Computers and Electrical Engineering 55 (2016) 12–23

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Classification of speech under stress using harmonic peak to energy ratio

Suman Deb*, S. Dandapat

Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India

ARTICLE INFO

Article history:
Received 25 February 2016
Revised 21 September 2016
Accepted 22 September 2016

Keywords:
Harmonic peak
Signal energy
Speech under stress
Binary-cascade

ABSTRACT

This paper explores the analysis and classification of speech under stress using a new feature, harmonic peak to energy ratio (HPER). The HPER feature is computed from the Fourier spectra of speech signal. The harmonic amplitudes are closely related to breathiness levels of speech. These breathiness levels may be different for different stress conditions. The statistical analysis shows that the proposed HPER feature is useful in characterization of various stress classes. Support Vector Machine (SVM) classifier with binary cascade strategy is used to evaluate the performance of the HPER feature using simulated stressed speech database (SSD). The performance results show that the HPER feature successfully characterizes different stress conditions. The performance of the HPER feature is compared with the mel frequency cepstral coefficients (MFCC), the Linear prediction coefficients (LPC) and the Teager-Energy-Operator (TEO) based Critical Band TEO Autocorrelation Envelope (TEO-CB-Auto-Env) features. The proposed HPER feature outperforms the MFCC, LPC and TEO-CB-Auto-Env features. The combination of the HPER feature with the MFCC feature further increases the system performance.

1. Introduction

Speech under stress is defined as the speech produced under any stress conditions, which perturbs speech production from the neutral condition. There are several reasons that cause stress. Some of the reasons are workload, glottal abnormalities, task demand, noisy environment (Lombard effect), specific emotions such as sad, angry and anxiety [1]. This may results in alteration of speech production mechanism from the neutral condition. Due to this, the performance of speech recognition or speaker recognition decreases under stress conditions. Analysis of speech under stress may improve the performance of speech recognition or speaker recognition. Therefore, analysis of speech under stress is very useful for man-machine interaction.

Analysis of speech under stress can be divided into two parts, feature extraction part and modeling/classification part. In feature extraction part, the desired information is extracted. The modeling/classification part consists of two stages, the training stage and the testing stage. During training, the parameters of the model are updated. In the testing stage, a score is calculated and based on that we make a classification decision. Various modeling techniques have been used for classification of speech under stress. Hidden Markov Model (HMM) [2,3], Artificial Neural Network (ANN) [3] and Support Vector Machine (SVM) [4,5] are used extensively. The performance of speech under stress classification depends on the model

* Corresponding author.
E-mail addresses: suman.2013@iitg.ernet.in (S. Deb), samaren@iitg.ernet.in (S. Dandapat).

http://dx.doi.org/10.1016/j.compeleceng.2016.09.027
0045-7906/

chosen as well as the type of the feature used. Researchers have done many experiments in this regard. First, continuous features including energy, timing and pitch related features provide important cues about various stress conditions [1]. The mel frequency cepstral coefficients (MFCC) feature capturing vocal tract information has been used for speech under stress classification [6,7]. The Linear prediction coefficients (LPC), derived from the linear source filtering concept, is tested for classification of speech under stress [7]. Zhou et al. have shown that Teager-Energy-Operator (TEO) based Critical Band TEO Autocorrelation Envelope (TEO-CB-Auto-Env) feature, derived from non-linear vortex flow through the vocal tract, successfully classify the speech under different stress conditions [2]. Zao et al. have used the pH time frequency feature for speech under stress classification [8]. It is found that pH feature successfully characterizes different stress conditions. Yao et al. have shown that the features, derived from the physical model, are effective in stress classification [9]. Searching for new feature is always a pivot part in classification of speech under stress.

In this work, we have proposed a new feature, harmonic peak to energy ratio (HPER), for classification of speech under stress. The harmonic amplitude is the index, which measures the breathiness level [10]. The breathiness of speech has been used extensively for the detection of different pathologies from the speech signal [10,11]. It is expected that different stress classes may have different breathiness levels. This gives us motivation to propose the HPER feature for analysis and classification of speech under stress. The major contribution of this paper is proposing a new feature, HPER, for speech under stress classification. Along with this, the other contributions are *i*) binary-cascade multi-classification approach using SVM classifier for SSD database and *ii*) a combination of the HPER and the MFCC features for further analysis of classification performance.

The organization of the paper is as follows. The analysis of the HPER feature for speech under stress classification is explained in Section 2. Section 3 discusses the performance of the HPER feature using SVM classifier, and the conclusion is made in Section 4.

2. Harmonic peak to energy ratio (HPER) for classification of speech under stress

This section discusses the process to compute the proposed feature, harmonic peak to energy ratio (HPER) (Section 2.1), the significance of the HPER feature using statistical analysis (Section 2.2), the details about the database (Section 2.3), the SVM classifier (Section 2.4) and the binary-cascade multi-class classification approach of stress classification (Section 2.5). Harmonic peak to energy ratio (HPER) is the amplitudes of the harmonics relative to the total energy of the speech signal. The energy distributions of different stress classes vary with frequency bands. The high-activation stress classes, like happiness and anger, are more concentrated around high frequency regions. On the other hand, low-arousal stress classes, such as sadness and boredom, are low-pitched signals [8]. Lower order harmonics correspond to the lower frequency components, where as higher order harmonics correspond to the higher frequency components. Therefore, amplitude of different harmonic capture the energy concentration at different frequency components. The HPER measures how the harmonic intensity (energy) varies with respect to the total energy. The harmonic amplitude of speech spectrum has also been used for analysis of breathy voice quality. The breathiness level may also be different for different stress conditions.

