Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Applied Sciences

ISSN 1812-5654
DOI: 10.3923/jas.2016.406.418

Research Article
Modified-VQ Features for Speech Emotion Recognition

Hemanta Kumar Palo and Mihir Narayan Mohanty

Department of Electronics and Communication Engineering, Siksha 'O' Anusandhan University Bhubaneswar, Odisha, India

Abstract

Objective: Features of the signal has major role for recognition, classification and detection task. Less number of features for effective recognition is the challenge that motivates the authors to proceed in this respect. In this study, a modified Vector Quantized (VQ) feature for emotional speech recognition has been proposed. **Methodology:** The proposed feature is based on statistical VQ and differential VQ statistics of frame-level prosodic features derived at utterance level. Further, the combination of frame-level baseline features, VQ based frame-level prosodic features and modified VQ prosodic features at utterance level are compared and analyzed. Neural network based classifiers as multilayer perceptron (MLP) and Radial Basis Function Network (RBFN) has been tested with proposed combinations. Standard Berlin emotional (EMO-DB) database and a locally collected emotional speech database have been used for validation of the methods. **Results:** The modified VQ feature combinations outperformed all other feature combinations in terms of classification accuracy and Mean Square Error (MSE). **Conclusion:** Result exhibited highest accuracy of 91.08% with RBFN and 89.93% with MLP classifiers respectively with modified VQ based feature combination for EMO-DB database. As against it the recognition was 90.38 and 88.05% with VQ based prosodic feature combination and 85.79 and 84.04% with frame level prosodic feature combination, respectively.

Key words: Speech emotion recognition, feature combination, vector quantization, radial basis function network, multilayer perceptron

Received: May 30, 2016 Accepted: July 15, 2016 Published: August 15, 2016

Citation: Hemanta Kumar Palo and Mihir Narayan Mohanty, 2016. Modified-VQ features for speech emotion recognition. J. Applied Sci., 16: 406-418.

Corresponding Author: Hemanta Kumar Palo, Department of Electronics and Communication Engineering, Siksha 'O' Anusandhan University Bhubaneswar, Odisha, India Tel: 8895138864, 9437056742

Copyright: © 2016 Hemanta Kumar Palo and Mihir Narayan Mohanty. This is an open access article distributed under the terms of the creative commons attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Human expressions from faces, physical movements, voices and cultural artifacts can be used for emotion perception and recognition[1]. Among these, recognition using diversified spoken language is often complex hence least analyzed. Many application areas such as human resource management, criminal investigation, bio-medical engineering, banking and finance, gaming and computer tutoring mechanism etc., rely on speech emotions more often to remain viable. It requires an effective recognition model to enhance their applicability. However, absence of distinct and reliable features, genuine database, a well-defined theoretical platform that relates human being to emotions and efficient classifier makes these emotional models more complex. Hence forth, developing a reliable recognition system remains an ever-growing challenge for researchers. Arguably, the claimed accuracy is widely varied among literatures based on these factors[2-15]. It poses further ambiguities in terms of recognition authenticity.

Anger, boredom, happiness, fear, surprise and neutral are mostly discussed emotional categories in this area[2-4,8-26]. A review on related literatures suggests for an effective use of features, which will describe these speech emotions adequately[15,22,26]. Either features extracted at frame-level or at utterance-level are often attempted[2,9,12,14,16-36]. Nevertheless, the recognition of these speech emotions is below 70% with individual features tested alone[6,15,23-24,28,30,36]. It made speech engineers to explore effective feature reduction and combination mechanisms as an alternative for enhanced performance[2-3,15,22,32]. Few conventional feature reduction mechanisms as statistical techniques, Principle Component Analysis (PCA), vector quantization has been often discussed in literatures[2,5,11,15-16,18,20,22,25,33-39]. Use of PCA based reduced features has claimed accuracy of 40.7% (for male) and 36.4% (for female)[16], below[23] 73 and 79.9%[4] for emotional speech recognition based on acoustic features. However, it is not possible to guarantee the projected variables in a physical sense using PCA. This makes the selected variables difficult to interpret. Further, reduced features using PCA tends to be less reliable[25]. Statistical methods have been quite successful in reducing features for effective recognition[15,20]. Nevertheless, loss of temporal information in describing speech emotion makes the method less efficient[2,36]. Vector Quantization (VQ), has been successfully employed for speech coding, speaker recognition and emotional speech detection[11,22,39]. Further, reduced features using VQ tend to be superior to both linear and non-linear PCA based features[5]. Hence, VQ method of feature reduction along with statistical techniques are our preferred choice, hence opted in this study.

Possible combination using effective individual feature has been another major source for enhanced accuracy. Prosodic feature combination has provided 62.7% accuracy[39]. The authors suggested use of VQ method of reduction along with MFCC with an improved accuracy of 71.7% in their study. Reported accuracy of 70% with spectral and prosodic combination[6], 83.55% with acoustic, semantic and prosodic combination[37], 95.55% with VQ based spectral and time-frequency feature combination[22], 88.7% with spectral combination[21], 83.1 and 92% using linguistic and acoustic level fusion[31] are few works that motivated the authors in this direction. However, further improvement in performance was experienced in our earlier effort involving both feature reduction and combination mechanism together[22]. However, to determine a possible combination among hundreds and even thousands of available features is quite cumbersome. Literature favors for possible combination of features that provide complementary information[15,22]. We have approached our problem based on these factors. The benefits of both reduction and combination concept are explored. Selection of the classifier is judiciously done after extensive study in this area to make the classifier conducive with our proposed feature extraction methods.

Gaussian Mixture Model (GMM) classifiers are capable of generating adequate model for emotions involving large feature sets[2]. However, for reduced feature sets neural network based classifiers outperform both GMM and HMM[2,22,31]. Hence, to maintain compatibility with our proposed reduced features, NN based classifiers have been chosen for this study.

Following are the few issues investigated in this piece of work. Firstly, VQ has been used to reduce the chosen prosodic features. Secondly, VQ is modified with its statistical and differential values to further reduce the features. However, it is observed that individual reduction mechanism has not improved the recognition much as reported in[2,6,15,26,33]. Hence, modified reduction mechanism based on statistics of both VQ and differential VQ are used to derive the proposed features. Use of differential VQ has enhanced feature robustness by involving temporal dynamic that is lost during statistical analysis. Thirdly, the mechanism used is validated in a cross corpus platform. For the purpose, a locally collected database has been tested along with standard Berlin EMO-DB database using our propose features. Fourthly, selection of efficient reduction and classifier design parameter for optimum efficiency has been iteratively made. Finally, possible combinations based on similar natured features both at frame-level and utterance level including our proposed set up is made.

