International Journal of Engineering & Technology, 7 (4.39) (2018) 112-116

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Comparative Analysis of Neural Networks for Speech Emotion Recognition

Hemanta Kumar Palo[1], Mihir N. Mohanty[2*]

[1,2]Departmet of Electronics and Communication Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be University, Bhubaneswar
**Corresponding author E-mail: mihirmohanty@soa.ac.in*

Abstract

This paper aims to investigate the ability of Neural Network (NN) models in recognizing speech emotions. Extensive simulation of NN models such as the Radial Basis Function Network (RBFN), the Multilayer Perceptron (MLP), and the Probabilistic Neural Network (PNN) has been carried out to determine the Speech Emotion Recognition (SER) Accuracy of emotional states such as anger, happiness, sadness, and boredom. The utterances for these states are chosen from the standard Berlin (EMO-DB) database. The efficient Cepstral domain vocal tract system features such as the Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCCs), the Perceptual Linear Prediction coefficients (PLP) are put to test for their emotional discriminating ability with the proposed setup. These features are extracted at a frame-level and are clustered into their corresponding Vector Quantized (VQ) coefficients to get rid of the redundant information before simulating the chosen classifiers. The NN based identification system models are experienced with the desired level of SER accuracy as these classifiers remain effective for low-dimensional feature sets. An improved accuracy of 83% has been observed with the PNN using the LPCCVQ feature sets as compared to 82% with the RBFN and 78% with the MLP. Amongst the derived feature sets, the LPCCVQ remains more reliable in characterizing the intended speech emotions while the PNN has outperformed other NN classifiers in the classification category as revealed from our results.

Keywords:*Neural Network; Speech Emotion Recognition; Feature Extraction; Feature Reduction; Classification*

1. Introduction

Amongst many independent modalities of expressive emotions, such as posture, gesture, speech and facial expressions, this piece of work will address on spoken contents of the voice signal. Speech happens to play a pivotal role, as it is the lone effective channel of communication via the phone. These emotions contain different spectral, prosodic and acoustic properties. The extraction of relevant and discriminate features describing speech emotions for effective characterization and classification remains a challenge to the community. A judicious selection of the most relevant features by removing redundant data has been an area of focus researchers are struggling with during the last few decades. Attributes to decrease in high dimensional data calculation in feature extraction and selection methods employed remained vital for classifier performance as of today.

It has been observed that the frequency analysis of a signal can provide more emotional relevant information than the time domain analysis. This is due to the fact that, the size, and shape of the vocal tract and vocal folds which vibrate and resonates differently based on the arousal states of an emotion. The spectral/Cepstral contents have evidenced with a better emotional description as compared to the prosodic representation as reported in many pieces of literature [1, 2]. Among many spectral techniques, the LP based features such as the LPC, LPCC, MFCC, and PLP have been widely acknowledged as standard and effective features in the field of speech, and emotional analysis. However, these frame-level features contain irrelevant data due to the presence of the amplitude and energy information in a signal. Such data redundancy can overload the classifiers with large storage space and requires more time for modeling the desired emotions. It creates space for a suitable feature selection algorithm for faster and efficient recognition of the intended emotions. Over the years, many feature selection algorithms such as the Principal Component Analysis, Fisher Discriminant Ratio (FDR), Vector Quantization (VQ), multi-class Linear Discriminant Analysis (LDA), Sequential Forward Selection (SFS), statistical techniques, i-vector, Harmony search, etc. have been explored by many researchers [1,3-5]. The PCA components are unable to discriminate among classes, in which case the LDA is a preferable choice as the latter can provide both within the class and between class information by projecting the training samples onto the fisher basis vectors. However, in these techniques, the representing variables are the projected ones, hence cannot be guaranteed in a true sense. On the contrary, the non-linearity dependency, vector dimensionality, correlation, and probability density function shape in clustering the features make the VQ a more versatile technique. It has outperformed the conventional PCA in clustering emotional data with better SER accuracy [6-9]. The technique is likely to approximate the human hearing mechanism when applied in conjunction with spectral based features for better emotional modeling. This has motivated the authors to explore the application domain of VQ to modify the vocal tract system features (LPCC, MFCC, and PLP) towards enhanced SER accuracy. The modified features have been tested and compared for classification accuracy with a few of the efficient NN models.

