Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 394083, 13 pages
http://dx.doi.org/10.1155/2015/394083

Research Article

Improved Emotion Recognition Using Gaussian Mixture Model and Extreme Learning Machine in Speech and Glottal Signals

Hariharan Muthusamy,[1] Kemal Polat,[2] and Sazali Yaacob[3]

[1]*School of Mechatronic Engineering, Universiti Malaysia Perlis (UniMAP), Campus Pauh Putra, 02600 Perlis, Perlis, Malaysia*
[2]*Department of Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Abant Izzet Baysal University, 14280 Bolu, Turkey*
[3]*Universiti Kuala Lumpur Malaysian Spanish Institute, Kulim Hi-Tech Park, 09000 Kulim, Kedah, Malaysia*

Correspondence should be addressed to Hariharan Muthusamy; hari@unimap.edu.my

Received 25 August 2014; Accepted 29 December 2014

Academic Editor: Gen Qi Xu

Recently, researchers have paid escalating attention to studying the emotional state of an individual from his/her speech signals as the speech signal is the fastest and the most natural method of communication between individuals. In this work, new feature enhancement using Gaussian mixture model (GMM) was proposed to enhance the discriminatory power of the features extracted from speech and glottal signals. Three different emotional speech databases were utilized to gauge the proposed methods. Extreme learning machine (ELM) and *k*-nearest neighbor (*k*NN) classifier were employed to classify the different types of emotions. Several experiments were conducted and results show that the proposed methods significantly improved the speech emotion recognition performance compared to research works published in the literature.

1. Introduction

Spoken utterances of an individual can provide information about his/her health state, emotion, language used, gender, and so on. Speech is the one of the most natural form of communication between individuals. Understanding of individual's emotion can be useful for applications like web movies, electronic tutoring applications, in-car board system, diagnostic tool for therapists, and call center applications [1–4]. Most of the existing emotional speech database contains three types of emotional speech recordings such as simulated, elicited, and natural ones. Simulated emotions tend to be more expressive than real ones and most commonly used [4]. In the elicited category, emotions are nearer to the natural database but if the speakers know that they are being recorded, the quality will be artificial. Next, in the natural category, all emotions may not be available and difficult to model because these are completely naturally expressed. Most of the researchers have analysed four primary emotions such as anger, joy, fear, and sadness either in stimulated domain or in natural domain. High emotion recognition accuracies were obtained for two-class emotion recognition (high arousal versus low arousal), but multiclass emotion recognition is still disputing. This is due to the following reasons: (a) which speech features are information-rich and parsimonious, (b) different sentences, speakers, speaking styles, and rates, (c) more than one perceived emotion in the same utterance, and (d) long-term/short-term emotional states [1, 3, 4].

To improve the accuracy of multiclass emotion recognition, a new GMM based feature enhancement was proposed and tested using 3 different emotional speech databases (Berlin emotional speech database (BES), Surrey audio-visual expressed emotion (SAVEE) database, and Sahand Emotional Speech database (SES)). Both speech signals and its glottal waveforms were used for emotion recognition experiments. To extract the glottal and vocal tract characteristics from the speech waveforms, several techniques have been proposed [5–8]. In this work, we extracted the glottal waveforms from the emotional speech signals by using inverse filtering and linear predictive analysis [5, 6, 9]. Emotional speech signals and its glottal waveforms were decomposed into 4 levels using discrete wavelet packet transform and relative

energy and entropy features were calculated for each of the decomposition nodes. A total of 120 features were obtained. Higher degree of overlap between the features of different classes may degrade the performance of classifiers which results in poor recognition of speech emotions. To decrease the intraclass variance and to increase the interclass variance among the features, GMM based feature enhancement was proposed which results in improved recognition of speech emotions. Both raw and enhanced features were subjected to several experiments to validate their effectiveness in speech emotion recognition. The rest of this paper is organized as follows. Some of the significant works on speech emotion recognition are discussed in Section 2. Section 3 presents the materials and methods used. Experimental results and discussions are presented in Section 4. Finally, Section 5 concludes the paper.