2.1. Harmonic peak to energy ratio (HPER)

Harmonic peak to energy ratio (HPER) is defined as a ratio of harmonic peaks to the total energy of the speech signal. The HPER feature vector consists of $HPER_i$ elements, $\mathbf{HPER} = [\mathbf{HPER_1}, \mathbf{HPER_2}, \ldots, \mathbf{HPER_M}]^{\mathbf{T}}$, where $i = 1, 2, \ldots, M$ and M is the number of harmonics. The steps of the proposed feature extraction method are described as follows

i) The speech signal is decomposed into a number of frames of 20 ms length with 10 ms frame shift.
ii) Each frame is multiplied with a hamming window to reduce the signal discontinuities at both the ends.
iii) The pitch frequency for each frame is calculated using autocorrelation method. The complete step of pitch estimation is as follows: For a given signal $x(n)$, the autocorrelation $R_x(m)$ is defined as [12]

$$R_x(m) = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{n=-N}^{N} x(n)x(n+m). \tag{1}$$

Thus, if the signal $x(n)$ is periodic with period T, the autocorrelation is also periodic i.e. $R_x(m) = R_x(m+T)$. For a non-stationary signal like speech, the autocorrelation is defined on short segments of speech signal, and it is given by [12]

$$R_l(m) = \frac{1}{N} \sum_{n=0}^{N'-1} [x(n+l)w(n)][x(n+l+m)w(n+m)] \tag{2}$$

where $0 \le m \le M_0 - 1$, N represents the section length being analyzed, N' is the total number of samples used in $R_l(m)$ computation, $w(n)$ is the hamming window, l represents the starting sample index of the frame and M_0 represents the total number of autocorrelation points. In pitch estimation, N' is normally set as $N' = N - m$, so that only N samples of the frame $(x(l), x(l+1), \ldots, x(l+N-1))$ are used for autocorrelation estimation. From the autocorrelation

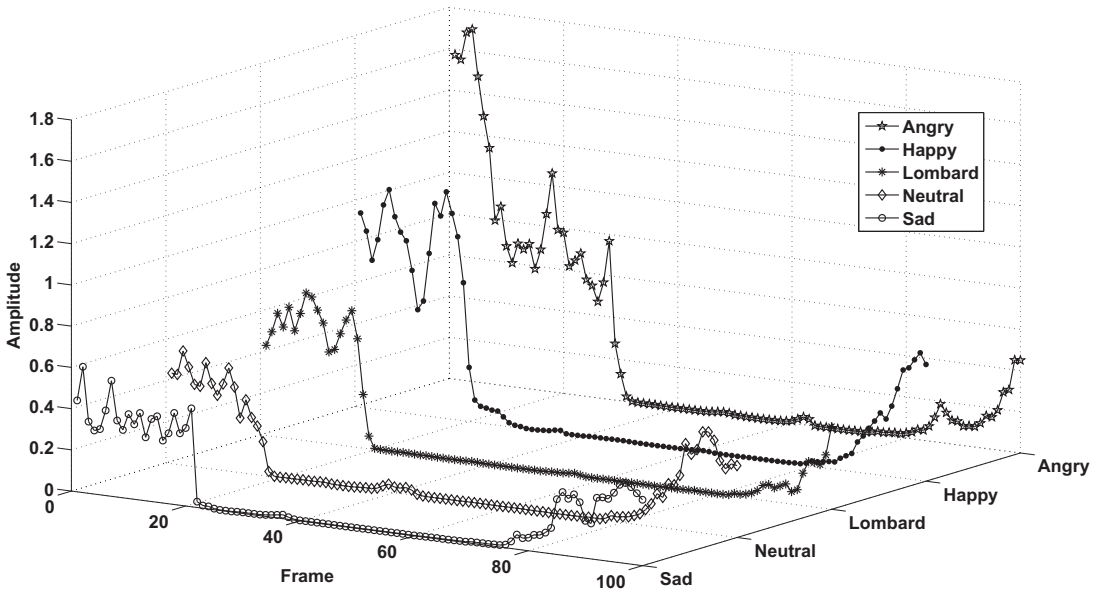

Fig. 1. Contours of the mean of the HPER feature with five different stress classes.

function, the pitch period (T_0) is computed by finding the time lag of the second largest peak from the central peak using the peak picking algorithm. After that, pitch frequency (f_0) is obtained as

$$f_0 = \frac{1}{T_0}. \tag{3}$$

Similarly pitch frequency is calculated for all the speech frames.

iv) The median of pitch frequencies is calculated by arranging all the pitch frequencies in ascending order and picking the middle one.

v) The Fourier spectra for each frame is estimated using N-point Discrete Fourier Transform (N-point DFT). For any signal $x(n)$, the DFT is given by

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp^{-j\frac{2\pi}{N}kn} \tag{4}$$

where $k = 0, 1, \ldots, N-1$. In this work, 1024-point DFT (N = 1024) is performed for estimation of Fourier spectrum.

vi) From the Fourier spectrum, the first harmonic (H_1) that falls within 5% of the median pitch frequency is estimated [10].

vii) Similar to step (vi), we estimate the ith harmonic (H_i) that falls within 5% of $i \times H_1$ ($i = 2, 3, \ldots, M$). The harmonics are also computed by considering the range that falls within 2% and 10% of $i \times H_1$. The maximum performance is achieved with the range that falls within 5% of $i \times H_1$.

viii) The energy (E) of each speech frame $x_f(n)$ is calculated using Eq. (5).

$$E = \sum_{n=1}^{N} x_f^2(n) \tag{5}$$

where N is the total number of samples in the speech frame.

ix) The harmonic peak to energy ratio ($HPER_i$) is evaluated as

$$HPER_i = \frac{H_i}{E} \tag{6}$$

where $i = 1, 2, \ldots, M$.