MATERIALS AND METHODS

Proposed methods: From literature survey, it is clear that combination of features can enhance the recognition level[6,21-22,37,39]. However, it also suggest for possible combination using feature bearing similar characteristics[15,22]. The feature chosen should provide complementary information or else should provide supplementary in nature. Hence, in this study, few prosodic features as Zero Crossing Rate (ZCR), fundamental frequency (F0), autocorrelation coefficients and the STE (short time energy) bearing either complementary or supplementary information are selected for possible combinations. Researchers have applied different reduction techniques to a single database in most cases. In this case, although some trends may surface naturally but a consensus among designated features are quite difficult to prove[32]. The reason may be attributed to the reduced features, which may vary among different databases.

This study includes the proposed modified-VQ feature extraction techniques and the classification scheme used for efficient recognition of selected speech emotions with effective design parameters. The proposed scheme involves following steps:

Proposed feature extraction technique: In this study, 3 feature extraction techniques are used. Further, the suitable combinations are tested for better accuracy. Finally, the comparisons among all these techniques are shown in this study. These are:

- Baseline frame-level feature combination
- Reduced VQ based frame-level feature combination
- Modified VQ utterance-level feature combination using both statistical VQ and differential VQ feature statistics

Baseline features: The frame-level features of basic prosodic features are considered and those are Zero Crossing Rate (ZCR), fundamental frequency (F0), autocorrelation coefficients and the Short Time Energy (STE). These are extracted by considering the frame-level signal.

Energy of the emotional speech signal indicates the arousal level and presence of higher frequency component in the signal. It is informative for recognition of speech emotions. To approximate the non-stationary speech signal x(n) into a quasi-periodic signal at short interval, STE has been computed here. The STE is found from:

$$E_{ST}(m) = \sum_{m=0}^{M-1} |x(m)|^2 \quad (1)$$

where, m = 1, 2,..., M is total number of features in a frame.

In order to account for the periodicity information of emotions in speech signal, auto-correlation coefficients (ACF) has been quite useful. Further, these features will provide complementary energy information for effective feature combination. These coefficients can be estimated with a time lag τ as:

$$S(\tau) = \frac{1}{M}\sum_{n=0}^{M-1} x(m)x(m+\tau) \quad (2)$$

The F0 has been mostly effective in describing human speech emotions[2,26,33]. Auto-correlation method has been used to find F0 in this study since it is a robust, simple and more reliable method[33]. The ACF will attain its maximum value at x(m) = x(m+τ). Peaks are obtained at τ = IT for a signal x(m) with period T where, I is an integer. Among ACF, lower peaks are manifested with increase in with highest value being observed for S(τ) = S(0). Therefore, the F0s can be computed at τ = T by finding the location at which the peak exists. The F0 feature is extracted from whole utterance of an emotional class in this study.

The transition of the emotional speech signal around zero axis can be additional information for recognition of the signal. The information can be supplementary in nature for efficient feature combination. This information can be obtained using ZCR as given by:

$$Z = \sum_{m=1}^{M} \text{sgn}\,(x(m)) - \text{sgn}\,(x(m-1)) \quad (3)$$

In extracting these mentioned features, a frame width of 30 ms overlapped with 10 ms has been chosen here. Windowed signal from the frames are obtained using Hamming window to remove edge effect and prevent signal loss. This widow has approximately twice the bandwidth compared to rectangular window hence preferred here[10]. For any baseline features, the number of features in an utterance is given by:

$$N_1 = M \times N \quad (4)$$

where, m = 1, 2,..., M denotes number of features in a frame of an utterance and n = 1,2,...,N = number of frames in an utterance. For U numbers of utterances of an emotional class, a total of M×N×U features are obtained. The STE, ZCR, ACF frame-level features of an emotional class thus obtained can be represented as B = {E_{STn}(u), Z_n(u), S(τ)n(u)}, u = 1,2,..., U.

Different combinations have been used from these prosodic baseline features. The combined feature sets are represented as B = {B_1,B_2,...,B_{11}} and is given below:

$$B_1 = \{E_{STn}(u), Z_n(u)\}, B_2 = \{E_{STn}(u), F0\},$$
$$B_3 = \{E_{STn}(u), S(\tau)_n(u)\}, B_4 = \{Z_n(u), S(\tau)_n(u)\},$$
$$B_5 = \{F0, S(\tau)_n(u)\}, B_6 = \{F0, Z_n(u)\},$$
$$B_7 = \{E_{STn}(u), Z_n(u), F0\}, B_8 = \{E_{STn}(u), F0, S(\tau)_n(u)\},$$
$$B_9 = \{S(\tau)_n(u), Z_n(u), F0\}, B_{10} = \{E_{STn}(u), Z_n(u), S(\tau)_n(u)\},$$
$$B_{11} = \{E_{STn}(u), F0, Z_n(u), S(\tau)_n(u)\}$$

The combination mechanism to be effective needs features of similar nature[15,22]. Hence, prosodic features providing energy information such as STE and ACF are combined. Pitch and ZCR provide supplementary information hence will increase the available information. Use of these features in our approach for combination enhanced the classifier performance.

Reduced features using VQ: The VQ has been an effective data compression technique found from literature[11,18,22,37]. The method outperformed both linear and non-linear PCA in removing redundant features[5]. This has been explored to reduce the features for effective recognition and faster process. This is formulated as follows.

Consider the ZCR features of an emotional class represented as Z. The baseline ZCR frame-level features comprises of $B = \{Z_1, Z_2,..., Z_U\}$ number of source feature vectors obtained from all the utterances U of an emotional class. In this study, k-dimensional of each feature vector is taken, i.e., $Z_u = \{z_{u,1}, z_{u,2},..., z_{u,k}\}$, u = 1, 2,..., U. In VQ, these source vectors need to be mapped into another vector space consisting of code vectors comprising of finite number of regions or clusters. These code vectors form a codebook $R = \{r_1, r_2,..., r_Y\}$, where, Y is number of code vectors in the codebook. Each k-dimensional code vector is then represented by $ry = \{r_{y,1}, r_{y,2},..., r_{y,k}\}$, y = 1, 2,..., Y. The associated encoding region of the code vector r_y is represented by a q_y, y = 1, 2,..., Y with the partition space indicated as $C = \{q_1, q_2,..., q_Y\}$. In this case, if the source feature vector Z_U belongs to the encoded space q_y, then the approximation of the feature vector if $G(Z_U) = ry$ if $Z_U \in q_y$. The deviation of the features from the centroid can be found by estimating the average distortion using a squared-error distortion measure given by:

$$D = \frac{1}{Uk}\sum_{u=1}^{U}\left\|Z_u - G(Z_u)\right\|^2 \quad (5)$$

For optimal result the distortion, need to be minimized. This requires the fulfillment of both the nearest neighbor and centroid condition. The LBG algorithm has been efficient in satisfying the above two condition for obtaining optimum codebook design[17] hence, is opted in this study. The designing steps for codebook using this algorithm are as follows:

