A Method for Modelling and Simulation the Changes Trend of Emotions in Human Speech

Reza Ashrafidoost[1] Saeed Setayeshi[2]

[1]Department of Computer Science IAU, Science and Research University, Iran, r.ashrafidoost@srbiau.ac.ir

[2]Amirkabir University of Technology, Iran, setayesh@aut.ac.ir

Abstract

One of the fastest and richest methods, which represents emotional profile of human beings is speech. It also conveys the mental and perceptual concepts between humans. In this paper we have addressed the recognition of emotional characteristics of speech signal and propose a method to model the emotional changes of the utterance during the speech by using a statistical learning method. In this procedure of speech recognition, the internal feelings of the individual speaker are processed, and then classified during the speech. And so on, the system classifies emotions of the utterance in six standard classes including, anger, boredom, fear, disgust, neutral and sadness. For that reason, we call the standard and widely used speech database, EmoDB for training phase of proposed system. When pre-processing tasks done, speech patterns and features are extracted by MFCC method, and then we apply a classification approach based on statistical learning classifier to simulate changes trend of emotional states. Empirical experimentation indicates that we have achieved 85.54% of average accuracy rate and the score 2.5 of standard deviation in emotion recognition.

Keywords: emotional speech modelling, speech recognition, human-computer interaction (HCI), gaussian mixture model (GMM), Mel frequency cepstral coefficient

1 Introduction

The manners of speaking have eminent role in human communications, which are the natural methods to express the emotion and feeling in conversation. Equally important, the tone of voice is a method to express the state of emotion of the speaker. Once, an utterance expresses the word with an emotion that makes his tone of speech change, the meaning of the word is accomplished. Up to date, Emotion recognition of speech is one of the challenging fields in modeling systems which are based on human computer user interface. These systems could simulate the feelings including uttered speech, if equipped with intelligent emotional recognition techniques and algorithms (Cowie et al., 2001). Using this kind of systems would outline the attributes of uttered speech including psychological and cognitive background, and the emotions of speaker. This approach provides the possibility for intelligent or adaptive system designers to design machines, which make suitable automatic reactions in accordance with natural human needs at different situations. Evidently, one of the major areas has attracted loads of attentions to these systems is automatic emotion recognition (AER) of human speech.

Scientists, who have been working on voice and speech technology for the past four decades, have now good understandings of the voice analysis, human speech modeling and speech processing-based systems; this leads them to develop various practical applications in this field. With regard to the capabilities which provided by speech signal analysis, researchers in the field of artificial intelligence, robotics and human-computer interaction (HCI) could design machines which would be useful to develop tools and systems, which are related to human natural behavior. Some of these systems would be similar to responsive and adaptive systems, speech production, speech simulation, evaluations systems, security and surveillance systems, speaker recognition systems, human-robot interactive systems (HRI), and generally the environments which are equipped with smart workplace systems. To achieve this purpose, it is needed to automate data collection from users to get optimal performance of these systems and could perform services real-time and compatible with user's needs.

In this paper, we propose an approach which could acquire the attributes of the utterance and his emotional changes by studying the patterns of speech signal. This information is used to recognize the conceptual characteristics of speech which wrapped in the voice of humans by an intelligent machine. In this study we apply speech corpus of utterances from EmoDB as input, and then processing of speech signal recognition begins, some pre-processing tasks are performed on raw speech signal, and then desired features are extracted by Mel Frequency Cepstral Coefficient (MFCC) method. Then, the attributes of each uttered word of speech is elicited

DOI: 10.3384/ecp17142479

separately by the Learning Gaussian Mixture Model (LGMM), as the innovative classification approach. These emotional states, which stand for the emotional state of speaker for each word during speech, are labeled and arranged side by side in according to speech stream. To end with, we could delineate the trend of emotional states of speaker during speech or conversation.