Fig. 1 shows the contours of the mean of the HPER feature with five stress classes using simulated stressed speech database (SSD) for one person. The details about the database are explained in Section 2.3. The study of the HPER contours can provide useful information for different stress conditions. The mean values of the HPER features for a fixed number of 100 overlapping frames are used for contour plots. From the figure, it is observed that the amplitude values vary with different stress classes. Contours of the angry and happy classes are higher than those of the Lombard, neutral and sad classes. For all stress classes, the contour of the angry class has highest peak value, where as the contour of the neutral and

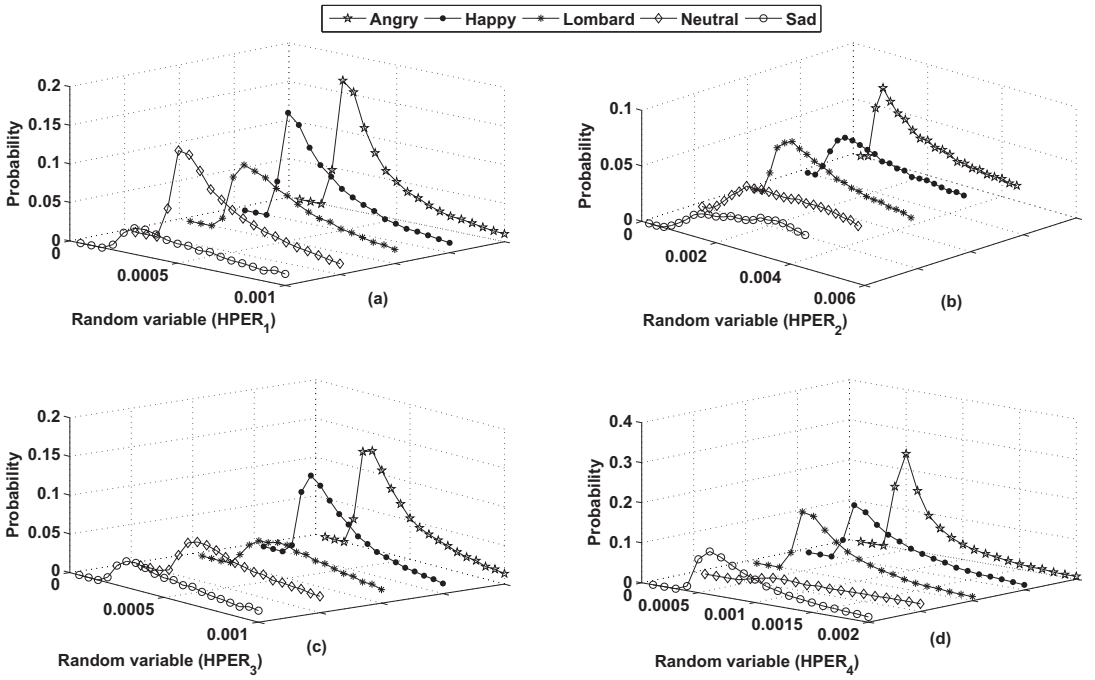

Fig. 2. Probability densities of four HPER features. (a) $HPER_1$ probability densities. (b) $HPER_2$ probability densities. (c) $HPER_3$ probability densities. (d) $HPER_4$ probability densities.

Table 1
Mean and Variance values of four HPER features for five stress classes (Mean and variance values have a multiplication factors of 10^{-4} and 10^{-5}, respectively).

CLASS ↓	$HPER_1$		$HPER_2$		$HPER_3$		$HPER_4$	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Angry	12.2	29.4	10.2	22.0	7.7	9.5	5.7	7.4
Happy	12.4	23.9	11.0	15.5	7.4	7.3	5.5	4.4
Lombard	12.8	19.9	11.9	19.4	8.1	7.7	6.3	4.5
Neutral	12.7	22.2	11.9	17.9	8.9	11.2	6.1	5.7
Sad	12.9	18.3	11.7	11.1	8.7	8.4	6.9	4.9

sad classes have lower peak values. Similar variations have been reported by Ramamohan and Dandapat [13]. These results prompt us to do statistical analysis of the HPER feature.

2.2. Statistical analysis of the HPER feature

The statistical analysis of the HPER feature is useful for analyzing the discrimination capabilities among various stress conditions. The statistical analysis is carried out by evaluating the pdf characteristics and estimating the mean and the variance values. Fig. 2 shows the probability densities (pdf) of four HPER features ($HPER_1$ Fig. 2(a), $HPER_2$ Fig. 2(b), $HPER_3$ Fig. 2(c) and $HPER_4$ Fig. 2(d)) for five different stress classes in 3D graph. The probability densities (pdf) of HPER features are calculated from the training data sets, each training stress class contains approximately 496 speech files. The pdf of random variable describes the relative likelihood to take on a given value. From the figure, it is observed that the pdf characteristics are different for different stress classes. For all the stress classes, angry class has highest peak value as found in contour analysis. It is noticed that the angry class has lower mean values compared to other stress classes, where as sad class has higher mean values. Variance values are also different. These results show that the HPER feature has the capability to distinguish among different stress classes. From the figure, it is further noticed that the pdf characteristics of HPER features are very similar to Gaussian distribution. These probability characteristics explore the qualitative differences among various stress classes. To analyze the quantitative differences, the mean and the variance values of the HPER features are evaluated for various stress conditions. Table 1 shows the mean and the variance values of the four HPER features for five different stress classes. It is noticed that different stress classes have different mean and variance values. The mean values of the HPER features are lower for angry and happy classes, compared to other classes. The neutral and sad classes have higher mean values of the HPER features.

Table 2
Results of T-Test using HPER, MFCC, TEO-CB-Auto-Env and LPC features.

Feature index →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
HPER	5.25	9.47	5.75	15.00	14.13	12.42	28.75	36.71	29.75	18.50	15.51	14.34	23.11	21.4	19.81
MFCC	26.52	23.25	31.49	11.93	21.22	17.15	9.65	11.19	11.02	4.55	10.11	5.67	9.19	12.73	11.29
TEO-CB-Auto-Env	11.96	28.93	17.39	6.29	13.84	15.12	6.35	8.37	26.56	19.78	12.92	19.15	12.93	8.36	6.24
LPC	24.89	15.93	12.43	18.93	10.42	18.45	14.93	19.11	21.88	16.43	7.98	10.43	19.12	14.83	9.98