- **Compute the centroid:** For the input feature B, let the codebook size Y = 1. Set the splitting parameter to a small value. The centroid is computed using:

$$r_1^* = \frac{1}{Uk}\sum_{u=1}^{U} Z_u \quad (6)$$

This 1-vector codebook is obtained by averaging the entire training data B hence does not require any iteration. Now, the average distortion is computed using the relation:

$$D^* = \frac{1}{Uk}\sum_{u=1}^{U}\left\|Z_u - r_1^*\right\|^2 \quad (7)$$

- **Splitting:** For j = 1, 2,..., Y, set $r_j^{(0)} = r_j^*(1+\beta)$ and $r_{Y+j}^{(0)} = r_j^*(1-\beta)$. Double the codebook size i.e., Y = 2Y
- **Steps for iteration:** For j = 0, let $D^{(0)} = D^*$
  - For u = 1, 2,..., U, compute the minimum of $\| Z_u - r_y^{(j)} \|^2$, y = 1, 2,..., Y. Let this minimum value is achieved for an index y*, then set $G(Z_u) = r_{y^*}^{(j)}$
  - For y = 1, 2,..., Y, the code vector is updated as:

$$r_y^{j+1} = \frac{\sum_{G(Z_u)=r_y^{(j)}} Z_u}{\sum_{G(Z_u)=r_y^{(j)}} 1} \quad (8)$$

  While satisfying the above condition, it is essential to include at least one input feature vector in each coding region to prevent the denominator of Eq. 8 becoming zero
  - Now change j = j+1. Estimate the average distortion again using

$$D^{(j)} = \frac{1}{Uk}\sum_{u=1}^{U}\| Z_u - G(Z_u)\|^2 \quad (9)$$

  - If $\frac{\left(D^{(j-1)} - D^{(j)}\right)}{D^{(j-1)}} > \beta$, then return to step (i)
  - Further, set $D^* = D^{(j)}$. For y, set $r_y^* = r_y^{(j)}$ as the final code vectors
- Steps 2 and 3 are repeated until we obtain the desired code vectors

Codebook size of 2^3, 2^4, 2^5 and 2^6 has been tested in this experiment. After few manipulation, a codebook size of 2^4 has been opted as a tradeoff between storage space, computation time and reconstruction quality. In addition, a maximum of 20 iterations for each codebook size has been chosen to suit above tradeoff. Different splitting percentage between 0.01 and 0.05 has been tested for the codebook design. A value of 0.02 splitting percentage proved to provide the reasonable required accuracy with a 0.75 rate of reduction in split size on completion of each splitting in our case. To maintain required degree of precision a threshold of 0.002 is chosen for distance measure.

The proposed VQ based frame-level features are extracted using following steps.

- Consider the ZCR feature indicated as Z. The ZCR values of each frame are given by $Z_m = Z_1, Z_2,..., Z_M$, m = 1, 2,....., M. Where, M is number of features in a frame. The VQ is applied to each frame to extract a single VQ coefficient of that frame indicated as $VZ_n = VZ_1, VZ_2,..., VZ_N$, n = 1,2,..., N. The set of VQ based frame-level features obtained this way can be represented as . Where, number of frames in an utterance

Therefore, the total number of VQ based ZCR features of an utterance is given by:

$$N_2 = N \quad (10)$$

This way the feature is reduced from M×N to N by applying VQ. For an emotion consisting of u = 1, 2,..., U such utterances, all VQ coefficients are extracted and stored. So, for an emotional class total VQ based ZCR features extracted can be represented as $\{VZ_n(u) = VZ_n(1), VZ_n(2),..., VZ_n(U)\}$, u = 1, 2,..., U. Hence, the total number of VQ based ZCR features of an emotional class thus become N×U.

- Similarly, for STE and autocorrelation coefficient corresponding VQ based features are extracted. The VQ based features for STE, ZCR, ACF of an emotional class thus obtained can be represented as $V = \{VE_{STn}(u), VZ_n(u), VS(\tau)_n(u)\}$
- However, F0 is extracted at utterance level in this study

Different combinations have been used from these VQ based features. The combined feature sets are represented as $V = \{V_1, V_2,..., V_{11}\}$ and is given below:

$$V_1 = \{VE_{STn}(u), VZ_n(u)\}, V_2 = \{VE_{STn}(u), F0\},$$
$$V_3 = \{VE_{STn}(u), VS(\tau)_n(u)\},$$
$$V_4 = \{VE_{STn}(u), VZ_n(u), VS(\tau)_n(u)\}$$
$$V_5 = \{VE_{STn}(u), VZ_n(u), F0\}, V_6 = \{VE_{STn}(u), F0, VS(\tau)_n(u)\},$$
$$V_7 = \{VE_{STn}(u), F0, VZ_n(u), VS(\tau)\ n(u)\}, V_8 = \{F0, VZ_n(u)\}$$
$$V_9 = \{F0, VS(\tau)_n(u)\}, V_{10} = \{VS(\tau)_n(u), VZ_n(u), F0\},$$
$$V_{11} = \{VZ_n(u)+VS(\tau)_n(u)\}$$

Modified VQ features: A modified VQ feature based on statistical VQ and differential VQ based reduction method is proposed. The formulation of the modified feature vector is explained below.

Statistical VQ for reduction: For further reduction of VQ features we have adopted the statistical method since the utterances used are of different duration. Further, statistical values are capable to distinguish emotions based on arousal level. Parameters as mean, standard deviation along with the minimum, maximum values and range of the signal are computed. These parameters will provide the exact information regarding the VQ features without losing major information content. From N number of VQ features in an utterance the corresponding statistical VQ based ZCR features at utterance level are SVZ_n (mean), SVZ_n (minimum), SVZ_n (SD), and SVZ_n (range). For an emotional class having U utterances all such mean values for each feature category are stored and represented for ZCR by SVZ_n (mean) = {SVZ_1 (mean), SVZ_2 (mean),..., SVZ_U (mean)}, u = 1, 2,..., U. Similarly, other statistical values of VQ based ZCR features at utterances level for an emotional class are stored. Similar features for STE and ACF are extracted. Correspondingly, the statistical features of VQ based STE and ACF are gathered from all utterances of an emotional class.

