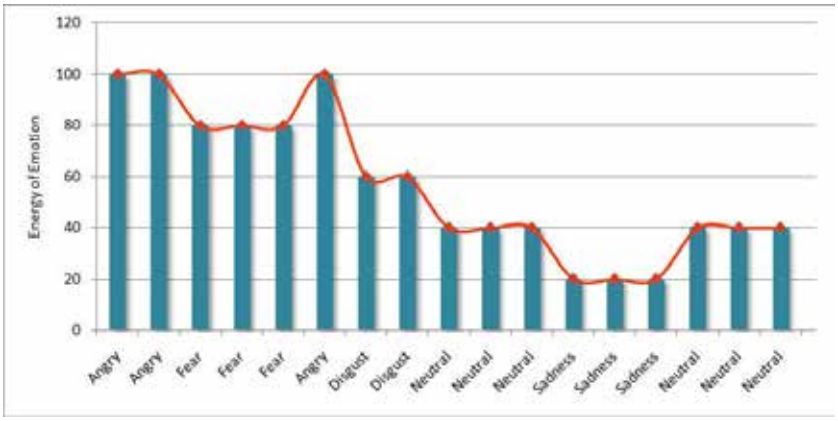

Fig. 5: Sample trend diagram of emotion states of utterance during speech

just performed once, when the application begins to run. At this stage of emotional classification, the previously mentioned steps are performed on feature vector, which has been acquired for each single word in the uttered speech signal separately. Then the utterance emotion during expressing the particular word in speech is distinguished.

At the final stage of this classification level of proposed approach, the markers of emotional classes as the emotional states are acquired to the number of the uttered words in the entire speech text. These emotional markers could be different for each uttered word due to changes of utterance emotional states, during the speech.

Finally, it is represented the emotional states changes trend of utterance during the speech. So, it is obtained a trend of emotional changes automatically, during the speech using proposed approach. This trend represents the mood changes of the speaker, when expressing talk or speech.

3. Tracking Changes of Emotions

Emotional information, which is embedded in the speech signal is derived from uttered speech input or from parts thereof, this uttered emotional information is expressive for an emotional state of a speaker and its changes [9].

The system could measure changes in the utterance emotional states by applying the proposed approach. Fig. 5 illustrates the trend of an instance speech. This diagram shows the emotional state changes and feelings of the speaker during expressing speech. As shown in the Fig. 5, it is also noticeable that the mood of the speaker changes between anger, fear, disgust, neutral and sadness during speech.

V. EXPERIMENTS

1. Test Results

In this paper, it is proposed a method for emotion recognition during the speech or talk, using the extracted features from the uttered speech signal. Considered emotional classes are based on standard classification in behavioral and speech sciences. These classes include anger, boredom, disgust, fear, neutral and sadness. Applying and testing this method on the data in EmoDB, the results are investigated using Cross-validation method and are denoted by the evaluation parameters in the form of accuracy rates.

To recall, the recognition accuracy is an evaluation method, which means how close the measured value to the actual accurate value is. This measure indicates the percentage rate of emotion recognition accuracy for each input speech signal in the test phase, to total emotional speech data in the training phase of speech recognition process [16].

$$\text{Accuracy} = \frac{\text{Correctly Recognized Speech Emotions}}{\text{Total Trained Speech Emotions}} \times 100 \quad (14)$$

Table 1: Recognition accuracy rate on EmoDB

Emotion	Classification(%)
Angry	82.94
Boredom	86.78
Disgust	83.81
Fear	80.54
Neutral	86.87
Sadness	87.52

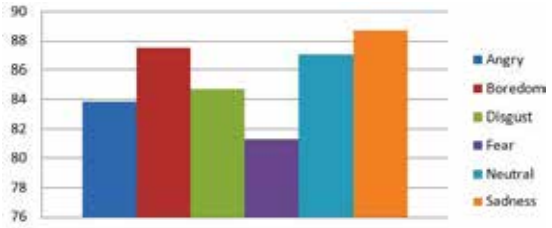

Fig. 6: Comparison chart of recognition rates obtained for each emotion

These assessments are provided for each of six emotional states in the domain of emotional classes, according to the equation 14. The results are obtained based on performance of the proposed method on Berlin emotional speech corpus (EmoDB). The acquired recognition accuracy rates are represented in the table 1.