The contour plots (Fig. 1), the pdf characteristics (Fig. 2), and the mean and the variance evaluation (Table 1) show that the proposed HPER feature can differentiate among various stress classes. Statistical measures, T-Test and F-score, are evaluated to further quantify the discrimination capability of the proposed HPER feature with the mel frequency cepstral coefficients (MFCC), the Linear prediction coefficients (LPC) and the TEO-CB-Auto-Env features. The MFCC, TEO-CB-Auto-Env and LPC features have been used for speech under stress analysis [2,6,7]. The MFCCs are evaluated using a filter bank of 22 mel-filters [14,15]. Delta (Δ) of MFCCs and delta-delta ($\Delta\Delta$) of MFCCs are also calculated [16]. Thus, the resulting MFCC feature vector is of 39 dimensions. The TEO-CB-Auto-Env feature is computed using a filter bank of 39 Gabor band-pass filter, followed by TEO operator [2]. After that, each TEO profile is segmented into frames of 20 ms length with a shift of 10 ms, and then the normalized autocorrelation is evaluated for each frame. Therefore, the TEO-CB-Auto-Env feature will be of 39 dimensions (corresponding to 39 Gabor filters). The LPCs are evaluated using LP analysis of the speech signal [17,18]. In LP analysis, current speech sample $s(n)$ is predicted from the last p samples as a linear combination of the samples, and it is given by $\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k)$, where a_k represents the LP coefficient (LPC) and p is the model order. The a_k values are evaluated by minimizing total prediction error. In this work, we have chosen $p = 18$. Normally, value of p is selected based on sampling frequency of speech signal [19] ($p = f_s/1000 + 2 = 16000/1000 + 2 = 18$, where f_s is the sampling frequency of speech). Therefore, the resulting LPCs feature vector will be of 18 dimensions.

T-Test calculates a probability that the two feature vector sets are from different categories [20]. In T-Test, a score, t-value, is calculated. A larger t-value demonstrates higher discrimination capability between the data sets. The t-value between two data sets (X_1 and X_2) is calculated as [20]

$$t = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{V_1^2}{N_1} + \frac{V_2^2}{N_2}}} \tag{7}$$

where $\bar{X_1}$ and $\bar{X_2}$ are the mean of X_1 and X_2 respectively, V_1 and V_2 are their corresponding variance values, N_1 and N_2 represent the number of samples of data sets X_1 and X_2 respectively. Since t-value is calculated between two classes and SSD database contains 5 stress classes, a total of 10 combination is possible (angry vs happy, angry vs Lombard, angry vs neutral, angry vs sad, happy vs Lombard, happy vs neutral, happy vs sad, Lombard vs neutral, Lombard vs sad and neutral vs sad). The t-values, evaluated with one combination (angry vs happy), are explained as follows. Let, X_1 and X_2 be the HPER features of angry and happy classes respectively, and $\hat{X}_1$ and $\hat{X}_2$ are their corresponding means. The variance values, V_1 and V_2, are calculated from the HPER features, X_1 and X_2, respectively. After that, t-value is calculated using Eq. (7). Similarly, t-value is calculated for the remaining 9 combinations. The final t-value is calculated as the average of the t-values obtained with all the 10 combinations. Table 2 shows the t-values, evaluated on 15 HPER features, 15 MFCC features, 15 TEO-CB-Auto-Env features and 15 LPC features. It is observed that 10 HPER features have higher t-values than MFCC features, 9 HPER features have higher t-values than TEO-CB-Auto-Env features and 10 HPER features have higher t-values than LPC features. That means, t-values of the proposed HPER features are higher compared to those of the MFCC and the TEO-CB-Auto-Env features in majority cases. These T-test results suggest that the stress information can be better captured using HPER feature, compared to that captured using MFCC, TEO-CB-Auto-Env and LPC features. In F-score, a score value is calculated between two data sets [21]. Higher score value implies that the two data sets are more discriminating. The F-score between two classes (X_1 and X_2) is defined as

$$F(i) = \frac{(\mu_{1i} - \mu_i)^2 + (\mu_{2i} - \mu_i)^2}{\frac{1}{N_1-1}\sum_{k=1}^{N_1}(x_{1k,i} - \mu_{1i})^2 + \frac{1}{N_2-1}\sum_{k=1}^{N_2}(x_{2k,i} - \mu_{2i})^2} \tag{8}$$

where μ_{1i} and μ_{2i} are the mean of the ith feature of the X_1 and X_2 respectively, μ_i is the mean of the ith feature of the whole datasets, $x_{1k,i}$ and $x_{2k,i}$ are the ith feature of the kth instance of the datasets, X_1 and X_2, respectively. The datasets X_1 and X_2 have the N_1 and N_2 number of instances respectively. As the F-score is calculated between two classes, therefore, F-score is evaluated for all the 10 combinations as discussed during T-test evaluation. The final F-score is the average of the F-score values obtained with all the combinations. Table 3 shows the F-score values of 15 HPER features, 15 TEO-CB-Auto-Env features, 15 MFCC features and 15 LPC features. It is observed that, in majority cases, the HPER features have higher score values than the MFCC features, the TEO-CB-Auto-Env features and the LPC features. These results, evaluated using T-Test and F-score, suggest that the proposed HPER feature can have better discrimination capability than the MFCC, the TEO-CB-Auto-Env and the LPC features.

Table 3
F-score values using HPER, MFCC, TEO-CB-Auto-Env and LPC features (Each score value has a multiplication factor of 10^{-3}).

Feature index →	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
HPER	0.73	0.52	6.10	6.25	25.68	36.33	32.88	20.98	6.55	2.37	4.46	7.30	10.33	13.90	20.38
MFCC	27.18	3.98	46.60	5.51	26.33	2.93	33.50	2.73	1.05	2.45	0.11	0.02	0.07	0.03	0.12
TEO-CB-Auto-Env	27.81	59.65	26.13	33.10	10.01	5.62	23.75	20.15	1.14	10.35	3.53	1.18	1.65	1.08	1.65
LPC	24.54	15.62	17.39	4.29	4.10	5.10	6.63	4.23	5.43	8.93	3.86	5.43	11.99	8.43	12.93

2.3. Database

In this work, the simulated stressed speech database (SSD), recorded by Shukla et al. [22], has been used. Fifteen speakers (10 male speakers and 5 female speakers) participated in data recording. Before the data recording, the subjects were suggested to think about the contextual situation where the respective stress conditions (such as angry, happy and sad) are elicited. The sampling frequency, used in data recording, is 16 kHz with a resolution of 16 bits/sample. Recording is done for 33 Hindi keywords. These keywords are semantically balanced. The data are recorded in two sessions. The time-duration between the two sessions is at least one week. The database consists of five stress classes: angry, Lombard, neutral, happy and sad. Each word is recorded with all the five stress classes. Lombard speech is defined as the speech produced under noisy conditions. To record the Lombard speech, additive noise is played through the headphone to the subject. All the human subjects participated to produce all the five stress classes in their speech. The database contains approximately 3100 speech files. Each of the stress classes has approximately 620 speech files.