The statistical VQ based utterance level features for STE, ZCR and ACF of each utterance of an emotional class thus obtained can be represented as $S_{VQ_u} = \{SVE_{STu}, SVZ_u, SVS(\tau)_u\}$.

The statistical features are computed using following generalized formulas for VQ based ZCR features is given by:

$$\text{Mean} = \frac{1}{N}\sum_{n=1}^{N} VZ_n \quad (11)$$

$$\text{Range} = \max(VZ_n) - \min(VZ_n) \quad (12)$$

$$\text{SD} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(VZ_n - \text{mean})^2} \quad (13)$$

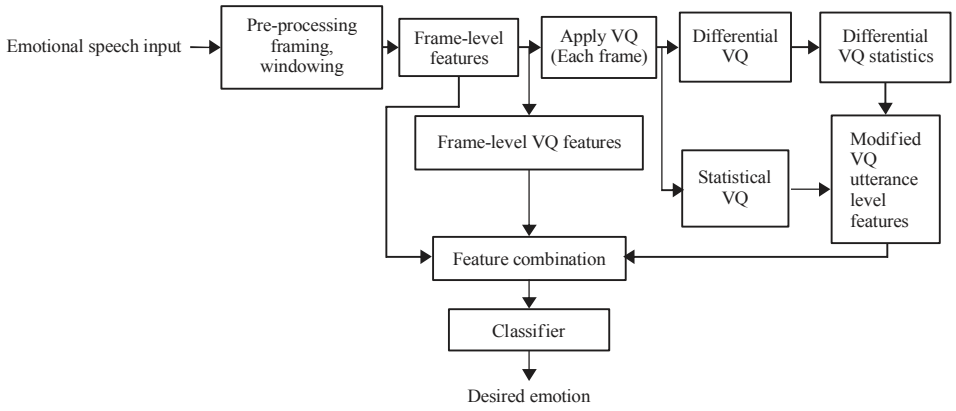

Fig. 1: Proposed feature extraction technique

This way number of VQ based features is reduced to statistical VQ features for each utterance of an emotional class. For U utterances of an emotional class, a total of statistical VQ features are obtained.

Differential VQ for reduction: However, the accuracy using statistical features suffers since the temporal information among features is completely lost during its extraction[2]. To compensate this problem, the features are reduced by taking the difference between two consecutive frame-level VQ features of an utterance. This way the temporal dynamics has been retained in the reduced features. The resultant features designated as differential_VQ and are extracted for ZCR features as:

$$d_{VQu} = Vz_n\text{-}VZ_{n-1},\ n = 1, 2,..., N \quad (14)$$

The number of d_{VQ_u} of an utterance for each emotional class is thus become N-1.

To include the statistical nature of the differential_VQ features, again their mean, maximum, minimum, standard deviation and range for each utterance is computed as before. This way, both temporal dynamics and statistical characteristics of the features has been involved in deriving differential VQ statistics, Sd_{VQ_u}. Number of statistical differential VQ based features (Sd_{VQ_u}) of each utterance is thus limited to $N_4 = 5$.

Proposed modification: To derive the modified features, the statistical VQ based features and statistical differential VQ based features for each utterance of an emotional class has been used. The modified features of each utterance can be represented as . Total number of features obtained in this way for an utterance can be given by:

$$N_4 = N_3+N_4 = 5+5 = 10 \quad (15)$$

For each emotional class comprising of U utterances, the total modified VQ based ZCR features at utterance level of an emotional class can be estimated as:

$$\text{Total modified VQ features}\left(M_{VQ}(Z)\right)_u = UXS_{VQ_u} + UXSd_{VQ_u} = 10\ U \quad (16)$$

Similar modified features are also extracted for VQ based STE and ACF features. The modified VQ based utterance level features for STE, ZCR and ACF of each utterance of an emotional class thus obtained can be represented as $(M_{VQ})_u = \{(M_{VQ}(Z))_u, (M_{VQ}(E_{ST}))_u, (M_{VQ}S(\tau)_u)\}$.

Different combinations have been used from these modified VQ features. The combined feature sets are represented as $M_{VQ} = \{M_{VQ1}, M_{VQ2},..., M_{VQ11}\}$ and is given below:

$$M_{VQ1} = \{(M_{VQ}(E_{ST}))_u, (M_{VQ}(Z))_u\},\ M_{VQ2} = \{(M_{VQ}(E_{ST}))_u, F0\},$$
$$M_{VQ3} = \{(M_{VQ}(E_{ST}))_u, (M_{VQ}S(\tau))_u\},$$
$$M_{VQ4} = \{(M_{VQ}(E_{ST}))_u, (M_{VQ}(Z))_u, (M_{VQ}S(\tau))_u\},$$
$$M_{VQ5} = \{(M_{VQ}(E_{ST}))_u, (M_{VQ}(Z))_u, F0\},$$
$$M_{VQ6} = \{(M_{VQ}(E_{ST}))_u, F0, (M_{VQ}S(\tau))_u\},$$
$$M_{VQ7} = \{(M_{VQ}(E_{ST}))_u, F0, (M_{VQ}(Z))\_u, (M_{VQ}S(\tau))_u\},$$
$$M_{VQ8} = \{F0, (M_{VQ}(Z))_u\},\ M_{VQ9} = \{F0, (M_{VQ}S(\tau))_u\},$$
$$M_{VQ10} = \{(M_{VQ}S(\tau))_u, (M_{VQ}(Z))_u, F0\},$$
$$M_{VQ11} = \{(M_{VQ}(Z))_u, (M_{VQ}S(\tau))_u\}$$

The proposed feature extraction technique has been shown in Fig. 1.