The choice of NN classifiers is made based on evidences and their advantages over the Gaussian Mixture Models or Hidden Markov Models in a reduced feature space [10]. The classifiers have better self-learning ability and regulating property in describing complex

input/output relationship between emotional states [11]. Their parallel structure suits speech analysis application as these signals have frequencies occurring in parallel. The use of the VQ clustering algorithm helps to maintain the desired level of compatibility with the NNs. The derived low-dimensional feature sets are likely to take less time and storage space during training of the classifiers. In the process, the effectiveness of three NNs such as the MLP, RBFN, and the PNN is considered for the proposed analysis. As compared to the MLP, the input to the hidden space is non-linearly mapped in RBFN whereas it is linearly mapped in MLP. The use of a table look-up interpolation scheme in a reduced feature space makes the RBFN more versatile than the MLP. It is faster to train the RBFN as the network does not back-propagate the error during classification modeling, unlike MLP [12]. Among the discussed classifiers, the PNN has better statistical properties, higher speed of learning, and lower computation time than other NNs such as the MLP or the RBFN [13]. The effectiveness of the PNN as a pattern recognizing tool in the field of SER is attributed to the following. First: the ability of the network to approach the Bays' optimal solution makes it faster with better convergence. Second: the absence of any local minima issues assists the network to generate the required predicted class accurately. Third: it is not necessary to retrain the network when the size of the training data varies. Fourth: the ominous presence of parallel structure in this network makes them reluctant to outliers. Finally, the network remains accurate and faster than other conventional machine learners in pattern recognition application. Quick convergence, simple training, and ease of implementation have been three motivation factors for choosing this classifier in this work.

The organization of the paper is as follows: Section 2 elaborates the feature modification algorithm proposed to simulate the chosen classifiers. This section also briefly explains the chosen database and classification schemes employed in this work. Section 3 analyzes the findings of this work. Section 4 concludes the work with necessary future directions.

2. Materials and Methods:

The derived feature extraction algorithm, choice of an emotional dataset and the classification models used to compare the SER accuracy have been explained in this section.

2.1. The Key Feature Extraction Algorithm

To form the desired feature vectors, the signal '$x(n)$' under consideration is pre-processed, normalized and mean subtracted to minimize the speaker and environmental variability. The frame level features are extracted from each sampled signal using a frame size of 30ms with 50% overlapping between frames. The signal passes through a Hamming window before further processing.

The LPCC uses a log scale that takes into consideration the human hearing mechanism. The coefficients are thus more discriminating and can provide better emotional relevant information than the conventional LPCs. The LPCC coefficients can be derived from the LP coefficients features using the relation

$$\begin{aligned} L_0 &= \ln r \\ L_p &= a_p + \sum_{j=1}^{p-1}\left(\frac{j}{p}\right) L_i\, a_{p-i}\,, \text{for } 1 < p < r,\ \ j = 1,2,\dots,r \\ L_p &= \sum_{j=p-r}^{p-1}\left(\frac{j}{p}\right) L_j\, a_{p-i}\,, \qquad \text{for } \ p > r \end{aligned} \tag{1}$$

where p denotes the number of samples in a windowed frame and r representsthe number of previous samples to be combined linearly. The terms a_j denote the LP coefficients. Out of each frame, sixteen number of LPCC coefficients are retained for further simulation as it can provide most of the perceptual emotional relevant information. The feature matrix of an emotional class so formed can be represented as

$$L(e) = \begin{Bmatrix} L_{1,1}(1) & \dots & L_{1,16}(1) \\ \vdots & \ddots & \vdots \\ L_{F,1}(1) & \dots & L_{F,16}(1) \\ L_{1,1}(2) & \dots & L_{1,16}(2) \\ \vdots & \ddots & \vdots \\ L_{F,1}(2) & \dots & L_{F,16}(2) \\ \vdots & \ddots & \vdots \\ L_{1,1}(U) & \dots & L_{1,16}(U) \\ \vdots & \ddots & \vdots \\ L_{F,1}(U) & \dots & L_{F,16}(U) \end{Bmatrix}, e = 1,2,\dots\ ,E \tag{2}$$