Consequently, it is calculated the analytical

and statistical parameters, which stem from obtained result. it has been achieved to the score 2.52 as the standard deviation in accuracy rates of the six emotional states. This result shows that the proposed approach represents high degree of stability in the speech emotion recognition. Also, it is achieved the scores, 6.36, 0.0298, 6.98 and 84.74 as variance, dispersion coefficient, variation range and geometric mean, respectively. As well as, the acquired dispersion coefficient also emphasizes on the sustainability of the proposed system.

It is also clearly depicted the comparison of the recognition rates in the Fig. 6. This diagram illustrates that the highest rate of emotion recognition is achieved in evaluating emotions of anger and boredom, and the lowest one belongs to the emotion of fear. This result eventually is expected and probable because of the distinctive attributes of the emotions of anger and Boredom, and Fear. Accordingly, the attributes of these emotions, which extracted from acoustical signal of speech, are similar that conform to the same pattern. This similarity in patterns obviously can be deduced from this comparison diagram.

2. Evaluation and Comparison

There are various kinds of emotional speech recognition systems, which have been proposed based on well-known classifiers such as, Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), K-nearest neighbors (KNN), Support Vector Machine (SVM), and Artificial Neural Network (ANN), and also the combination of these methods. Each of them, has advantages and restrictions. In this paper, it is applied the GMM classifier as the basis of the proposed model, in which it is achieved reliability and stability over the other similar speech emotion recognition (SER) systems.

One of the main concerns that earlier researches came across in the emotional speech recognition systems, was that there is a large difference between accuracy rates among the results for emotions recognition. But, in according to the test results, the proposed system shows stable upshots, which obviates this chronic problem in emotional speech recognition researches.

In this results, the accuracy rates are not so far from each other and the recognition rates for all six emotions are close together. This outcome indicates that the proposed system shows the reliable results to recognize different emotions, which contained in single speech signal.

Some of similar articles have been reviewed, in which the GMM-based classifier and EmoDB are used to recognize emotions from the speech signals. First, in 2012, Ashish B. Ingale et al. proposed a method, in which they have used mixture of classification approaches, and then compared the results by applying them on the same speech corpus (EmoDB), and obtained acceptable results [23]. In this research, the GMM method has been compared with ANN-based, HMM-based and SVM-based methods and their results. In according to this article, the best performance belongs to GMM-based model, whereas in the speaker independent recognition (similar to our method) they have attained the minimum recognition accuracy rate for the best features as 78.77%. In the other study, in the year 2013, the authors proposed a model of GMM-SVM based approach to recognize emotional speech signals. According to the results, they have been achieved 78.27% of recognition accuracy for the neutral emotion in the finest situation [24].

Table 2: Recognition rates (%) comparison

Emotion / Classifier	Angry	Boredom	Disgust	Fear	Neutral	Sadness
LGMM	82.94	86.78	83.81	80.54	86.87	87.52
GMM	92	-	-	50	73	89

By reviewing the mentioned models, it shows that our proposed approach outperforms these methods in regards to the average accuracy rates.

It is worth noting that the most important contribution of our work is to propose a learning-based GMM method, which in according to the accuracy rates, its stability is the forte point of this approach. As mentioned, we have achieved the better average accuracy rate results in compare to mentioned analogous researches.