By using the proposed approach, the emotional states of utterance are classified in six standard emotional classes. These classes include, anger, boredom, fear, disgust, neutral and sadness. Despite the context of expressed talk during the speech, the system could detect and track the trend of credible internal emotional states of utterance. Emotional speech recognition using this method of classification provides the precise results and high accuracy in emotion recognition and its changes trend. The most prominent goal of this article is to propose an approach based on an innovative learning method of Gaussian Mixture Model in emotional recognition of speech to extract internal emotions and feelings of the speaker. This, is performed by processing the speech signal, and then represent changes trend of emotional states. This approach of speech patterns processing, could be used in intelligent systems which closely interact with users to predict the emotional states of them. The systems which equipped with this capability could be used in the fields like medical, educational, surveillance systems or intelligent work places.

The reminder of the paper is structured as follows, Section 2, delineates some of recent related works in this field and their specifications. Section 3, illustrates the overall view of methods and concepts using in this article; including Emotion recognition; feature extraction method, MFCC; and the Gaussian Mixture Model. Section 4 introduces the database we have used during the test and train phases (Burkhardt et al.,2005). In Section 5, the computational architecture of proposed approach is presented and described the structure of method. Section 6 provides tentative results and performance measurements, and finally Section 7 draws conclusion remarks.

2 Related Works

Recognition of emotion in speech and tracking its changes trend to disclose 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 researchers have focused specifically on localizing emotion transition wrapped in speech. Most of them focused on acoustical features of the speech signal. For example, using troughs and peaks in the profile of fundamental frequency; intensity and boundaries of pauses; and energy of signal were the popular clues to design emotion classifiers. So, we are focusing on some of recent outstanding researches in this field and briefly investigate their specifications.

Anguera et al. (2011) proposed an approach to detect speaker change, which using two consecutive fixed-lengths windows, modeling each by Gaussian Mixture Model and distance-based methods, such as Generalized Likelihood Ratio (GLR), Kullback-Leibler (KL) divergence, and Cross Log Likelihood Ratio (CLLR) have been investigated. In 2013, another study performed by C.N. van der Wal and W. Kowalczyk who proposed a system to measure changes in the emotional states of the utterance automatically by analyzing voice of speaker. They represented the obtained results by visualizing them in 2-D space. In this study the Random Forest algorithm was applied for classification and regression problems. Their results show 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). They also claimed that the proposed system performs to classify negative emotions and provides better performance.

Besides, in the other studies for extracting emotion from speech, a number of useful methods like SVM (Fergani et al., 2008), Variational Bayes free energy (Valente, 2005) and factor analysis (Kenny et al.,2010) have used. However, it seems that these methods require large databases for testing and training phases to be effective.

3 Preliminaries

3.1 Emotion Recognition

Different moods and emotional feelings reflected in the voice of speakers are represented by the special patterns of acoustical features in speech signals. This means that the worthwhile information wrapped with emotional states of the utterance is encoded in acoustical speech signal of the voice of speaker. This information would be decoded and then embedded emotions disclosed and could be perceived and feel once receiving by audiences. Therefore, the first step to design the automatic emotional recognition systems is to find out how to encode the emotional states which expresses by the speaker in the speech. This work is done by extracting the most discriminator features from speech samples in training phase. Then the classification

DOI: 10.3384/ecp17142479

method resolves this issue and decodes the data in order to recognize the class of particular emotional state (Yang and Lugger, 2009).

3.2 Feature Extraction (MFCC)

Cepstrum coefficients of Mel frequency is the representation of the speech signals which extracts the non-linear frequency components of the human auditory system. This method converts linear spectrum of speech signal to non-linear frequency scale which is called "Mel". At the first stage of our proposed method, pre-processing tasks are performed on the raw speech input signal using windowing techniques (Kowalczyk and van der Wal, 2013). The windowing is done after providing Discrete Fourier Transform (DFT) of each frame to obtain the spectrum scale of speech signal (Motamed, 2014). Then, frequency wrapping is used to convert spectrum of speech to Mel scale where the triangle filter bank at uniform space is achieved (Rahul et al., 2015). These filters multiplied by the size of spectra and eventually obtained MFCCs. In this paper 20 filter banks and 12-MFCC are used for feature extraction. Mel-scale frequency conversion equation is determined in