2.4. Support vector machine(SVM) classifier

Support Vector Machine (SVM) classifier has been used extensively for speech under stress classification [4,23]. Use of convex quadratic optimization makes it possible in achieving a globally optimal solution. In SVM, kernel function is used. Due to this, data vectors are mapped into a higher dimensional space, where the data vectors can be linearly separable. This is an advantage of using SVM classifier compared to other classifiers. SVM is basically a binary classifier. For multi-class problem, it is accomplished through three different strategies, “one-against-one”, “one-against-all” and “DAGSVM (binary cascade schema)” [24].

The HPER features ($\mathbf{f}_l \in \mathbf{R}^M$), extracted from the speech signal, are arranged in matrix $\mathbf{F} \in \mathbf{R}^{L \times M}$, and then send to the SVM input, where L represents the total number of feature examples and M is the feature length. Each feature example has a label $y_l \in \{-1, 1\}$. At the training stage, the primal optimization problem is stated as [24,25]

$$\text{minimize}\left(\frac{\mathbf{w}^{\mathbf{T}}\mathbf{w}}{2} + C\sum_{l=1}^{L}\varepsilon_l\right) \tag{9}$$

subject to:

$$y_l(\mathbf{w}^{\mathbf{T}}\phi(\mathbf{f}_l) + b) \geqslant 1 - \varepsilon_l \tag{10}$$

$$\varepsilon_l \geqslant 0 \tag{11}$$

where C is constant and it is called as regularization parameter. The Lagrangian can be used to solve the primal problem [25]. After the Lagrangian, the dual optimization problem is given by [25]

$$\text{maximize } Q(\alpha) = \sum_{l=1}^{L}\alpha_l - \frac{1}{2}\sum_{l=1}^{L}\sum_{j=1}^{L}\alpha_l\alpha_j y_l y_j \mathbf{K}(\mathbf{f}_l, \mathbf{f}_j) \tag{12}$$

subject to:

$$0 \leq \alpha_l \leq C \tag{13}$$

$$\sum_{l=1}^{L}\alpha_l y_l = 0 \tag{14}$$

The solution of the above dual problem gives the Lagrange multiplier (α_l). Using this value of α_l, the weight ($\mathbf{w} = \sum_{l=1}^{L}\alpha_l y_l \mathbf{f}_l$) and bias ($b = y_l - \mathbf{w}^{\mathbf{T}}\phi(\mathbf{f}_l)$) are evaluated. At the testing stage, the mth example output (y_m) of the given test vector $\mathbf{f}_{test}$ is given as

$$y_m = \text{sgn}\left[\sum_{l=1}^{l^*}\alpha_l y_l \mathbf{K}(\mathbf{f}_l, \mathbf{f}_{test})\right] \tag{15}$$

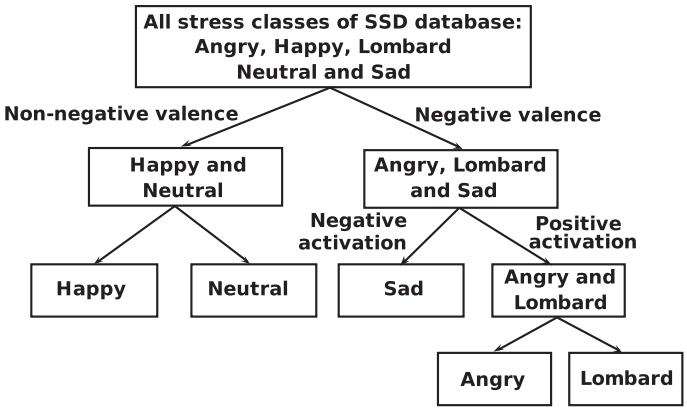

Fig. 3. Binary cascade schema for multi-class classification using SVM.

where l^* and $\mathbf{K}(\mathbf{f}_l, \mathbf{f}_{test})$ correspond to the total number of support vectors and the kernel function respectively. In this work, we have used radial basis kernel function and it is given by [26,27]

$$\mathbf{K}(\mathbf{f}_l, \mathbf{f}_{test}) = \exp\left(\frac{\|\mathbf{f}_l - \mathbf{f}_{test}\|^2}{2\sigma^2}\right) \tag{16}$$

where σ represent the Gaussian width. To select the super-parameters (C and σ), the SVM classifier is trained using 80% of speech files and then it is validated using remaining 20% speech files with different values of C and σ. The values of $C = 1$ and $\sigma = 0.48$ provide minimum training error. The SVM classifier is tested with three different kernel function, the radial basis kernel function, the polynomial kernel function and the linear kernel function. Maximum performance is achieved with the radial basis kernel function with $C = 1$ and $\sigma = 0.48$.

2.5. Binary-cascade multi-class classification approach

In this subsection, the binary-cascade multi-class classification schema for SSD database, based on the dimensional descriptions of different stress conditions, are analyzed. The stress classes can be divided based on the dimension descriptors, valence-activation. The first descriptor is valence, which is associated with a measure of pleasure, ranging from positive to negative. The second descriptor is activation, which is a measure of how dynamic the stress or emotional state is. Based on the descriptor valence, the stress classes are divided into two categories, positive-valence and negative-valence. The positive-valence category includes happy, neutral, surprise and pleasure classes. On the other hand, the negative-valence category includes angry, sad, fear, disgust, despair and nervous. In this work, we have used binary cascade strategy for multi-class classification using the dimensional descriptor, valence-activation, because it has several advantages [23]. Fig. 3 shows the approach of multi-class classification using the binary cascade schema for SSD database. The SSD database contains five stress classes, angry, Lombard, sad, neutral and happy. At first, a classification is made between the stress classes based on negative and non-negative valence. The negative valence category includes angry, Lombard and sad classes, where as non-negative valence category contains happy and neutral classes. Next, a classification is carried out based on the valence-activation. The positive activation contains angry and Lombard classes, where as sad belongs to the negative activation. And at the last stage, we classify the positive activation category between angry and Lombard classes.