Where U represents the number of utterances in an emotional state, F is the number of frames per utterance, and E denotes the number of emotional states. The LPCC feature set extracted this way has a size of $UF \times 16$. The VQ clustering algorithm is then applied to the first coefficient of each frame individually to obtain a single LPCC-VQ coefficient. Thus, a signal comprising of $F -$ number of first LPCC coefficient can be represented by a single LPCCVQ coefficient and so on. This way an LPCC-VQ feature vector of size 16×1 per utterance of an emotion is formed from sixteen columns of the LPCC feature matrix using equation (1). The size of the LPCC-VQ vector of an emotional state with a $U -$ number of utterances becomes $16U \times 1$ for simulation of the classifier. The LPCC-VQ feature vector of an emotional utterance can be represented as

$$L_V(u) = \{L_V(1), L_V(2), \dots, L_V(16)\}, u = 1,2,\dots\ U \tag{3}$$

For the extraction of the MFCC coefficients, the windowed spectrums of the signal $x_F(t)$ under consideration with the fundamental frequency f is wrapped into a Mel frequency scale using the relation

$$f_M = 2595 \log_{10}\left(\frac{f}{700} + 1\right) \tag{4}$$

The resultant Mel-filter coefficients are fed to a logarithm block to approximate the logarithmic human hearing mechanism. The log-Mel coefficients are converted into the Discrete Cosine Transform (DCT) coefficients to obtain the MFCCs finally. It is observed that the higher order DCT parameters change very fast and tend to degrade the performance of the recognizer. Hence, these coefficients are dropped and only the first sixteen coefficients are retained similar to LPCC technique. The VQ clustering algorithms is then applied to the frame-level MFCCs similar to eqution (3).

To compute the PLP coefficients, a Bark frequency scale has been used to wrap the windowed signal $x_F(t)$instead of a Mel-scale in MFCC using the relation

$$f_B(w) = 6\, ln\left[\frac{w}{1200\pi} + \left[\left(\frac{w}{1200\pi}\right)^2 + 1\right]^{1/2}\right] \tag{5}$$

Where w denotes the angular frequency of the windowed signal under consideration. The PLP technique is capable of providing the critical band spectral resolution and gives due importance to an equal loudness level. The use of the intensity-loudness power law algorithm makes it more versatile than the LPC technique in describing the emotions in a speech signal. Further, the coefficients deemphasize higher order frequencies, unlike LPCs that considers only higher frequencies. This makes the PLP features more correlated for better emotional portrayal. The modified feature extraction method is shown in Fig. 1.

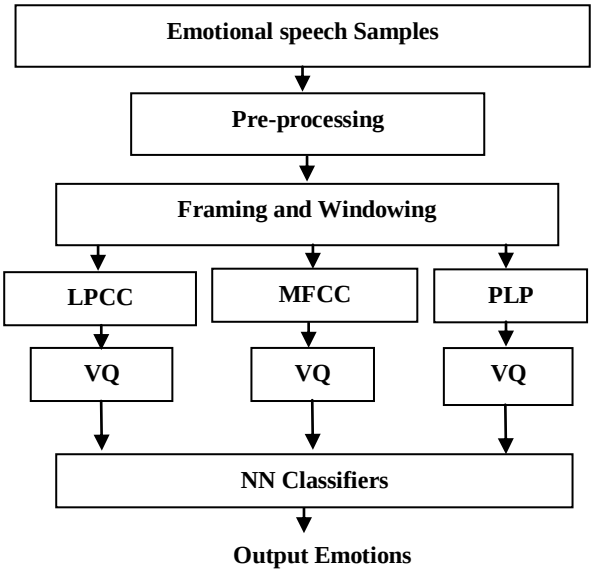

Fig. 1: The Modified Feature Extraction Method

2.2. The Classification Method:

Three efficient NN classifiers such as the Multilayer Perceptron (MLP), Radial Basis Function Network (RBFN) and the Probabilistic Neural Network (PNN) have been chosen for the proposed work. The ability of these classifiers has been compared with the derived low-dimensional feature sets for better SER accuracy. A brief on the machine learning algorithms have been given below.
The MLP updates the weights and biases using the error back-propagation algorithm during training and classification purpose. The input feature vectors pass through the hidden layer to the output and back-tracked to minimize the error between the input and the target vectors. A negative gradient cost function (δ) is used to update the synaptic weights ΔW as per the relation

$$\Delta W = -\beta \frac{\partial \delta}{\partial w} \tag{6}$$

whereβ signifies the back propagation learning parameter.
Similar to the MLP, the RBFN has input, hidden, and the output layers. However, in this case, the function$\gamma_h(\cdot)$ is formulated with the Euclidean distance norm using the relation

$$\gamma_h(\cdot) = \sum_{h=1}^{H} \gamma_h \left\| r_{m,h} - c_h \right\| \tag{7}$$