2. Previous Works

Several speech features have been successfully applied for speech emotion recognition and can be mainly classified into four groups such as continuous features, qualitative features, spectral features, and nonlinear Teager energy operator based features [1, 3, 4]. Various types of classifiers have been proposed for speech emotion recognition such as hidden Markov model (HMM), Gaussian mixture model (GMM), support vector machine (SVM), artificial neural networks (ANN), and k-NN [1, 3, 4]. This section describes some of the recently published works in the area of multiclass speech emotion recognition. Table 1 shows the list of some of the recent works in multiclass speech emotion recognition using BES and SAVEE databases.

Though speech related features are widely used for speech emotion recognition, there is a strong correlation between the emotional states and features derived from glottal waveforms. Glottal waveform is significantly affected by the emotional state and speaking style of an individual [10–12]. In [10–12], researchers have investigated that the glottal waveform was affected due to the excessive tension or lack of coordination in the laryngeal musculature under different emotional states and the speech produced under stress. The classification of clinical depression using the glottal features was carried out by Moore et al. in [13, 14]. In [15], authors have obtained 85% of the correct emotion recognition rate by using the glottal flow spectrum as a possible cue for depression and near-term suicide risk. Iliev and Scordilis have investigated the effectiveness of glottal features derived from the glottal airflow signal in recognizing emotions [16]. The average emotion recognition rate of 66.5% for all six emotions (happiness, anger, sadness, fear, surprise, and neutral) and 99% for four emotions (happiness, neutral, anger, and sadness) was achieved. He et al. have proposed wavelet packet energy entropy features for emotion recognition from speech and glottal signals with GMM classifier [17]. They achieved average emotion recognition rates for BES database between 51% and 54%. In [18], prosodic features, spectral features, glottal flow features, and AM-FM features were utilized and two-stage feature reduction was proposed for speech emotion recognition. The overall emotion recognition rate of 85.18% for gender-dependent and 80.09% for gender-independent was achieved using SVM classifier.

Several feature selection/reduction methods were proposed to select/reduce the course of dimensionality of speech features. Although all the above works are novel contributions to the field of speech emotion recognition, it is difficult to compare them directly since division of datasets is inconsistent: the number of emotions used, the number of datasets used, inconsistency in the usage of simulated or naturalistic speech emotion databases, and lack of uniformity in computation and presentation of the results. Most of the researchers have commonly used 10-fold cross validation and conventional validation (one training set + one testing set) and some of them have tested their methods under speaker-dependent, speaker-independent, gender-dependent and gender-independent environments. In this regard, the proposed methods were validated using three different emotional speech databases and emotion recognition experiments were also conducted under speaker-dependent and speaker-independent environments.

3. Materials and Methods

3.1. Emotional Speech Databases. In this work, three different emotional speech databases were used for emotion recognition and to test the robustness of the proposed methods. First, Berlin emotional speech database (BES) was used which consists of speech utterances in German [19]. 10 professional actors/actresses were used to simulate 7 emotions (anger: 127, disgust: 45, fear: 70, neutral: 79, happiness: 71, sadness: 62, and boredom: 81). Secondly, Surrey audio-visual expressed emotion (SAVEE) database [20] was used and it is an audio-visual emotional database which includes seven emotion categories of speech utterances (anger: 60, disgust: 60, fear: 60, neutral: 120, happiness: 60, sadness: 60, and surprise: 60) from four native English male speakers aged from 27 to 31 years. 3 common, 2 emotion-specific, and 10 generic sentences from 15 TIMIT sentences per emotion were recorded. In this work, only audio samples were utilized. Lastly, Sahand Emotional Speech database (SES) was used [21] and it was recorded at Artificial Intelligence and Information Analysis Lab, Department of Electrical Engineering, Sahand University of Technology, Iran. This database contains speech utterances of five basic emotions (neutral: 240, surprise: 240, happiness: 240, sadness: 240, and anger: 240) from 10 speakers (5 male and 5 female). 10 single words, 12 sentences, and 2 passages in Farsi language were recorded which results in a total of 120 utterances per emotion. Figures 1(a)–1(d) show an example of portion of utterance spoken by a speaker in the four different emotions (neutral, anger, happiness, and disgust). It can be observed from the figures that the structure of the speech signals and its glottal waveforms are considerably different for speech spoken under different emotional states.