RESULTS AND DISCUSSION

Standard EMO-DB database has been tested with proposed features along with a locally collected regional Oriya speech emotional database. Due to unavailability of Oriya database, it has been collected from different sources. Emotional speech utterances of angry, fear, happy and sad

Table 1: Details of utterances used from the database

Database	References	Language	Details	Emotions	Linguistic nature
SAVEE	Haq and Jackson[8]	British English	4 male actors, 4 emotions, 240 utterances, Age: 27-31	Angry, fear, happiness and sadness	Audio
Local database	N/A	Oriya (Indian)	4 emotions, 64 utterances, different sources, Age: 25-40	Angry, fear, happiness and sadness	Audio

Table 2: Impact of dimension reduction with different feature extraction techniques

Proposed features	Baseline	VQ based	Differential VQ	Statistical VQ	Differential VQ statistics	Modified VQ
Dimension/utterance	M×N	N	N-1	5	5	10
Dimension/emotion	$\sum_{u=1}^{U} M \times N$	$\sum_{u=1}^{U} N$	$\sum_{u=1}^{U} N-1$	5×U	5×U	10×U
Accuracy	Lowest	Lower	Low	High	Higher	Highest
MSE	Highest	Higher	High	Low	Lower	Lowest

M: No. of features/frame, N: No. of frames/utterance, U: No. of utterances/emotion

states are analyzed. Sixty utterances of EMO-DB database and 16 utterances of Oriya database for each emotion are used for extraction of features and further classification. The local database is resampled to that of EMO-DB database to make the comparing platform similar. The details of the utterances used are shown in Table 1.

Neural network based classifiers as RBFN and MLP has been used to model the speech emotions in this study. These have self-learning ability and hence suitable for emotional speech recognition. Further these are parallel structures and suitable for speech signals that have frequencies occurring in parallel. These classifiers minimize error using weights and biases and are suitable for reduced features. Conventional classifiers like HMM/GMM can adequately model if feature dimensions are large but their performance degrades for reduce features. Since use of VQ at frame-level and modified VQ at utterance level reduced the feature dimension, NNs suit us. Further, the correlation among features are poorly defined by HMM as compared to NNs. Due to these factors, these classifiers can converge easily to optimal solution than other conventional classifiers.

For our purpose, optimum results have been achieved with a learning parameter of 0.01 with 0.1 moment rate using MLP. The number of epochs has been set to 50 after testing the classifier with 20, 30, 40, 50, 55 and 65 numbers of epochs. The number of iteration was chosen by taking into account the speed of response and computational complexity to achieve the desire result. Number of input layer neurons equals to available number of input source feature vectors has been maintained for both MLP and RBFN classifiers. The source features are collected from all four classes of emotions. The output neuron equals to 40% of source feature for each emotional category has been selected for classification. Hidden nodes of 10, 15, 20, 30 and 40 have been tested for both the classifiers. Highest performance with 20 and 40 hidden nodes for local database and EMO-DB database, respectively has been experienced in this study.

For RBFN a single hidden layer is sufficient to model the emotions[22] hence, selected for our purpose. Mean square error goal of 0.0 has been used with addition of extra neurons to hidden layer until we achieved the desired MSE. Spread of 0.01, 0.5, 1.0, 1.5 and 2.1 has been tested for this network. Optimum smoothing has been obtained with a spread of 2.1 in this study hence, maintained.

The classifiers are simulated with 70/15/15%, 60/20/20% and 50/25/25% training, validation and testing ratios. A ratio of 60/20/20% founds to be most effective in this experiment. The observation using proposed feature extraction methods are described below.

In deriving the modified features, we have considered the utterance level statistics of VQ based features. Statistical features are quite successfully applied in the field of emotional speech recognition. These features are able to distinguish low arousal emotions against high arousal emotion. Since this piece of work takes into consideration both arousal emotional levels, hence statistical techniques proved suitable in our case. However, in describing the desired emotions at utterance level, the temporal dynamics between features and between frames are completely lost. Hence, we have considered differential VQ features based on the variation of feature values between frames. Thus, the rate at which the VQ based features change between frames has been taken into account. This way both statistical and dynamical characteristics of the VQ based features are included. It has resulted in enhanced accuracy. Next to it, the statistics of differential VQ features are estimated to reduce the features further in modified feature set. Reduction of features in proposed methods helped in selection of features. Hence, the cross validation are executed faster. In turn, this has enhanced the compatibility of the features to the chosen NN classifiers. The impact of dimension reduction using the different feature extraction techniques has been described in Table 2.

The RBFNN classification accuracy is found to be 85.79% (Table 3), 90.38% (Table 4) and 91.08% (Table 5) with frame-level feature combination without VQ, VQ based feature combination and modified VQ based feature combination, respectively using EMO-DB database. Thus, modified VQ based feature combinations outperformed all other feature combinations. It is followed by VQ based prosodic feature combination in terms of classification accuracy. As compared to this corresponding MLP classification accuracy is 84.04% (Table 6), 88.05% (Table 7) and 89.93% (Table 8) with frame-level feature combination without VQ with VQ based feature combination and with modified VQ based feature combination respectively. These results are experienced with E+ F0 +Z+ACF feature combination, which out performed all other possible combinations. The results are better than similar works attempted using by earlier researchers[6,29,37].

Table 3: Performances of RBFN classifier with frame level prosodic feature combination using EMO-DB database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	79.72	76.97	74.20	75.65	76.64	0.782
Z+ACF	76.37	75.02	73.17	74.79	74.84	0.815
E+F0	73.12	72.50	71.69	72.42	72.43	0.859
F0+Z	74.69	73.76	71.90	73.22	73.39	0.836
E+ACF	78.58	77.87	76.61	77.60	77.67	0.775
F0+ACF	73.31	73.37	72.41	73.10	73.05	0.840
E+Z+ACF	83.21	81.24	80.43	82.11	81.75	0.613
E+F0+ACF	82.19	80.52	78.16	79.42	81.07	0.661
ACF+Z+F0	81.22	80.62	79.06	80.24	80.29	0.720
E+Z+F0	80.75	80.64	78.64	79.87	79.98	0.728
E+F0+Z+ACF	87.13	86.32	84.42	85.27	85.79	0.444

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Table 4: Performances of RBFN classifier with VQ based prosodic feature combination using EMO-DB database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	81.27	80.77	79.46	80.25	80.44	0.706
Z+ACF	81.92	80.64	79.43	80.11	80.53	0.683
E+F0	75.85	75.42	73.73	75.10	75.03	0.813
F0+Z	75.78	75.35	75.18	75.29	75.40	0.808
E+ACF	81.40	80.33	80.13	80.33	80.55	0.680
F0+ACF	76.07	75.12	74.91	75.23	75.33	0.810
E+Z+ACF	87.04	85.38	85.10	85.77	85.82	0.436
E+F0+ACF	85.26	84.17	83.01	84.19	84.16	0.487
ACF+Z+F0	83.89	83.14	82.25	83.29	83.14	0.535
E+Z+F0	84.44	83.62	81.90	82.77	83.18	0.514
E+F0+Z+ACF	90.88	90.72	89.56	90.34	90.38	0.352