$$M\ (f) = 1125\ \ln(1 + \frac{f}{700}) \qquad (1)$$

and the transpose equation of Mel frequency transformation is showed in

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

3.3 Gaussian Mixture Model (GMM)

In statistical sciences, the mixture model is considered as a probabilistic model which is used to represent existence of the subsets of classes which belong to the larger population. A Bayesian model like GMM is one of the special cases of these statistical models. GMM modeling technique is straightforward but so efficient. Therefore, these capabilities are significant due to its ability of forming soft approximations and curved shapes of any form of distribution in random data. This model is used as a successful model in different systems, especially in the field of speech recognition and speaker identification systems.

Accordingly, the Gaussian Mixture Modeling first invented by N. Day and later by J. Wolfe at the late 60's (Wolfe, 1970) known as Expectation-Maximization (EM) algorithm (Ververidis and Kotropoulos, 2005). Hence, the main reason of using this model in the wide range of intelligent systems is the ability of this technique to model the data classes or the distribution form of acoustical observations of the speaker (Alaie et al., 2015). According to

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

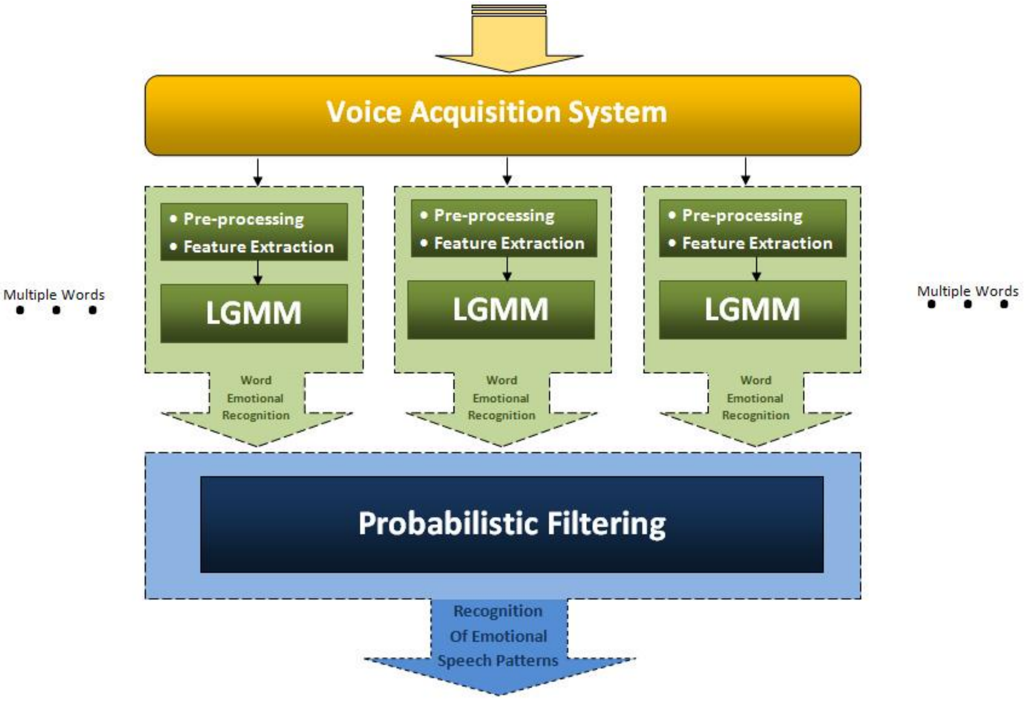

Figure 1. Block diagram of emotional speech recognition routine.

DOI: 10.3384/ecp17142479

in the GMM likelihood function which has been used for D-dimensional feature vector, x is a weighted sum of K multivariate Gaussian components, $f_i(x)$, is D×1 for each mean vector (μ_i) and D×D covariance matrix (Σ_i).