3.2. Features for Speech Emotion Recognition. Extraction of suitable features for efficiently characterizing different emotions is still an important issue in the design of a speech

Table 1: Some of the significant works on speech emotion recognition.

Ref. number	Database	Signals	Number of emotions	Methods	Best result
[49]	BES	Speech signals	Anger, boredom, disgust, fear, happiness, sadness, and neutral	Nonlinear dynamic features + prosodic + spectral features + SVM classifier	82.72% (females) 85.90% (males)
[50]	BES	Speech signals	Neutral, fear, and anger	Nonlinear dynamic features + neural network	93.78%
[51]	BES	Speech signals	Anger, boredom, disgust, fear, happiness, sadness, and neutral	Modulation spectral features (MSFs) + multiclass SVM	85.60%
[30]	BES	Speech signals	Anger, boredom, disgust, fear, happiness, sadness, and neutral	Combination of spectral excitation source features + autoassociative neural network	82.16%
[27]	BES	Speech signals	Anger, boredom, disgust, fear, happiness, sadness, and neutral	Combination of utterancewise global and local prosodic features + SVM classifier	62.43%
[52]	BES	Speech signals	Anger, boredom, disgust, fear, happiness, sadness, and neutral	LPCCs + formants + GMM classifier	68%
[28]	BES	Speech signals	Anger, boredom, fear, happiness, sadness, and neutral	Discriminative band wavelet packet power coefficients (db-WPPC) with Daubechies filter of order 40 + GMM classifier	75.64%
[53]	BES	Speech signals	Anger, boredom, disgust, fear, happiness, sadness, and neutral	Low level audio descriptors and high level perceptual descriptors with linear SVM	87.7%
[54]	BES	Speech signals	Anger, boredom, disgust, fear, happiness, sadness, and neutral	MPEG-7 low level audio descriptors + SVM with radial basis function kernel	77.88%
[55]	SAVEE	Speech signals	Anger, surprise, sadness, happiness, fear, disgust, and neutral	Mel-frequency cepstral coefficients + signal energy + correlation based feature selection + SVM with radial basis function kernels	79%
[56]	SAVEE	Speech signals	Anger, surprise, sadness, happiness, fear, disgust, and neutral	Energy intensity + pitch + standard deviation + jitter + shimmer + kNN	74.39%
[57]	SAVEE	Speech signals	Anger, surprise, sadness, happiness, fear, disgust, and neutral	Audio features + LDA feature reduction + single component Gaussian classifier	63%
[20]	SAVEE	Speech signals	Anger, surprise, sadness, happiness, fear, disgust, and neutral	Pitch + energy + duration + spectral + Gaussian classifier	59.2%

emotion recognition system. Short-term features were widely used by the researchers, called frame-by-frame analysis. All the speech samples were downsampled to 8 kHz. The unvoiced portions between words were removed by segmenting the downsampled emotional speech signals into nonoverlapping frames with a length of 32 ms (256 samples) based on the energy of the frames. Frames with low energy were discarded and the rest of the frames (voiced portions) were concatenated and used for feature extraction [17]. Then the emotional speech signals (only voiced portions) are passed through a first-order low pass filter to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing [22]. The first-order preemphasis filter is defined as