On the other hand, to emphasize on our achievements in recognition stability, we have reviewed wide range of recent similar articles in the field of emotional speech recognition, in which they have applied GMM and EmoDB as well. In 2015, R. Lanjewar et al. published an article, in which they recognized emotional states in speech using the GMM and K-NN. They also used the EmoDB for train and test phases as it has been applied in this paper. Based on their results, they have achieved good results in accuracy rates in some of emotional states such as anger and sadness, 92% and 89%, respectively [25]. There are better results in compare with our work in these two emotions, but in according to the accuracy rates for the other four comparable emotional states, our proposed approach shows the better performance in accuracy rates. The point is, the outcome, which it has been attained in all emotional states in this paper, are free of huge gaps in recognition system, whereas in the mentioned research there are the huge furrows, about 67%, between recognition rates (table 2). This advantage of our proposed approach emphasizes on stability of this method for all emotional states in speech recognition and indicates the reliability of the proposed system.

VI. CONCLUSIONS

It has been demonstrated an approach for speech emotion recognition (SER) using the innovative classification method, which is based on a probabilistic method to obtain changes trend of the speaker emotional states. To this end, it has been applied a modified version of GMM, as a basis for this approach of emotion classification, which has been entitled as the Learning Gaussian Mixture Model (LGMM). Also, it has been used 12-MFCC method to extract speech features, and have used SFS method to select the features efficiently from the raw audio signal of speech. Besides, the Berlin emotional speech corpus database (EmoDB) has been applied for training and testing the proposed method of emotion recognition. The main motivation of this research is to recognize the trend of feeling changes in emotions of speaker during the speech. A prominent advantage using this method is to depict a clear and informative view of emotional behavior of the speaker, regardless to the speech context, instant deeds or contrived behaviors during the speech, with high degree of accuracy rates. This approach benefits of using MFCC in feature extraction and its combination with SFS, which leads to more accurate results. This method of feature extraction also demonstrates acceptable performance in noisy environments. However, the recognition accuracy could be decreased a bit in the very noisy situations. Compared to the conventional methods in the field of emotional speech recognition, despite of the limited number of train and test samples in the database, the obtained results using the proposed approach allows us to achieve admissible consequences in recognition accuracy and run time.

This work could be explored and that could perhaps further improve the classification accuracy or reduce the computational complexity in the future experiments. Despite the stability of our approach, it could be enhanced to achieve higher average accuracy rates with the same admissible stability. In order to improve the performance of this emotional speech recognition system, the following potential extensions are proposed. There exist speech feature classification techniques such as Hidden Markov Model (HMM), Support Vector Machine (SVM) and Probabilistic Neural Network (PNN) proposed in recent papers, which could use to help improvements of classification part of our approach.

So, the performance and capabilities of the recognition rate could be further improved. Also, further research on the cognitive characteristics of the speech signals in emotional and conceptual classification methods could be expedient. As a final point, there are only few studies that deliberated applying multiple classifier system for speech emotion recognition. It is believed that this research direction has to be further explored.

REFERENCES

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., "Emotion recognition in human-computer interaction", IEEE Signal Processing magazine, vol. 18, no. 1, pp. 32-80, January 2001.

[2] S. Wu, T. H. Falk, W. Chan, "Automatic speech emotion recognition using modulation spectral features", Journal of Speech Communication, May, 2011, vol. 53, pp. 768–785.

[3] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, "Speaker Diarization: A Review of Recent Research", IEEE Transactions on Audio, Speech, and Language Processing. DOI: 10.1109/TASL.2011.2125954.

[4] C.N. Van der Wal, W. Kowalczyk, "Detecting Changing Emotions in Human Speech by Machine and Humans", Springer Science and Business Media, NY - Applied Intelligence, December 2013. DOI: 10.1007/s10489-013-0449-1.

[5] B. Fergani, M. Davy, and A. Houacine, "Speaker diarization using one-class support vector machines," Speech Communication, vol. 50, pp. 355-365, 2008. DOI: 10.1016/j.specom.2007.11.006.