In (3), λ_k represents parameters of GMM and include K components in order to the restricted 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

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

In (4) Φ_i= (μ_i,Σ_i) represents the parameters for i-th Gaussian density and A^{Tr} is the inversion of matrix A. In general, GMM could be identified with its associated parameters, the parameters are; λ_k = (c_i,Φ_i, i=1,…,K) .

4 Database

The emotional speech database which is provided by Berlin University is a standard collection of speech corpus, which is used widely in voice sciences and speech processing scientific resources. This database includes audio recordings of ten actors and actresses (five males and five females) who have pronounced sentences with seven standard classes of emotions in German. These seven classes of emotions include anger, disgust, fear, happiness, neutral, sadness and boredom. In this process, each actor has been asked to express one out of ten predetermined sentences which has more vowels with dedicated emotion (Burkhardt et al., 2005). 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 factors. This method makes it possible to select best sentences which represent the most similar emotions to real natural emotions of speakers with particular emotional states. Also, it performs more accurate recognition with precision higher than 80% and natural selection with more than 60% of choices to increase performance and accuracy of this database (Burkhardt et al., 2005). In this experience we have used 454 enounced emotions with respect to sextet standard emotions which exist in EmoDB.

5 Proposed Approach

The approaches, which commonly are used for speech processing, have derived from the methods that are known as pattern recognition. In particular, each moment of speech signal stream, represents the encoded data which leads to that the analytic works on speech emotion recognition (SER) are closely similar to pattern recognition cycle. To begin with, the words uttered in the input speech signal are analyzed separately and performed the routine to emotion recognition. Then the changes in trend of emotional states determine the prevailed emotional feelings of utterance during the lecture or conversation. This result is performed by the probabilistic filtering method to boost up classification accuracy. The overall view of the proposed approach illustrated in the Figure 1.

5.1 Emotion classification (LGMM)

By using this method, the emotion that is laid in uttered single word, is determined. The main purpose of speech processing by this approach is to recognize emotional states of the speaker, and model the trend of its changes during the long speech. The first level of emotion recognition cycle represents in Figure 2.

At this stage, the pre-processing tasks, including windowing are performed and also silent frames are removed from the input speech signal, and then required features of speech signal are extracted using MFCC method for each single word. In the next step, feature selection is performed to convert obtained coefficients into the required coefficients. This causes to decrease the size of feature vector and prevents curse of dimensionality at the classification process. Then these features are used as an input vector to the classifier. We use a type of Gaussian Mixture Model, which we have modified it to perform learning as the leaning-based GMM. We have called this method as LGMM.

In this paper we propose an expanded derivation of Gaussian Mixture Model to provide classes of emotions using combination of Gaussian densities. The motivation which convinced us to use this type of GMM was that Gaussian components can represent some of spectral shapes of speech signal which depend to general emotions of utterance. Another reason is that the capabilities of Gaussian combinations are so reasonable for stochastic density modeling similar to modeling of speech signals.

To describe mathematically, 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

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

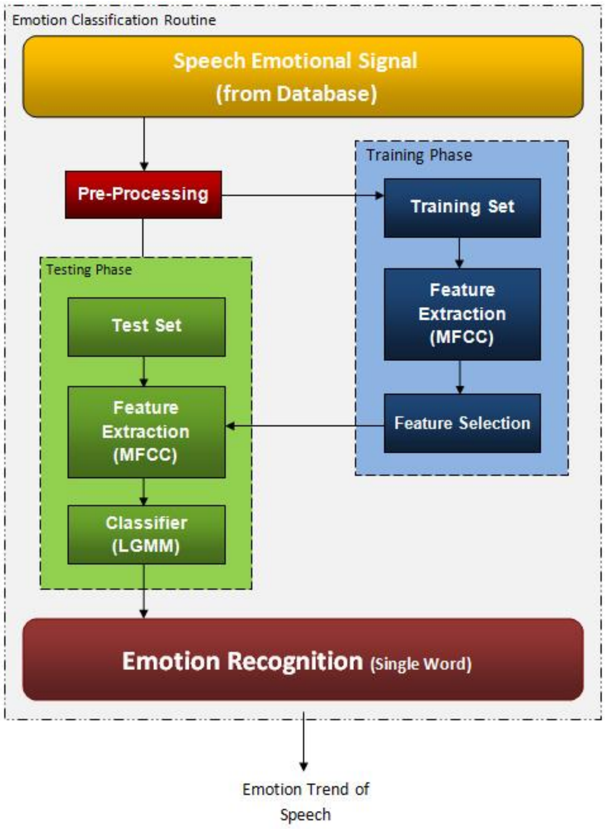
Emotion Classification Routine
Speech Emotional Signal (from Database)
Pre-Processing
Training Phase
Training Set
Feature Extraction (MFCC)
Feature Selection
Testing Phase
Test Set
Feature Extraction (MFCC)
Classifier (LGMM)
Emotion Recognition (Single Word)
Emotion Trend of Speech