$$H(z) = 1 - a * z^{-1} \quad 0.9 \leq a \leq 1.0. \tag{1}$$

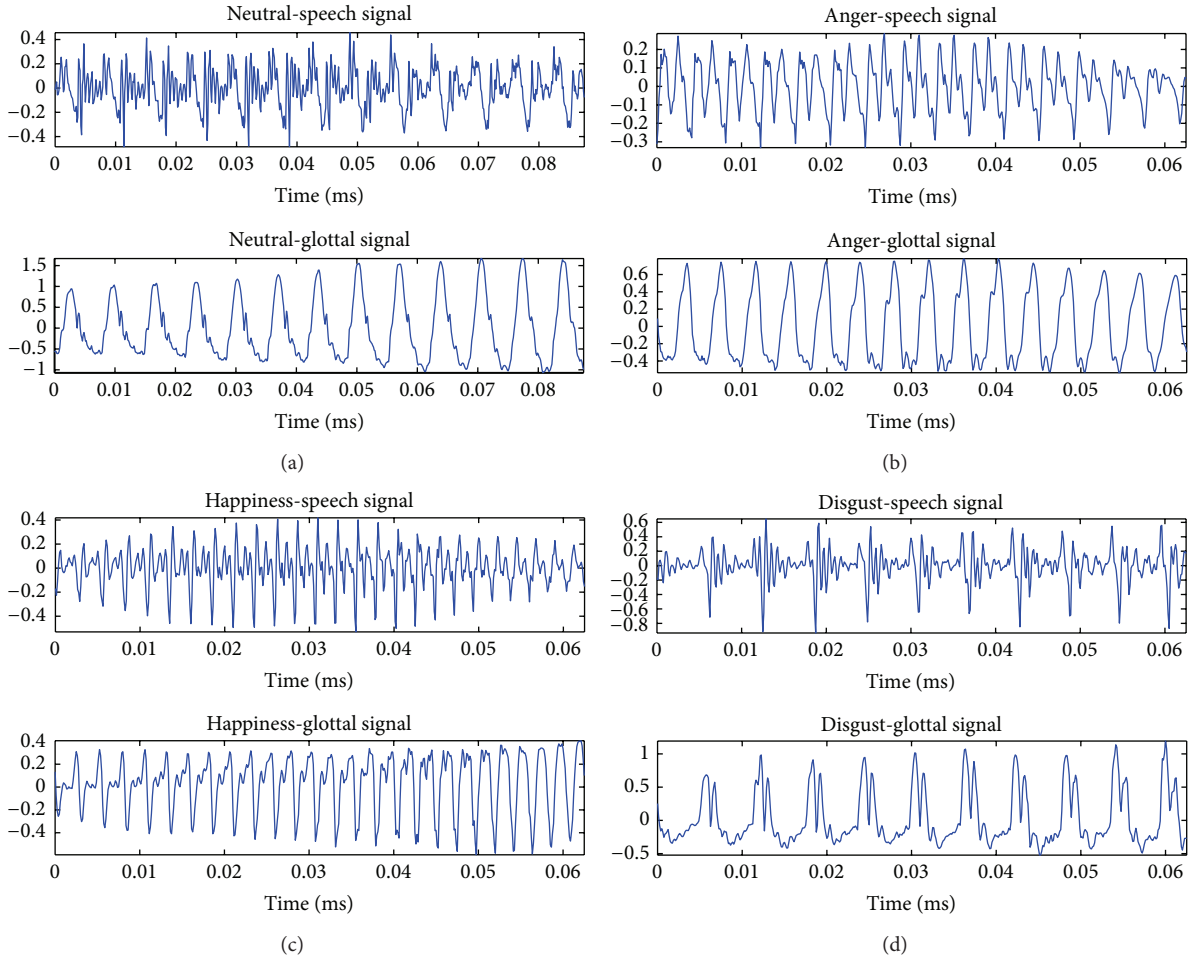

Figure 1: ((a)–(d)) Emotional speech signals and its glottal waveforms (BES database).