[6] F. Valente, "Variational Bayesian Methods for Audio Indexing," PhD. dissertation, Universite de Nice-Sophia Antipolis, 2005. DOI: 10.1007/11677482_27.

[7] P. Kenny, D. Reynolds, and F. Castaldo, "Diarization of telephone conversations using factor analysis," Selected Topics in Signal Processing, IEEE Journal of, vol. 4, pp. 1059-1070, 2010.

[8] S. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Audio Speech Language Processing. 28, 357–366, 1980.

[9] Wojtek Kowalczyk, C. Natalie van der Wal., "Detecting Changing Emotions in Natural Speech", Springer Science Business Media New York, Appl Intell (2013) 39:675–691 DOI:10.1007/s10489-013-0449-1.

[10] Sara Motamed, Saeed Setayeshi, "Speech Emotion Recognition Based on Learning Automata in Fuzzy Petri-net", Journal of mathematics and computer science, vol. 12, August 2014.

[11] Rahul B. Lanjewar, Swarup Mathurkar, Nilesh Patel, "Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques," In Procedia Computer Science 49 (2015) pp. 50-57, DOI: 10.1016@j.procs.2015.04.226, 2015.

[12] J. H. Wolfe, "Pattern clustering by multivariate analysis," Multivariate Behavioral Research, vol. 5, pp. 329-359, 1970.

[13] D. Ververidis, C. Kotropoulos, "Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm," IEEE International Conference on Multimedia and Expo, Amsterdam, 2005. DOI:10.1109/ICME.2005.1521717.

[14] H. Farsaie Alaie, L. Abou-Abbas, C. Tadj, "Cry-based infant pathology classification using GMMs," Speech Communication (2015), DOI: 10.1016/j.specom.2015.12.001, 2015.

[15] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., "A database of german emotional speech", INTERSPEECH, pp.1517–1520, 2005.

[16] Yang, M.Lugger, "Emotion recognition from speech signals using new harmony features," Special Section on Statistical Signal & Array Processing, vol. 90, Issue 5, May 2010, pp. 1415–1423, DOI: 10.1016/j.sigpro.2009.09.009.

[17] A. Rabiee, S. Setayeshi, "Robust and optimum features for Persian accent classification using artificial neural network," in the proceedings of the 19th international conference on Neural Information Processing - Volume Part IV. DOI: 10.1007/978-3-642-34478-7_54.

[18] J. Kittler, "Feature set search algorithms," Journal of Pattern Recognition and Signal Process, 1978, pp. 41–60.

[19] R. Ashrafidoost, S. Setayeshi, "A Method for Modelling and Simulation the Changes Trend of Emotions in Human Speech", In Proc. of 9th European Congress on Modelling and Simulation (Eurosim), Sep.2016, p.444-450, DOI:10.1109/EUROSIM.2016.30.

[20] L. R. Welch, "Hidden Markov models and the Baum-Welch algorithm," IEEE Information Theory Society Newsletter vol. 53, pp. 1, 10-13, Dec 2003.

[21] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture," in Proc. 2004 IEEE Int. Conf. Acoustics, Audio and Signal Processing, May 2004, vol. 1, pp. 577-580.

[22] C. M. Lee, S. Narayanan, "Towards detecting emotion in spoken dialogs," IEEE Trans. Speech and Audio Processing, vol. 13, no. 2, pp. 293-303, 2005. DOI: 10.1016/j.specom.2010.08.013

[23] A. B. Ingale, D. S. Chaudhari, "Speech Emotion Recognition," International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-1, March 2012.

[24] A. S. Utane, S. L. Nalbalwar, "Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine," International Journal of Scientific & Engineering Research, Volume 4, Issue 5, May-2013.

[25] R. B. Lanjewar, S. Mathurkar, N. Patel, "Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model (GMM) and K-Nearest Neighbour (NKK) Techniques," Elsevier, Procedia Computer Science, December 2015, DOI: 10.1016/j.procs.2015.04.226.