Where $h = 1,2,\dots,H$ are the number of hidden layers RBF units, $\|\bullet\|$ is the Euclidean distance measure, $x_{m,h}$ denotes the input feature vector, c_h is the RBF centre, γ_h is the RBFN activation function, and $1 \le m \le M$ represents the training pattern index. A Gaussian activation function has been used in this work as given by

$$\gamma_h(r) = exp\left(\frac{\left\|r_{m,h} - c_h\right\|^2}{\sigma_h}\right) \tag{8}$$

whereσ_h represents the spread factor and r is the input pattern.
The PNN consists of the input, pattern, summation and the output layers [14]. For a given input pattern$r_{m.n}$, a smoothing parameter σ is used to obtain the pattern layer output as given by.

$$O_{m,n}(z) = \frac{1}{(2\pi)^{\frac{s}{2}}\sigma^s} exp\left[-\frac{(r_m - r_{m,n})^T (r - r_{m,n})}{2\sigma^2}\right] \tag{9}$$

Where r is the input pattern with pattern vector dimension$s$. A maximum likelihood of the chosen pattern 'r' is then computed by the third layer (summation layer) neurons. The layer summarizes and computes the average of all the neuron outputs corresponding to a particular emotional class and classifies the pattern'r' as belonging to that class.

$$p_o(r) = \frac{1}{(2\pi)^{\frac{s}{2}}\mu^s} \frac{1}{F_m} \sum_{n=1}^{F_m} exp\left[-\frac{(r_m - r_{m,n})^T (r - r_{m,n})}{2\sigma^2}\right] \tag{10}$$

whereF_m is the number of features in the designated emotional class.

2.3. The Chosen Database:

For extraction of the derived feature sets and classification purpose, the widely popular Berlin (EMO-DB) has been chosen [15]. The database comprises of fear, sadness, anger, boredom, disgust, happiness, and neutral emotional states simulated by ten (five male and five female) professional actors. A sampling frequency of 48 kHz down-sampled to 16 kHz has been used to record the emotional utterances and evaluated for their emotional relevance by subjective judgments. From each emotional category, forty-five numbers of utterances are chosen for the proposed analysis. These emotions are anger, boredom, sadness and happiness. The choice of the dataset is made on the basis that a number of authenticated works in this field have been validated using this dataset which can create a sound comparing platform.

3. Results and Discussion:

The classifiers are simulated with feature vectors of all emotions fed as input and the feature vector of each emotion fed as output for testing the recognition accuracy. The individual accuracy is averaged out to obtain the overall accuracy of the emotional states. For classification purpose, 70% of the input feature vectors is used for training. For validation and testing the classifier 15% each of the input feature vectors has been used. A comparison of recognition error among different emotional states with the derived feature sets has been made for the MLP classifier in Table 1. The classifier has shown a lower error for the happy and angry states as compared to the bore and sad states. Between the derived feature sets, the LPCCVQ has shown to outperform all other feature extraction techniques as observed from this Table. The reason being, the LPCC technique considers the acoustic properties of a speech signal uniformly at all frequencies as compared to the MFCCs or the PLPs that use either Mel or Bark scales in describing the signal contents. Arguably, the accuracy level using LPCCs has been better as there is every possibility of emotional information to be present uniformly across the entire frequency band.

Table 1: Comparison of Recognition Error with the Derived Feature Sets using MLP Classifier

Feature	Bore	Angry	Happy	Sad	Average
LPCCVQ	0.25	0.22	0.21	0.23	0.22
MFCCVQ	0.29	0.24	0.23	0.28	0.26
PLPVQ	0.33	0.28	0.25	0.30	0.29

Similar results have been obtained with the RBFN and the PNN classifiers with the LPCCVQ feature sets. The classification error with respect to different emotional states using the derived feature sets has been shown in Table 2 for RBFN.

Table 2: Comparison of the Recognition Error with the Derived Feature Sets using RBFN Classifier

Feature	Bore	Angry	Happy	Sad	Average
LPCCVQ	0.17	0.21	0.14	0.22	0.18
MFCCVQ	0.21	0.24	0.22	0.25	0.23
PLPVQ	0.23	0.28	0.25	0.32	0.27

Table 3 provides the classification error with respect to different emotional states using the derived feature sets for the PNN classifier.