The commonly used a value is $15/16 = 0.9375$ or 0.95 [22]. In this work, the value of a is set equal to 0.9375. Extraction of glottal flow signal from speech signal is a challenging task. In this work, glottal waveforms were estimated based on the inverse filtering and linear predictive analysis from the preemphasized speech waveforms [5, 6, 9]. Wavelet or wavelet packet transform has the ability to analyze any nonstationary signals in both time and frequency domain simultaneously. The hierarchical wavelet packet transform decomposes the original emotional speech signals/glottal waveforms into subsequent subbands. In WP decomposition, both low and high frequency subbands are used to generate the next level subbands which results in finer frequency bands. Energy of the wavelet packet nodes is more robust in representing the original signal than using the wavelet packet coefficients directly. Shannon entropy is a robust description of uncertainty in the whole signal duration [23–26]. The preemphasized emotional speech signals and glottal waveforms were segmented into 32 ms frames with 50% overlap. Each frame was decomposed into 4 levels using discrete wavelet packet transform and relative wavelet packet energy and entropy features were derived for each of the decomposition nodes as given in (4) and (7). Consider the following:

$$\mathrm{EGY}_{j,k} = \log_{10}\left(\frac{\sum \left|C_{j,k}\right|^2}{L}\right), \tag{2}$$

$$\mathrm{EGY}_{\mathrm{tot}} = \sum \mathrm{EGY}_{j,k}, \tag{3}$$

$$\text{Relative wavelet packet energy, RWPEGY} = \frac{\mathrm{EGY}_{j,k}}{\mathrm{EGY}_{\mathrm{tot}}}, \tag{4}$$

$$\mathrm{EPY}_{j,k} = -\sum \left|C_{j,k}\right|^2 \log_{10}\left|C_{j,k}\right|^2, \tag{5}$$

$$\mathrm{EPY}_{\mathrm{tot}} = \sum \mathrm{EPY}_{j,k}, \tag{6}$$

$$\text{Relative wavelet packet entropy, RWPEPY} = \frac{\mathrm{EPY}_{j,k}}{\mathrm{EPY}_{\mathrm{tot}}}, \tag{7}$$

where $j = 1, 2, 3, \ldots, m$, $k = 0, 1, 2, \ldots, 2^m - 1$, m is the number of decomposition levels, and L is the length of wavelet packet coefficients at each node (j, k). In this

work, Daubechies wavelets with 4 different orders (db3, db6, db10, and db44) were used since Daubechies wavelets are frequently used in speech emotion recognition and provide better results [1, 17, 27, 28]. After obtaining the relative wavelet packet energy and entropy based features for each frame, they were averaged over all frames and used for analysis. Four-level wavelet decomposition gives 30 wavelet packet nodes and features were extracted from all the nodes which yield 60 features (30 relative energy features + 30 relative entropy features). Similarly, the same features were extracted from emotional glottal signals. Finally, a total of 120 features were obtained.

3.3. Feature Enhancement Using Gaussian Mixture Model. In any pattern recognition applications, escalating the interclass variance and diminishing the intraclass variance of the attributes or features are the fundamental issues to improve the classification or recognition accuracy [24, 25, 29]. In the literature, several research works can be found to escalate the discriminating ability of the extracted features [24, 25, 29]. GMM has been successfully applied in various pattern recognition applications particularly in speech and image processing applications [17, 28–36]; however its capability of escalating the discriminative ability of the features or attributes is not being extensively explored. Different applications of GMM motivate us to suggest GMM based feature enhancement [17, 28–36]. High intraclass variance and low interclass variance among the features may degrade the performance of classifiers which results in poor emotion recognition rates. To decrease the intraclass variance and to increase the interclass variance among the features, GMM based clustering was suggested in this work, to enrich the discriminative ability of the relative wavelet packet energy and entropy features. GMM model is a probabilistic model and its application to labelling is based on the assumption that all the data points are generated from a finite mixture of Gaussian mixture distributions. In a model-based approach, certain models are used for clustering and attempting to optimize the fit between the data and model. Each cluster can be mathematically represented by a Gaussian (parametric) distribution. The entire dataset z is modeled by a weighted sum of M numbers of mixtures of Gaussian component densities and is given by the equation