Figure 2. Block diagram of emotion recognition of proposed method for single word of speech.

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

$$b_i(\vec{\mathbf{x}}) = \frac{1}{(2\pi)^2|\Sigma^{\mathrm{i}}|^2}\exp\left\{-\frac{1}{2}(\vec{\mathbf{x}}-\vec{\upmu}_{\mathbf{i}})^T\textstyle\sum_i^{-1}(\vec{\mathbf{x}}-\vec{\upmu}_{\mathbf{i}})\right\} \qquad (6)$$

the $\vec{\mu}_{\mathbf{i}}$ represents the mean vector and Σ_{i} determines the covariance matrices. Also, combined weights of general probability rule, emphasize the concept that sum of probabilities is equal to 1 and satisfy the main statistical rule which is $\sum_{i=1}^{M} p_i = 1$.

The mathematically flexibility is the prominent advantage 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, we could use a set of GMMs to calculate probability of particular emotion which are prevailed by utterance. This method also concludes maximum likelihood estimation which should be determined a class-condition probability density function by providing a Bayesian classifier. For instance, the selection of initial model could be done by using 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. Just suppose we have a set of independent samples such as X={x₁, x₂,…,x_N} derives from a data distribution which is represented by probability density function like p(x;θ). In this function the θ is the set of parameters of PDF. The likelihood is represented in

$$\mathrm{L}(\mathrm{X};\,\theta)=\textstyle\prod_{\mathrm{x}=1}^{\mathrm{N}}\mathrm{P}(\mathrm{x}_{\mathrm{N}};\theta) \qquad (7)$$

This equation represents 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 $\hat{\theta}$ would maximize value of likelihood. We also have in

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

This function most often does not reach to its maximum value, but the algorithm mentioned in (9) analytically and mathematically is evident and clear. This equation also called likelihood function:

$$L(X;\theta) = \ln L(X;\theta) = \textstyle\sum_{n=1}^{N} \ln p(x_{\mathrm{N}};\theta) \qquad (9)$$

Due to uniformity of the logarithm function, a solution that has mentioned in (10) has similar usage to $L(X;\theta)$. According to these definitions, implementation steps of LGMM classifier is as described underneath. At the first point, the parameters are initialized, and then mathematical expectation is taken based on previous probabilities for i=1,…, n and then k=1,…,K are calculated by