Table 3: Comparison of the Recognition Error with the Derived Feature using PNN Classifier

Feature	Bore	Angry	Happy	Sad	Average
LPCCVQ	0.20	0.15	0.21	0.13	0.17
MFCCVQ	0.23	0.17	0.25	0.19	0.21
PLPVQ	0.25	0.21	0.27	0.23	0.23

Fig. 2 provides a graphical representation of training versus the testing accuracy of the PNN classifier using the LPCCVQ feature sets with the variation in smoothing parameter for the happy emotional state. A highest average testing accuracy of 69.7% has been observed by this state as observed from this Figure.

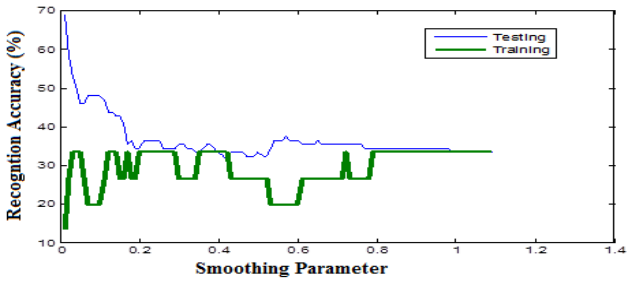

Fig. 2: PNN Recognition Accuracy versus the Smoothing Parameter for Sad Emotional State with LPCCVQ Feature sets

A comparison of classification accuracy has been made among the chosen classifiers with the derived spectral feature sets and is shown in Fig. 3. The PNN has shown to outperform the RBFN and the MLP in providing better emotional models as observed from this Table. The ability of the PNN classifier to approach the Bayes' optimal solution and absence of multi-parameter adjustment makes the system faster and less complex with better SER accuracy without any data over-fitting. Among the RBFN and the MLP, the former has shown better recognition accuracy as our results reveal. The possibility to use the table look-up interpolation scheme has favored the RBFN classifier with better accuracy as observed in this Figure.

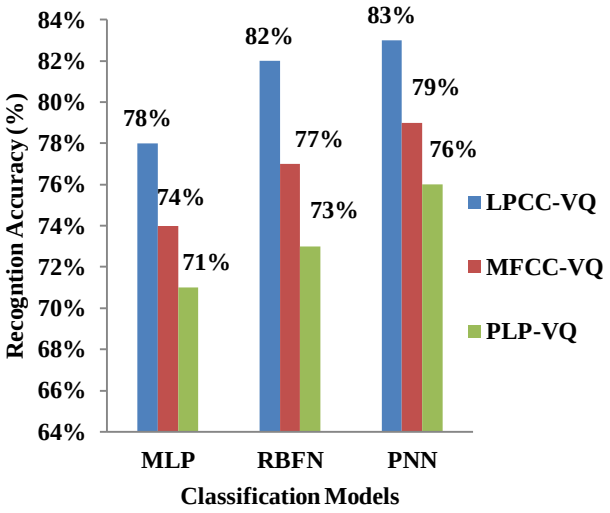

Fig. 3: Comparison of Recognition Accuracy with Different Classifiers using the Derived Feature Sets

A comparative table is provided in Table 4 for the proposed method with that of state-of-art feature extraction and classification algorithm applied in the field of SER.

Table 4: Comparison of the State-of-art Methods with the Proposed Method

Literature	Baseline features	Feature reduction (dimension)	Classifier	Highest accuracy	Database used	No of emotions
Wenjinget. al., 2009 [9]	Prosodic, 12 MFCC, 12 ΔMFCC	VQ Statistical	ANN ANN	71.7% 62.4%	Self-database	
Rao et. al., 2013[16]	LPCC MFCC	-	GMM	64% 63%	EMO-DB	
Kamińska et. al., 2013 [17]	MFCC PLP	Statistical	KNN	68.9% 61.8%	EMO-DB	7
Wu et. al., 2011 [5]	MFCC PLP	SFFS, LDA	SVM	76.5% 71.2%	EMO-DB	
Yuan et. al., 2015 [18]	Prosodic, MFCC	WLDA (3), PCA (161), LDA (3)	SVM	WLDA: 88.78% PCA: 80.40% LDA: 79.63%	UCI and CASIA	4
Quanet. al., 2017 [2]	Correlation, cepstral distance, MFCC, prosodic	PCA+ grid search	SVM	<80% (highest for 3 emotion)	EMODB	3, 4, 6
Proposed method	LPCCVQC MFCCVQC PLPVQC	VQ statistics	MLP RBFN PNN, DNN	Highest with PNN (83%) 79% 76%	EMO-DB	4