3. Results and discussions

In Section 2, the significance of the HPER feature is analyzed using the contour plots, the pdf characteristics, the quantitative results (means and variances) and the statistical measures T-Test and F-score. The performance of the HPER feature for classification of speech under stress is evaluated in this section.

3.1. Performance analysis of the HPER feature

The performance of the HPER feature is analyzed using SVM classifier with binary-cascade multi-class classification approach. Comparison is performed with the mel frequency cepstral coefficients (MFCC), the TEO-CB-Auto-Env and the Linear prediction coefficients (LPC) features. The performance is also compared with the combination of the HPER and the MFCC features. There are equal weights while combining the both of these features. The HPER feature, used in this work, is of 15 dimensions, where as both the MFCC and TEO-CB-Auto-Env feature vectors have 39 attributes. The LPC feature dimension is 18. All the classification results are presented using 5-fold cross-validation method. In 5-fold cross-validation, the complete data set is divided into 5-subsets, and hold-out validation are repeated 5 times. Each time, one subset is used as the testing

Table 4
Confusion matrix (%) for classification of speech under stress using HPER, MFCC, TEO-CB-Auto-Env, and the combination of HPER and MFCC features.

HPER feature					
STRESS	Angry	Happy	Lombard	Neutral	Sad
Angry	**82**	10	6	1	1
Happy	8	**79**	0	11	2
Lombard	13	2	**83**	0	2
Neutral	0	9	0	**84**	7
Sad	2	0	0	3	**95**
Average accuracy = 84.6					
MFCC feature					
STRESS	Angry	Happy	Lombard	Neutral	Sad
Angry	**74**	13	8	2	3
Happy	10	**77**	2	7	4
Lombard	9	3	**86**	0	2
Neutral	2	6	0	**87**	5
Sad	5	3	0	9	**83**
Average accuracy = 81.4					
TEO-CB-Auto-Env feature					
STRESS	Angry	Happy	Lombard	Neutral	Sad
Angry	**59**	13	19	4	5
Happy	21	**73**	0	4	2
Lombard	19	4	**66**	6	5
Neutral	6	11	0	**69**	14
Sad	7	2	2	18	**71**
Average accuracy = 67.6					
LPC feature					
STRESS	Angry	Happy	Lombard	Neutral	Sad
Angry	**54**	17	21	4	4
Happy	20	**62**	2	11	5
Lombard	16	3	**71**	6	4
Neutral	11	23	2	**57**	7
Sad	5	3	3	10	**79**
Average accuracy = 64.6					
HPER feature + MFCC feature					
STRESS	Angry	Happy	Lombard	Neutral	Sad
Angry	**86**	9	4	0	1
Happy	8	**80**	1	10	1
Lombard	7	2	**89**	0	2
Neutral	0	6	0	**91**	3
Sad	3	0	0	3	**94**
Average accuracy = 88.0					

set and the other 4 subsets are used together as a training set. The final accuracy is the average of the accuracies obtained along each of the five subsets.

Table 4 shows the classification results obtained with the HPER feature, the MFCC feature, the TEO-CB-Auto-Env feature, the LPC feature, and the combination of the HPER and the MFCC features. The HPER feature produces the higher recognition rates for angry (82%), happy (79%) and sad (95%) classes, compared to those obtained with the MFCC, the TEO-CB-Auto-Env and the LPC feature. The average recognition rate obtained with the HPER feature is 84.6%, which is higher than those achieved with the TEO-CB-Auto-Env, the MFCC and the LPC features. It has been further observed that the system performance increases significantly with the combination of the HPER and the MFCC features. The combination of the HPER and the MFCC features shows significant increase in performance for angry (86%), Lombard (89%) and neutral (91%) classes. The average recognition rate of 88% is achieved with the combined features, which is higher than the average recognition rates obtained with the HPER feature (84.6%) and the MFCC feature (81.4%). The HPER feature normally captures the intensity variations around the harmonics with respect to energy. The MFCC feature normally captures the information related to the vocal tract shape, and the perceptual mechanism (mel-scale) is incorporated during the calculation of the MFCC feature. The combination of these two features (HPER+MFCC) captures the intensity variations as well as vocal tract and perceptual information. This may be the reason for better recognition using the combination of the HPER and MFCC features. It is further observed that the HPER feature gives a higher recognition rate of 95% for sad class, compared to other classes. The HPER

Table 5
The best feature among HPER, MFCC, TEO-CB-Auto-Env and HPER + MFCC features.

STRESS ↓	Feature ↓
Angry	HPER + MFCC
Happy	HPER + MFCC
Lombard	HPER + MFCC
Neutral	HPER + MFCC
Sad	HPER

feature normally captures the variations of harmonic intensities with respect to the total energy. The sad class has more impact on the low frequency regions [8]. Due to this, lower harmonic components may have higher amplitude intensities than higher harmonics for sad class. Therefore, the variations in lower harmonics and higher harmonics with respect to the total energy are more discriminating for sad class. This is the possible reason for higher recognition rate for sad class. When the HPER feature is combined with the MFCC feature, the recognition performance of sad class decreases compared to that obtained with the HPER feature. To analyze this, t-values are calculated between sad and other classes. The average t-values between sad and other classes with the HPER feature is 17.79, where as for the MFCC feature, it is 12.12. When the HPER and MFCC features are combined, the average t-value becomes 15.01. The average t-value decreases from 17.79 (with HPER feature) to 15.01 (with HPER+MFCC features). This result suggests the higher recognition of sad class using the HPER feature compared to the combined features (HPER + MFCC). We also combined the HPER feature with the TEO-CB-Auto-Env and LPC features, but its impact on the recognition accuracy was little. Hence, we have reported the performance obtained with the combination of the HPER and the MFCC features only. In order to analyze the performance of the HPER feature with different sessions (i.e. with session variability), the SVM is trained with one session and tested with the other session. When the model is trained with session 1 data and tested with session 2 data, the average accuracy obtained is 77.92%. On the other hand, when the model is trained with session 2 data and the model is tested with session 1 data, the average accuracy obtained is 79.19%. Table 5 shows the optimal combination of the features among HPER, MFCC, TEO-CB-Auto-Env, LPC and HPER+MFCC for different stress classes. For all stress classes, either the HPER feature or the combination of HPER and MFCC features provide higher performance. These results establish that the proposed HPER feature is capable of classifying different emotions.