$$p(z_i \mid \theta) = \sum_{k=1}^{M} \rho_k f(z_i \mid \theta_k), \tag{8}$$

where z is the N-dimensional continuous valued relative wavelet packet energy and entropy features, ρ_k, $k = 1, 2, 3, \ldots, M$, are the mixture weights, and $f(z_i \mid \theta_k)$, $\theta_k = (\mu_k, \Sigma_k)$ $i = 1, 2, 3, \ldots, M$ are the component Gaussian densities. Each component density is an N-variate Gaussian function of the form

$$f(z_i \mid \theta_k) = \frac{1}{(2\pi)^{n/2} \left|\Sigma_i\right|^{1/2}} \cdot \exp\left[-\frac{1}{2}(z - \mu_i)^T \Sigma_i^{-1} (z - \mu_i)\right], \tag{9}$$

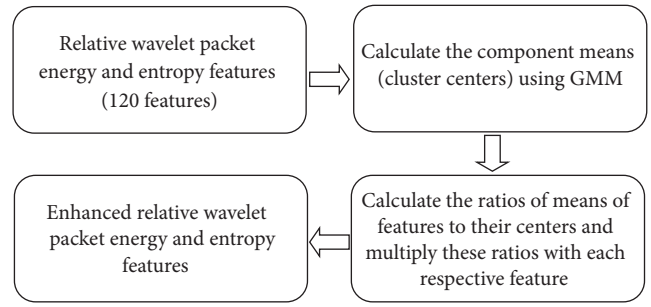

Figure 2: Proposed GMM based feature enhancement.

where μ_i is the mean vector and Σ_i is the covariance matrix. Using the linear combination of mean vectors, covariance matrices, and mixture weights, overall probability density function is estimated [17, 28–36]. Gaussian mixture model uses an iterative expectation maximization (EM) algorithm that converges to a local optimum and assigns posterior probabilities to each component density with respect to each observation. The posterior probabilities for each point indicate that each data point has some probability of belonging to each cluster [17, 28–36]. The working of GMM based feature enhancement is summarized (Figure 2) as follows: firstly, the component means (cluster centers) of each feature belonging to dataset using GMM based clustering method was found. Next, the ratios of means of features to their centers were calculated. Finally, these ratios were multiplied with each respective feature.

After applying GMM clustering based feature weighting method, the raw features (RF) were known as enhanced features (EF).

The class distribution plots of raw relative wavelet packet energy and entropy features were shown in Figures 3(a), 3(c), 3(e), and 3(g) for different orders of Daubechies wavelets ("db3," "db6," "db10," and "db44"). From the figures, it can be seen that there is a higher degree of overlap among the raw relative wavelet packet energy and entropy features which results in the poor performance of the speech emotion recognition system. The class distribution plots of enhanced relative wavelet packet energy and entropy features were shown in Figures 3(b), 3(d), 3(f), and 3(h). From the figures, it can be observed that the higher degree of overlap can be diminished which in turn improves the performance of the speech emotion recognition system.