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

Then maximization likelihood value is provided by

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

DOI: 10.3384/ecp17142479

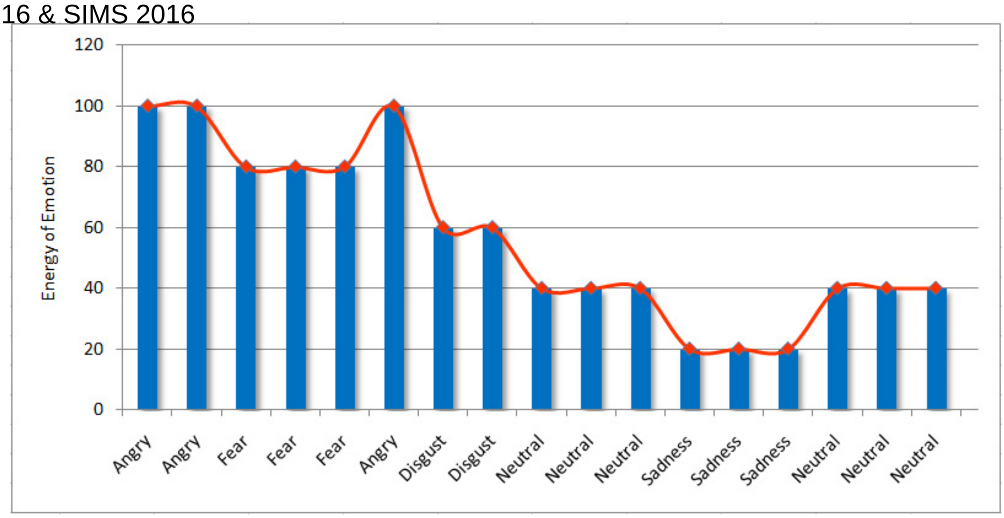

Figure 3. Sample trend diagram of emotion states of utterance during speech.

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

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

And as long as the data converge, steps of getting mathematical expectations and maximizations repeat iteratively. Besides, this data distribution is unknown for us at first, which in the next step; the features are obtained by applying MFCC method. These features are in the form of 12-dimensional space. It is also unknown for us the mode of this data and the number of peaks in its distribution.

So, in this way we begin using a Gaussian component for each emotional class and then calculate parameters. This phase of proposed approach called training phase which the learning tasks take place. Next, each component is divided into two parts and retrained repeatedly for each part, same as classical “divide and conquer” renowned method.

Divisions and trainings continue repeatedly until they reach the final number of required components. Another issue which we had faced 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 EM algorithm was used to model the Probability Density Function (PDF) of the emotional speech prosody features in (Schuller, 2004; Lee, 2005). By using this method, optimal Gaussian components are obtained at last in repeated iterations and training task of LGMM performed successfully.

Since we do not have enough data to calculate all parameters of complete covariance matrix, training of GMMs is performed using diagonal covariance matrices. It is also worth noting that the training phase just performs once when the application begins to run.

At this stage of emotional classification, all previously mentioned steps perform on feature vector, which obtained for each single word in uttered speech signal separately. Then the emotion of utterance during expressing the particular word in speech is recognized. At the final stage of this classification level of proposed approach, the labels of emotional classes as the emotional states are obtained to the number of uttered words in the whole speech. These emotional labels could be different for each uttered word due to the changes of emotional states of utterance during the speech or conversation. Finally, we depict the changes trend of the emotional states of utterance during the speech.

5.2 Trend of Changes in Emotional States

Emotional information which is embedded in speech signal is derived from expressed speech input or from parts thereof, this uttered emotion information being descriptive for an emotional state of a speaker and its changes (Kowalczyk and van der Wal, 2013).

Next, we have obtained a trend of emotional changes automatically, during the speech using proposed method. Modeling and simulation of this trend shows the changes of feelings of the speaker when expressing talk or speech. The system could measure changes in emotional states of the utterance by applying the proposed approach. Figure 3 illustrates the trend of an instance speech, which shows the changes of emotional states and moods of the speaker during expressing speech. In Figure 3, it is also obvious that the mood of the speaker changes between anger, fear, disgust, neutral and sadness during speech.

6 Experiments

In this paper, we propose a method for emotion recognition of utterances during the speech or talk using the extracted features of 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 the developed method and test it on the data in EmoDB, the results are investigated using Cross-validation method and represented with evaluation parameters in form of accuracy.

To remind, 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 test phase to total emotional speech data in training phase (Yang and Lugger, 2009). These assessments are provided for each of six emotional states in the domain of emotional classes. The results are obtained based on performance of proposed method on Berlin emotional speech database (EmoDB). The recognition accuracy rates are represented in Table 1.