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Table 5: Performances of RBFN classifier with modified VQ based feature combination using EMO-DB database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	82.53	81.85	81.70	82.21	82.07	0.609
Z+ACF	81.75	81.61	80.34	80.85	81.14	0.659
E+F0	76.55	76.33	74.01	75.32	75.55	0.810
F0+Z	82.62	80.51	79.01	80.11	80.56	0.671
E+ACF	83.61	82.72	82.12	82.39	82.71	0.562
F0+ACF	77.33	76.82	75.37	76.55	76.52	0.398
E+Z+ACF	89.88	88.60	87.35	87.92	88.44	0.379
E+F0+ACF	86.89	85.82	84.90	85.46	85.77	0.447
ACF+Z+F0	84.44	83.72	82.67	83.39	83.56	0.500
E+Z+F0	84.58	83.28	82.54	83.17	83.39	0.509
E+F0+Z+ACF	92.21	91.33	90.04	90.75	91.08	0.326

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Table 6: Performances of MLP classifier with frame level prosodic feature combination using EMO-DB database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	74.63	74.52	72.65	73.07	73.72	0.821
Z+ACF	74.05	73.57	71.98	72.87	73.12	0.843
E+F0	72.16	71.80	70.61	71.43	71.50	0.883
F0+Z	72.11	72.04	70.85	71.56	71.64	0.869
E+ACF	78.77	75.35	72.64	74.83	75.40	0.797
F0+ACF	72.55	71.54	71.03	71.30	71.61	0.872
E+Z+ACF	81.13	79.90	77.86	79.67	79.64	0.731
E+F0+ACF	80.57	79.44	77.80	78.91	79.18	0.747
ACF+Z+F0	79.38	79.40	76.92	78.00	78.43	0.764
E+Z+F0	79.86	78.78	77.63	78.59	78.72	0.758
E+F0+Z+ACF	85.66	84.51	82.05	83.93	84.04	0.493

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Table 7: Performances of MLP classifier with VQ based prosodic feature combination using EMO-DB database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	80.81	79.68	78.33	78.86	79.42	0.736
Z+ACF	79.82	79.70	77.41	78.30	78.81	0.751
E+F0	74.72	74.08	71.54	72.90	73.31	0.835
F0+Z	74.37	74.10	73.69	74.00	74.04	0.827
E+ACF	82.01	81.39	79.12	80.21	80.68	0.670
F0+ACF	74.22	74.07	73.23	73.85	73.84	0.821
E+Z+ACF	86.87	85.90	84.42	85.15	85.59	0.441
E+F0+ACF	84.05	83.78	82.44	83.14	83.35	0.542
ACF+Z+F0	82.52	81.65	80.40	81.32	81.47	0.622
E+Z+F0	82.61	81.50	79.82	80.87	81.20	0.640
E+F0+Z+ACF	89.16	87.85	87.53	87.66	88.05	0.386

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Table 8: Performances of MLP classifier with modified VQ based feature combination using EMO-DB database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	81.64	81.72	80.10	80.91	81.09	0.669
Z+ACF	82.08	80.70	78.83	79.66	80.32	0.715
E+F0	76.05	74.82	72.57	73.67	74.28	0.818
F0+Z	81.19	78.30	77.04	78.17	78.68	0.760
E+ACF	83.02	82.43	81.71	82.00	82.29	0.603
F0+ACF	76.31	75.28	75.19	75.23	75.50	0.793
E+Z+ACF	87.90	87.03	86.34	86.84	87.03	0.400
E+F0+ACF	85.39	84.51	83.57	83.96	84.36	0.485
ACF+Z+F0	83.24	82.80	81.45	82.31	82.45	0.589
E+Z+F0	84.20	83.03	81.92	82.45	82.90	0.531
E+F0+Z+ACF	91.42	89.48	88.52	90.31	89.93	0.365

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Similar trends have been observed for locally collected Oriya database (Table 9-14). Meager difference in classification accuracy between EMO-DB and locally collected database validates the authenticity of the feature extraction techniques proposed in this study.

Among both NN based classifier the accuracy level of RBFNN has been better than MLP network using both database with our proposed feature extraction techniques. The RBFN is superior to k-NN and MLP classifiers as suggested[13]. Further, RBFNNs have better training algorithm than MLP since at any time only a part of non-linear activation function is active for any given input vector[7,13,22]. Furthermore, it is easy to interpret the data using RBFNN. The MLP uses distributed learning as compared to localized learning approach adapted by RBFNN hence is slower[27].

Table 9: Performances of RBFN classifier with frame level prosodic feature combination using Oriya language database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	79.41	76.15	74.40	75.59	76.39	0.792
Z+ACF	76.04	74.96	73.11	74.21	74.58	0.818
E+F0	73.13	71.80	71.38	72.45	72.19	0.866
F0+Z	73.70	73.23	71.91	72.81	73.91	0.840
E+ACF	78.61	78.24	75.55	76.84	77.31	0.778
F0+ACF	73.85	72.60	71.72	72.31	72.62	0.845
E+Z+ACF	82.04	81.17	79.97	80.37	80.89	0.674
E+F0+ACF	81.33	80.32	79.33	79.76	80.19	0.725
ACF+Z+F0	80.46	79.70	78.25	78.52	79.23	0.737
E+Z+F0	80.73	80.11	77.97	78.05	79.22	0.751
E+F0+Z+ACF	86.46	84.82	83.20	84.15	84.66	0.480

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Table 10: Performances of RBFN classifier with VQ based prosodic feature combination using Oriya language database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	81.62	80.34	79.63	80.05	80.41	0.717
Z+ACF	81.18	80.36	79.90	80.03	80.37	0.722
E+F0	75.73	75.25	74.30	74.81	75.02	0.817
F0+Z	76.52	75.69	75.11	75.48	75.70	0.796
E+ACF	81.74	80.83	79.30	79.91	80.45	0.721
F0+ACF	76.37	75.01	74.16	74.35	74.97	0.817
E+Z+ACF	88.70	86.31	84.23	85.98	86.31	0.444
E+F0+ACF	85.48	84.65	84.50	84.66	84.82	0.479
ACF+Z+F0	85.04	82.96	81.20	82.79	83.00	0.528
E+Z+F0	84.45	83.58	82.74	83.25	83.51	0.521
E+F0+Z+ACF	91.02	90.78	88.70	89.57	90.02	0.361