3.4. Feature Reduction Using Stepwise Linear Discriminant Analysis. Curse of dimensionality is a big issue in all pattern recognition problems. Irrelevant and redundant features may degrade the performance of the classifiers. Feature selection/reduction was used for selecting the subset of relevant features from a large number of features [24, 29, 37]. Several feature selection/reduction techniques have been proposed to find most discriminating features for improving the performance of the speech emotion recognition system [18, 20, 28, 38–43]. In this work, we propose the use of stepwise linear discriminant analysis (SWLDA) since LDA is a linear technique which relies on the mixture model containing

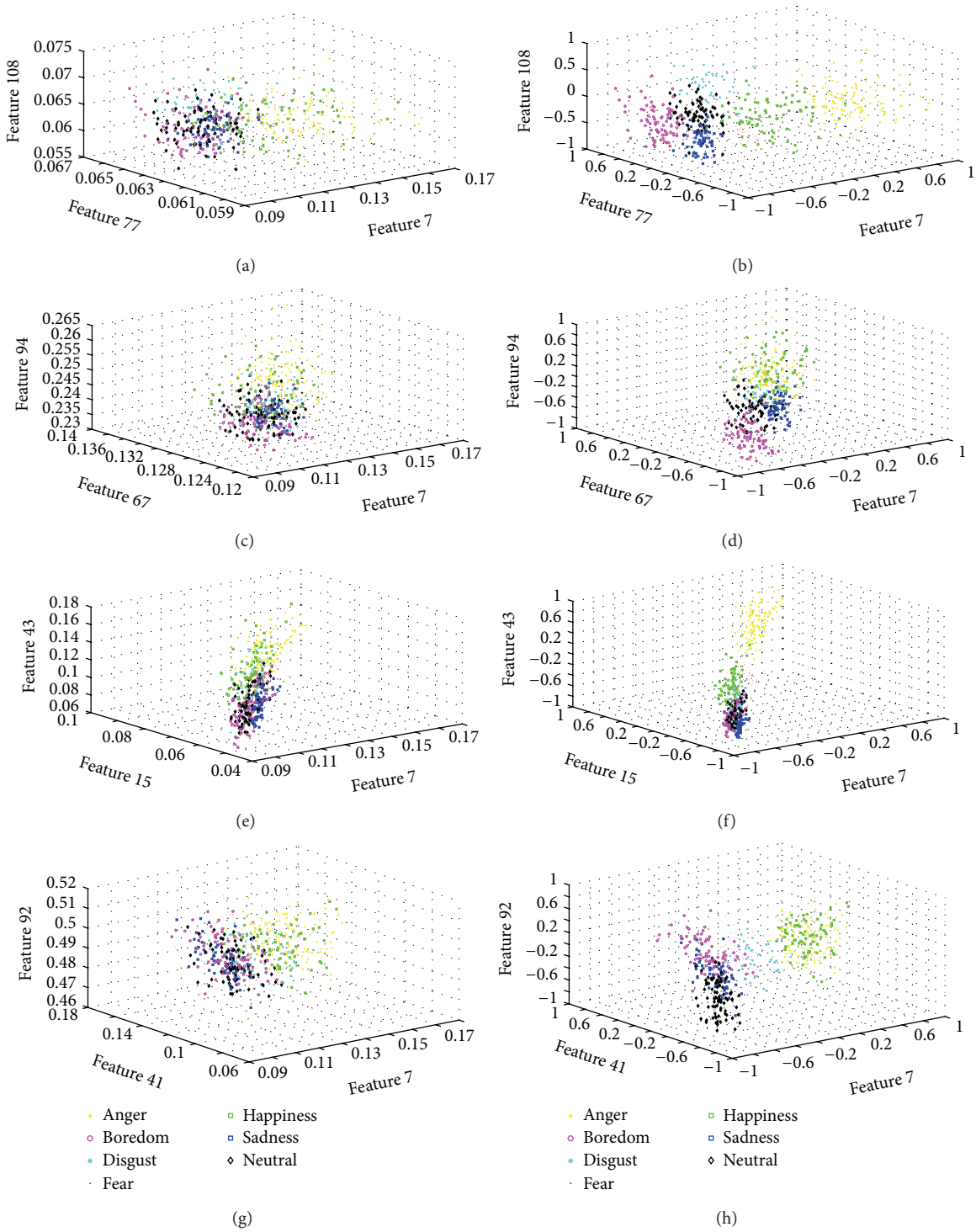

Figure 3: ((a), (c), (e), and (g)) Class distribution plots of raw features for "db3," "db6," "db10," and "db44." ((b), (d), (f), and (h)) Class distribution plots of enhanced features for "db3," "db6," "db10," and "db44".