In consequence, we have calculated analytical and statistical parameters which stem from obtained result. We also do have achieved the score 2.52 as the standard deviation for accuracy rates of the six emotional states. This result shows that our approach represents high degree of stability in emotion recognition from the speech. Also, we have achieved scores, 6.34, 0.0294, 7.42 and 85.50 as variance, dispersion coefficient, variation range and geometric mean, respectively. As well as, the acquired dispersion coefficient also emphasized on the sustainability of system.

Table 1. Recognition accuracy rate on EmoDB.

Emotion	*Classification*
Angry	83.86
Boredom	87.56
Disgust	84.69
Fear	81.32
Neutral	87.08
Sadness	88.74

7 Conclusions

We have demonstrated an approach for speech emotion recognition (SER) and modeling its emotional changes of the speaker using the innovative classification method, which is based on a probabilistic method. To this end, we applied a modified version of GMM as a basis for this approach of emotion classification, which we have named it as Learning Gaussian Mixture Model (LGMM). We have used 12-MFCC to extract features from the raw audio signal of speech. We also have used Berlin emotional speech corpus database (EmoDB) for training and testing the proposed method of emotion recognition. Due to the admissible results in recognition accuracy rates, which are obtained using the proposed method, we do offer this method to design and develop the emotional speech recognition-based systems, specially the systems which need to anticipate human behavior. The main motivation of this research is to simulate the trend of changes in feelings and emotions of speaker during the speech. A prominent advantage using this method is to depict a view of emotional behavior of the speaker regardless to speech context, instant events or factitious behaviors during the speech or conversation. Also, we have applied benefits of using MFCC for feature extraction, which this leads to more accurate results. This method of feature extraction also demonstrates good performance in noisy environments; however, the recognition accuracy could be a little decreased in the very noisy environments. 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 method allow us to achieve state-of-the-art consequences in recognition accuracy and run time.

References

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

F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss. A database of german emotional speech. *INTERSPEECH,* 12:1517–1520, 2005.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing magazine*, 18(1):32-80, 2001.

Hesam Farsaie Alaie, Lina Abou-Abbas, and Chakib Tadj. Cry-based infant pathology classification using GMMs. *Speech Communication*, DOI:10.1016/j.specom.2015.12.001, 77:28-52, 2015.

B. Fergani, M. Davy, and A. Houacine. Speaker diarization using one-class support vector machines. *Speech*

Communication, 50:355-365, 2008, DOI:10.1016/j.specom.2007.11.006.

P. Kenny, D. Reynolds, and F. Castaldo. Diarization of telephone conversations using factor analysis. *Selected Topics in Signal Processing, IEEE Journal of,* 4:1059-1070, 2010.

Wojtek Kowalczyk and C. Natalie van der Wal. Detecting Changing Emotions in Natural Speech. *Springer Science and Business Media New York, Appl Intell,* 39:675–691, 2013, DOI: 10.1007/978-3-642-31087-4_51.

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

C. M. Lee and S. Narayanan. Towards detecting emotion in spoken dialogs. *IEEE transaction Speech and Audio Processings,* 13(2):293-303, 2005.

Sara Motamed and Saeed Setayeshi. Speech Emotion Recognition Based on Learning Automata in Fuzzy Petri-net. *Journal of mathematics and computer science,* 12:173-185, 2014.

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 Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. (ICASSP '04).*

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

C.N. van der Wal and W. Kowalczyk. Detecting Changing Emotions in Human Speech by Machine and Humans. *Springer Science and Business Media, NY - Applied Intelligence,* 39(4):675-691, 2013, DOI: 10.1007/s10489-013-0449-1.

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

L. R. Welch. Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter* 53(1):10-13, 2003.

J. H. Wolfe. Pattern clustering by multi variant analysis. *Multivariable Behavior Res.,* 5:329-359, 1970.

Yang and M. Lugger. Emotion recognition from speech signals using new harmony features. *Special Section on Statistical Signal and Array Processing,* 90(5): 1415–1423, 2010, DOI: 10.1016/j.sigpro.2009.09.009.