4. Conclusion:

The work aims to investigate the ability of three efficient NN classifiers in modeling speech emotions such as anger, sadness, boredom and happiness. It modifies a few Cepstral domain features using VQ coefficients to simulate the classifiers for better SER accuracy. The LPCC features represented by their VQ coefficients have provided most discriminating features as experienced from our simulation. In the classification category, the PNN has outperformed other chosen NNs in recognizing the emotions considered in this work. Application of other feature extraction and selection techniques, combination mechanism and compatible classifiers may lead to better SER accuracy and provide future perspective in this direction.

References

[1] D Torres-Boza, M C Oveneke, F Wang, D Jiang, W Verhelst, and H Sahli, "Hierarchical sparse coding framework for speech emotion recognition," Speech Communication. Vol. 99, 80-89, 1st May 2018.

[2] C Quan, B Zhang, X Sun, and F Ren, "A combined cepstral distance method for emotional speech recognition," International Journal of Advanced Robotic Systems, Vol. 14, No. 4, Jul 2017.
[3] A Majkowski, M Kolodziej, R J Rak, and R Korczyński, "Classification of emotions from speech signal," In Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), IEEE, 276-281, 21st Sep 2016.
[4] H K Palo, M N Mohanty, and M Chandra, "Efficient feature combination techniques for emotional speech classification," International journal of speech technology, Vol.19, No.1, 135-150, Mar 2016.
[5] S Wu, T H Falk, and W Y Chan, "Automatic speech emotion recognition using modulation spectral features," Speech communication, Vol. 53, No.5, 768-785,1st May 2011.
[6] SG Koolagudi, Y S Murthy, and S P Bhaskar, "Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition," International Journal of Speech Technology, Vol 21, No.1, 167-183, Mar 2018.
[7] P Khanna, and M S Kumar, "Application of vector quantization in emotion recognition from human speech," In International conference on information intelligence, systems, technology and management, Springer, Berlin, Heidelberg, 118-125, 10th Mar 2011.
[8] H K Palo, and M N Mohanty, "Wavelet based feature combination for recognition of emotions," Ain Shams Engineering Journal, 28th Jan 2017 (in press).
[9] H Wenjing, L Haifeng, and G Chunyu, "A hybrid speech emotion perception method of VQ-based feature processing and ANN recognition," In Intelligent Systems, GCIS'09, WRI Global Congress on 2009, IEEE, Vol. 2, 145-149, 19th May, 2009.
[10] M. E. Ayadi, M SKamel, and FKarray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, Vol.44, No.3,572-587, 1st Mar 2011.
[11] S SHaykin, "Neural networks and learning machines," Upper Saddle River: Pearson, Vol.3, Nov 2009.
[12] H K Palo, and M N Mohanty, "Modified-VQ Features for Speech Emotion Recognition," Journal of Applied Sciences, Vol.16, No.9, 406-418, 15th Aug 2016.
[13] D F Specht, "Probabilistic neural networks," Neural networks, Vol.3, No.1, 109-118, 1st Jan 1990.
[14] H K Palo, M N Mohanty, and M Chandra, "New features for emotional speech recognition," In IEEE Power, Communication and Information Technology Conference (PCITC), 424-429, 15th Oct 2015.
[15] F Burkhardt, A Paeschke, M Rolfes, W F Sendlmeier, and B Weiss, "A database of German emotional speech," In Ninth European Conference on Speech Communication and Technology, Vol. 5, 1517–1520, 4th Sep 2005.
[16] KSRao, and SGKoolagudi, "Robust emotion recognition using pitch synchronous and sub-syllabic spectral features. SpringerBriefs in Speech Technology, 17-46, Springer, New York, NY, 2013.
[17] D Kamińska, T Sapiński, and A Pelikant, "Comparison of perceptual features efficiency for automatic identification of emotional states from speech," In Human System Interaction (HSI), 6th IEEE International Conference, 210-213, 6th Jun 2013.
[18] JYuan, LChen, TFan, and JJia, "Dimension reduction of speech emotion feature based on weighted linear discriminate analysis,"International Journal on Image Processing and Pattern Recognition, Vol.8, No.11, 299-308, 8th Nov 2015.