In summary, the proposed HPER feature improves the speech under stress classification by 3.93% with respect to the MFCC feature. The classification performance is further enhanced by 4.02% with the combination of the HPER and the MFCC features.

3.2. Performance comparisons with other classification methods

In Section 3.1, the performance of the HPER feature is analyzed and compared with the MFCC, TEO-CB-Auto-Env and LPC features the using SVM classifier. This section discusses the performance comparisons of the HPER feature using SVM classifier with those obtained with Hidden Markov Model (HMM), Gaussian Mixture Model (GMM) and Extreme Learning Machine (ELM) classifiers. The HMM and GMM classifiers have been used for classification of speech under stress [2,3,8]. In this work, a separate HMM is trained for each class. The HMM model is trained with 3-state left-to-right strategy with 256 mixtures per state [17,28]. The model is also tested with different mixtures and the maximum performance is achieved with 256 mixtures. A Gaussian Mixture Model (GMM) is defined as the weighted sum of M Gaussians. During training, the GMM parameters are updated using the expectation maximization (EM) algorithm. During testing, a likelihood value is calculated for each class and the test vector is assigned to the class which gives the maximum likelihood value. Extreme Learning Machine (ELM) model was first proposed for the single-hidden-layer feed-forward neural networks (SLFNs), and then it is extended to the generalized SLFNs [29]. The ELM output of generalized SLFNs is defined as $f_L(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta} = \sum_{\mathbf{i=1}}^{\mathbf{L}} \beta_\mathbf{i}\mathbf{h_i}(\mathbf{x})$, where $\boldsymbol{\beta}$ ($\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_L]^T$) is the weight vector between the hidden layer nodes and the output node, and $\mathbf{h}(\mathbf{x}) = [\mathbf{h_1}(\mathbf{x}), \mathbf{h_2}(\mathbf{x}), \ldots, \mathbf{h_L}(\mathbf{x})]$ represents the hidden layer output vector corresponding to the input vector $\mathbf{x}$. The value of $\boldsymbol{\beta}$ is obtained by solving the constrained optimization problem [29] and it is given by $\boldsymbol{\beta}^* = \mathbf{H}^*\mathbf{X}$, where $\mathbf{H}^*$ represents the Moore-Penrose generalized inverse matrix of the hidden layer output matrix $\mathbf{H}$. During testing, the decision is given by $D(\mathbf{x}) = \text{sgn}(\mathbf{h}(\mathbf{x})\boldsymbol{\beta}^*)$.

The recognition accuracies obtained with different features using HMM, GMM, ELM and SVM classifiers are shown in Table 6. For all the classifiers, the maximum average accuracy is obtained with HPER feature, compared to the MFCC, TEO-CB-Auto-Env and LPC features. The combination of the HPER and MFCC features further increases the recognition performance for all the classifiers. The HMM classifier with combined features (HPER+MFCC) gives an average recognition rate of 71.4%, which is higher than that obtained with GMM classifier (70.0%). HMM normally captures the temporal information (because of state-transition matrix), where as GMM is one-state HMM i.e. no state transition is in GMM. This is the reason for higher recognition rate of the HMM classifier than the GMM classifier. For all the classifiers using combined features (HPER+MFCC), the maximum recognition performance of 88.0% is achieved with SVM classifier. SVM is basically a binary classifier. In this work, SVM is used in binary-cascade multi-class classification approach, where SVM is used as a binary

Table 6
Recognition accuracies (in %) with different features using HMM, GMM, ELM and SVM classifiers (TEO† = TEO-CB-Auto-Env).

HMM Classifier						GMM Classifier					
Feature	HPER	MFCC	TEO†	LPC	HPER+MFCC	Feature	HPER	MFCC	TEO†	LPC	HPER+MFCC
Angry	69	64	70	61	76	Angry	58	70	66	50	70
Happy	63	62	64	54	77	Happy	69	67	64	42	68
Lombard	74	68	72	56	70	Lombard	75	69	70	49	73
Neutral	68	64	64	41	73	Neutral	65	63	67	43	67
Sad	70	70	69	68	61	Sad	71	66	67	65	72
Average	68.8	65.6	67.8	56.0	71.4	**Average**	67.6	67.0	66.8	49.8	70.0
ELM Classifier						SVM Classifier					
Feature	HPER	MFCC	TEO†	LPC	HPER+MFCC	Feature	HPER	MFCC	TEO†	LPC	HPER+MFCC
Angry	83	87	83	78	88	Angry	82	74	59	54	86
Happy	71	65	62	66	74	Happy	79	77	73	62	80
Lombard	60	60	62	58	61	Lombard	83	86	66	71	89
Neutral	73	72	74	69	75	Neutral	84	87	69	57	91
Sad	85	78	78	71	85	Sad	95	83	71	79	94
Average	74.4	72.4	71.8	68.4	76.6	**Average**	84.6	81.4	67.6	64.6	88.0

classifier. The SVM also uses kernel function, due to which linear separation of the data vectors is possible. This may be the reason for higher recognition rate of SVM classifier.