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Table 11: Performances of RBFN classifier modified VQ based feature combination using Oriya language database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	82.04	81.38	80.92	81.30	81.41	0.635
Z+ACF	81.35	81.09	80.67	80.76	80.97	0.667
E+F0	76.10	75.77	73.94	74.85	75.17	0.813
F0+Z	81.85	80.40	78.88	79.64	80.19	0.722
E+ACF	83.46	82.59	82.02	82.68	82.69	0.585
F0+ACF	77.03	76.25	75.30	75.89	76.12	0.790
E+Z+ACF	89.15	88.40	87.64	88.03	88.31	0.383
E+F0+ACF	86.45	88.34	87.51	88.17	85.62	0.466
ACF+Z+F0	84.32	83.50	82.46	82.97	83.31	0.522
E+Z+F0	84.22	83.58	82.83	83.04	83.42	0.541
E+F0+Z+ACF	91.54	91.03	89.62	90.15	90.59	0.339

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Combining similar information features provided improved recognition. Features carrying complementary energy information such as ACF and STE outperformed other similar combinations. Next to it, STE and ZCR feature combination proved better. Since ZCR distinguish voiced part of the signal from unvoiced parts, thus provides complementary energy information. Among three-feature category combination, STE+ZCR+autocorrelation provided highest accuracy for similar reason. This has been experience with all the proposed feature extraction technique and with both the classifiers.

Table 12: Performances of MLP classifier frame level prosodic feature combination using Oriya language database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	74.15	71.90	74.88	73.23	73.54	0.836
Z+ACF	73.85	73.22	71.39	72.44	72.73	0.849
E+F0	71.56	70.80	69.88	70.35	70.65	0.895
F0+Z	72.03	71.12	70.55	70.92	71.16	0.888
E+ACF	78.23	75.48	72.19	74.77	75.17	0.812
F0+ACF	71.87	70.92	69.41	71.05	70.81	0.892
E+Z+ACF	80.90	79.85	78.72	79.34	79.70	0.728
E+F0+ACF	80.47	79.35	77.86	78.43	79.03	0.752
ACF+Z+F0	80.06	79.31	76.55	77.62	78.39	0.765
E+Z+F0	79.25	77.42	79.82	78.38	78.72	0.760
E+F0+Z+ACF	85.46	84.62	82.51	83.73	84.08	0.495

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Table 13: Performances of MLP classifier VQ based prosodic feature combination using Oriya language database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	79.85	79.30	77.45	78.25	78.71	0.758
Z+ACF	79.72	78.85	76.06	77.82	78.11	0.773
E+F0	74.12	72.50	71.18	72.05	72.46	0.847
F0+Z	74.75	73.89	72.54	73.23	73.60	0.822
E+ACF	81.57	80.72	78.67	79.45	80.10	0.721
F0+ACF	74.62	73.86	72.14	72.90	73.38	0.838
E+Z+ACF	86.58	85.71	84.11	84.89	85.32	0.467
E+F0+ACF	84.39	83.76	82.23	81.99	83.09	0.545
ACF+Z+F0	82.68	81.59	80.36	80.91	81.39	0.639
E+Z+F0	82.68	81.37	79.75	80.59	81.10	0.665
E+F0+Z+ACF	88.76	87.80	87.48	87.69	87.93	0.389

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

Table 14: Performances of MLP classifier modified VQ based feature combination using Oriya language database

	Emotions (%)					
Features	Angry	Happy	Sad	Fear	Average accuracy (%)	MSE
E+Z	80.75	80.51	79.85	79.93	80.26	0.720
Z+ACF	81.68	79.77	78.82	80.07	80.09	0.729
E+F0	75.15	73.86	73.05	74.22	74.07	0.819
F0+Z	79.85	78.68	77.24	77.94	78.43	0.771
E+ACF	82.65	81.56	81.32	81.89	81.86	0.716
F0+ACF	75.67	75.38	74.70	74.89	75.16	0.814
E+Z+ACF	87.55	86.37	85.92	86.10	86.49	0.409
E+F0+ACF	84.83	83.54	82.90	83.35	83.66	0.502
ACF+Z+F0	82.65	81.82	81.16	82.09	81.93	0.611
E+Z+F0	82.80	82.27	81.55	81.93	82.14	0.602
E+F0+Z+ACF	90.51	89.37	88.96	89.17	89.50	0.372

E: Short time energy, F0: Fundamental frequency, ACF: Autocorrelation coefficient, Z: Zero crossing rate

CONCLUSION

The NN based classifiers chosen in this study, better classify reduced features as our results reveal. Improvements observed when VQ based frame-level reduced features are reduced further in modified features. This observation shows that feature combination enhances classification accuracy due to increase in available information. However, to select effective features for possible combination need to be judiciously planned. This is clear from our results, when features bearing similar information are considered for effective combination. Further, the recognition is better when both reduction and feature combination mechanism are applied together. This observation clearly demonstrated that it is possible to identify emotions in speech using limited number of data if the classifier is compatible to the extracted features. A meager difference in classification accuracy between locally collected and standard EMO-DB database has supported our claim.

REFERENCES

1. Albornoz, E. and D. Milone, 2016. Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles. IEEE Trans. Affective Comput. 10.1109/TAFFC.2015.2503757
2. El Ayadi, M., M.S. Kamel and F. Karray, 2011. Survey on speech emotion recognition: Features, classification schemes and databases. Pattern Recognit., 44: 572-587.
3. Chandrasekar, P., S. Chapaneri and D. Jayaswal, 2014. Automatic speech emotion recognition: A survey. Proceedings of the International Conference on Circuits, Systems, Communication and Information Technology Applications, April 4-5, 2014, Mumbai, pp: 341-346.
4. Cheng, X. and Q. Duan, 2012. Speech emotion recognition using Gaussian mixture model. Proceedings of the 2nd International Conference on Computer Application and System Modeling, July 27-29, 2012, Shanxi, China, pp: 1222-1225.
5. Fodor, I.K., 2002. A survey of dimension reduction techniques. Technical Report, pp: 1-18. https://computation.llnl.gov/casc/sapphire/pubs/148494.pdf
6. Fulmare, N.S., P. Chakrabarti and D. Yadav, 2013. Understanding and estimation of emotional expression using acoustic analysis of natural speech. Int. J. Nat. Lang. Comput., 2: 37-46.
7. Haykins, S., 2006. Neural Networks: A Comprehensive Foundation. 2nd Edn., Pearson Education, Delhi, India.
8. Haq, S. and P.J.B. Jackson, 2010. Multimodal Emotion Recognition. In: Machine Audition: Principles, Algorithms and Systems, Wang, W. (Ed.). IGI Global Press, USA., ISBN: 978-1615209194, pp: 398-423.
9. Han, K., D. Yu and I. Tashev, 2014. Speech emotion recognition using deep neural network and extreme learning machine. Proceedings of the 15th Annual Conference of the International Speech Communication Association, September 14-18, 2014, Singapore, pp: 223-227.
10. Iliou, T. and C.N. Anagnostopoulos, 2010. Classification on speech emotion recognition-a comparative study. Int. J. Adv. Life Sci., 2: 18-28.
11. Khanna, P. and M.S. Kumar, 2011. Application of Vector Quantization in Emotion Recognition from Human Speech. In: Information Intelligence, Systems, Technology and Management, Dua, S., S. Sahni and D.P. Goyal (Eds.). Springer, Berlin, Germany, ISBN: 978-3-642-19422-1, pp: 118-125.
12. Kim, J.C. and M.A. Clements, 2015. Multimodal affect classification at various temporal lengths. IEEE Trans. Affective Comput., 6: 371-384.
13. Kolodyazhniy, V., S.D. Kreibig, J.J. Gross, W.T. Roth and F.H. Wilhelm, 2011. An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions. Psychophysiology, 48: 908-922.
14. Kuchibhotla, S., H.D. Vankayalapati, R.S. Vaddi and K.R. Anne, 2014. A comparative analysis of classifiers in emotion recognition through acoustic features. Int. J. Speech Technol., 17: 401-408.
15. Koolagudi, S.G. and K.S. Rao, 2012. Emotion recognition from speech: A review. Int. J. Speech Technol., 15: 99-117.
16. Lee, C.M. and S.S. Narayanan, 2005. Toward detecting emotions in spoken dialogs. IEEE Trans. Speech Audio Process., 13: 293-303.
17. Linde, Y., A. Buzo and R.M. Gray, 1980. An algorithm for vector quantizer design. IEEE Trans. Commun., 28: 84-95.
18. Li, Y. and Y. Zhao, 1998. Recognizing emotions in speech using short-term and long-term features. Proceedings of the 5th International Conference on Spoken Language Processing, November 30-December 4, 1998, Sydney, Australia.
19. Luengo, I., E. Navas and I. Hernaez, 2010. Feature analysis and evaluation for automatic emotion identification in speech. IEEE Trans. Multimedia, 12: 490-501.
20. Nwe, T.L., S.W. Foo and L.C. De Silva, 2003. Speech emotion recognition using hidden Markov models. Speech Commun., 41: 603-623.
21. Pao, T.L., Y.T. Chen, J.H. Yeh and W.Y. Liao, 2005. Detecting emotions in Mandarin speech. Comput. Linguist. Chin. Lang. Process., 10: 347-362.
22. Palo, H.K., M.N. Mohanty and M. Chandra, 2016. Efficient feature combination techniques for emotional speech classification. Int. J. Speech Technol., 19: 135-150.
23. Palo, H.K., M.N. Mohanty and M. Chandra, 2015. Design of Neural Network Model for Emotional Speech Recognition. In: Artificial Intelligence and Evolutionary Algorithms in Engineering Systems, Suresh, L.P., S.S. Dash and B.K. Panigrahi (Eds.). Springer, India, ISBN: 978-81-322-2134-0, pp: 291-300.
24. Palo, H.K., M.N. Mohanty and M. Chandra, 2015. Use of different features for emotion recognition using MLP network. Adv. Intell. Syst. Comput., 332: 7-15.
25. Quan, C., D. Wan, B. Zhang and F. Ren, 2013. Reduce the dimensions of emotional features by principal component analysis for speech emotion recognition. Proceedings of the IEEE/SICE International Symposium on System Integration, December 15-17, 2013, Kobe, Japan, pp: 222-226.
26. Ramakrishnan, S., 2012. Recognition of Emotion from Speech: A Review. In: Speech Enhancement, Modeling and Recognition-Algorithms and Applications, Ramakrishnan, S. (Ed.). InTech Inc., Croatia, ISBN: 978-953-51-0291-5, pp: 121-138.
27. Santos, R.B., M. Rupp, S.J. Bonzi and A.M.F. Fileti, 2013. Comparison between multilayer feedforward neural networks and a radial basis function network to detect and locate leaks in pipelines transporting gas. Chem. Eng. Trans., 32: 1375-1380.

28. Schuller, B., S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Muller and S. Narayanan, 2013. Paralinguistics in speech and language-state-of-the-art and the challenge. Comput. Speech Lang., 27: 4-39.
29. Schuller, B., A. Batliner, D. Seppi, S. Steidl and T. Vogt *et al*., 2007. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. Proceedings of the 8th Annual Conference of the International Speech Communication Association, August 27-31, 2007, Antwerp, Belgium, pp: 2253-2256.
30. Schuller, B., A. Batliner, S. Steidl and D. Seppi, 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun., 53: 1062-1087.
31. 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. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 1, May 17-21, 2004, Montreal, Canada, pp: I-577-I-580.
32. Tahon, M. and L. Devillers, 2016. Towards a small set of robust acoustic features for emotion recognition: Challenges. IEEE/ACM Trans. Audio Speech Lang. Process., 24: 16-28.
33. Ververidis, D. and C. Kotropoulos, 2006. Emotional speech recognition: Resources, features and methods. Speech Commun., 48: 1162-1181.
34. Wang, K., N. An, B.N. Li, Y. Zhang and L. Li, 2015. Speech emotion recognition using Fourier parameters. IEEE Trans. Affective Comput., 6: 69-75.
35. Wang, K.C., 2015. Speech emotional classification using texture image information features. Int. J. Signal Process. Syst., 3: 1-7.
36. Wang, J.C., Y.H. Chin, B.W. Chen, C.H. Lin and C.H. Wu, 2015. Speech emotion verification using emotion variance modeling and discriminant scale-frequency maps. IEEE/ACM Trans. Audio Speech Lang. Process., 23: 1552-1562.
37. Wu, C.H. and W.B. Liang, 2011. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans. Affective Comput., 2: 10-21.
38. Wu, S., T.H. Falk and W.Y. Chan, 2011. Automatic speech emotion recognition using modulation spectral features. Speech Commun., 53: 768-785.
39. Wenjing, H., L. Haifeng and G. Chunyu, 2009. A hybrid speech emotion perception method of VQ-based feature processing and ANN recognition. Proceedings of the WRI Global Congress on Intelligent Systems, Volume 2, May 19-21, 2009, Xiamen, China, pp: 145-149.