4. Conclusion

From the above analysis, it is concluded that the HPER feature characterizes different stress conditions. The HPER feature captures the intensity variations of different harmonics as well as breathiness information, and these have been used for classification of speech under different stress conditions. The significance of the HPER feature for speech under stress classification is established using statistical analysis. The performance analysis is done using SVM classifier. Binary-cascade multi-class classification approach is used based on the valence-activation scale of different stress categories. In terms of classification rates, the HPER feature successfully classify different stress classes. There is a significant increase in performance with the combination of the HPER feature and the MFCC feature. The experiment result establishes the effectiveness of the HPER feature for analysis and classification of speech under stress.

References

[1] Hansen JH, Patil S. Speaker classification I. In: Speech under stress: analysis, modeling and recognition. Berlin, Heidelberg: Springer-Verlag; 2007. p. 108–37. ISBN 978-3-540-74186-2.
[2] Zhou G, Hansen J, Kaiser J. Nonlinear feature based classification of speech under stress. Speech Audio Process IEEE Trans 2001;9(3):201–16. doi:10.1109/89.905995.
[3] Ververidis D, Kotropoulos C. Emotional speech recognition: resources, features, and methods. Speech Commun 2006;48(9):1162–81.
[4] Wang K, An N, Li BN, Zhang Y, Li L. Speech emotion recognition using fourier parameters. Affect Comput IEEE Trans 2015;6(1):69–75. doi:10.1109/TAFFC.2015.2392101.
[5] Mariooryad S, Busso C. Compensating for speaker or lexical variabilities in speech for emotion recognition. Speech Commun 2014;57:1–12.
[6] Bou-Ghazale S, Hansen J. A comparative study of traditional and newly proposed features for recognition of speech under stress. Speech Audio Process IEEE Trans 2000;8(4):429–42. doi:10.1109/89.848224.
[7] Sezgin MC, Gnsel B, Kurt GK. Perceptual audio features for emotion detection. EURASIP J Audio Speech Music Process 2012;2012:16.
[8] Zao L, Cavalcante D, Coelho R. Time-frequency feature and AMS-GMM mask for acoustic emotion classification. Signal Process Lett IEEE 2014;21(5):620–4. doi:10.1109/LSP.2014.2311435.
[9] Yao X, Jitsuhiro T, Miyajima C, Kitaoka N, Takeda K. Modeling of physical characteristics of speech under stress. Signal Process Lett IEEE 2015;22(10):1801–5. doi:10.1109/LSP.2015.2434732.
[10] Hillenbrand J, Houde RA. Acoustic correlates of breathy vocal quality: dysphonic voices and continuous speech. J Speech Lang Hear Res 1996;39(2):311–21.
[11] Castillo-Guerra E, Ruiz A. Automatic modeling of acoustic perception of breathiness in pathological voices. Biomed Eng IEEE Trans 2009;56(4):932–40. doi:10.1109/TBME.2008.2007910.
[12] Rabiner LR. On the use of autocorrelation analysis for pitch detection. Acoust Speech Signal Process IEEE Trans 1977;25(1):24–33.
[13] Ramamohan S, Dandapat S. Sinusoidal model-based analysis and classification of stressed speech. Audio Speech Lang Process IEEE Trans 2006;14(3):737–46. doi:10.1109/TSA.2005.858071.
[14] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoust Speech Signal Process IEEE Trans 1980;28(4):357–66.
[15] Kopparapu SK, Bhuvanagiri KK. Recognition of subsampled speech using a modified mel filter bank. Comput Electr Eng 2013;39(2):655–62.
[16] Soong FK, Rosenberg AE. On the use of instantaneous and transitional spectral information in speaker recognition. Acoust Speech Signal Process IEEE Trans 1988;36(6):871–9.
[17] Rabiner LR, Juang B-H. Fundamentals of speech recognition, 14. PTR Prentice Hall Englewood Cliffs; 1993.
[18] Rahdari F, Eftekhari M, Mousavi R. A two-level multi-gene genetic programming model for speech quality prediction in voice over internet protocol systems. Comput Electr Eng 2016;49:9–24.
[19] Prathosh AP, Ananthapadmanabha TV, Ramakrishnan AG. Epoch extraction based on integrated linear prediction residual using plosion index. IEEE Trans Audio Speech Lang Process 2013;21(12):2471–80. doi:10.1109/TASL.2013.2273717.
[20] Zhang R, Li Y, Li X. Topology inference with network tomography based on *t*-test. Commun Lett IEEE 2014;18(6):921–4. doi:10.1109/LCOMM.2014.2317743.

[21] Polat K, Güneş S. A new feature selection method on classification of medical datasets: Kernel f-score feature selection. Expert Syst Appl 2009;36(7):10367–73.
[22] Shukla S, Dandapat S, Prasanna SR. Spectral slope based analysis and classification of stressed speech. Int J Speech Technol 2011;14(3):245–58. doi:10.1007/s10772-011-9100-x.
[23] Kotti M, Paternò F. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. Int J Speech Technol 2012;15(2):131–50.
[24] Hsu C-W, Lin C-J. A comparison of methods for multiclass support vector machines. Neural Netw IEEE Trans 2002;13(2):415–25. doi:10.1109/72.991427.
[25] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[26] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press; 2000.
[27] Mahendran G, Dhanasekaran R. Investigation of the severity level of diabetic retinopathy using supervised classifier algorithms. Comput Electr Eng 2015;45:312–23.
[28] Firooz SG, Almasganj F, Shekofteh Y. Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals. Comput Electr Eng 2016.
[29] Huang G-B, Zhou H, Ding X, Zhang R. Extreme learning machine for regression and multiclass classification. Syst Man Cybernet Part B 2012;42(2):513–29.

Suman Deb received his M.Tech. degree in signal processing from the Indian Institute of Technology Guwahati, India, in 2013. He is currently working as a Ph.D. student in the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, India. His research interest includes signal processing, speech signal processing, and pattern recognition.

S. Dandapat received the Ph.D. degree in electrical engineering from the Indian Institute of Technology Kanpur, India, in 1997. He is currently a Professor in the Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, India. His current research interests include digital signal processing, speech processing, biomedical signal processing, and medical image processing.