the correct number of components and has limited flexibility when applied to more complex datasets [37, 44]. Stepwise LDA uses both forward and backward strategies. In the forward approach, the attributes that significantly contribute to the discrimination between the groups will be determined. This process stops when there is no attribute to add in the model. In the backward approach, the attributes (less relevant features) which do not significantly degrade the discrimination between groups will be removed. F-statistic or p value is generally used as predetermined criterion to select/remove the attributes. In this work, the selection of the best features is controlled by four different combinations

of (0.05 and 0.1, SET1; 0.01 and 0.05, SET2; 0.01 and 0.001, SET3; 0.001 and 0.0001, SET4) p values. In a feature entry step, the features that provide the most significant performance improvement will be entered in the feature model if the p value < 0.05. In the feature removal step, attributes which do not significantly affect the performance of the classifiers will be removed if the p value > 0.1. SWLDA was applied to the enhanced feature set to select best features and to remove irrelevant features.

The details of the number of selected enhanced features for every combination were tabulated in Table 2. From Table 2, it can be seen that the enhanced relative energy and entropy features of the speech signals were more significant than the glottal signals. Approximately between 32% and 82% of insignificant enhanced features were removed in all the combination of p values (0.05 and 0.1, 0.01 and 0.05, 0.01 and 0.001, and 0.001 and 0.0001) and speaker-dependent and speaker-independent emotion recognition experiments were carried out using these significant enhanced features. The results were compared with the original raw relative wavelet packet energy and entropy features and some of the significant works in the literature.

3.5. Classifiers

3.5.1. k-Nearest Neighbor Classifier. kNN classifier is a type of instance-based classifiers and predicts the correct class label for the new test vector by relating the unknown test vector to known training vectors according to some distance/similarity function [25]. Euclidean distance function was used and appropriate k-value was found by searching a value between 1 and 20.

3.5.2. Extreme Learning Machine. A new learning algorithm for the single hidden layer feedforward networks (SLFNs) called ELM was proposed by Huang et al. [45–48]. It has been widely used in various applications to overcome the slow training speed and overfitting problems of the conventional neural network learning algorithms [45–48]. The brief idea of ELM is as follows [45–48].

For the given N training samples, the output of a SLFN network with L hidden nodes can be expressed as the following:

$$f_L\left(x_j\right) = \sum_{i}^{L} \beta_i g\left(w_i \cdot x_j + b_i\right), \quad j = 1, 2, 3, \ldots, N. \tag{10}$$

It can be written as $f(x) = h(x)\beta$, where x_j, w_i, and b_i are the input training vector, input weights, and biases to the hidden layer, respectively. β_i are the output weights that link the ith hidden node to the output layer and $g(\cdot)$ is the activation function of the hidden nodes. Training a SLFN is simply finding a least-square solution by using Moore-Penrose generalized inverse:

$$\widehat{\beta} = H\dagger T, \tag{11}$$

where $H\dagger = (H'H)^{-1}H'$ or $H'(HH')^{-1}$, depending on the singularity of $H'H$ or HH'. Assume that $H'H$ is not singular; the coefficient $1/\varepsilon$ (ε is positive regularization coefficient) is added to the diagonal of $H'H$ in the calculation of the output weights β_i. Hence, more stable learning system with better generalization performance can be obtained.

The output function of ELM can be written compactly as

$$f(x) = h(x) H' \left(\frac{1}{\varepsilon} + HH'\right)^{-1} T. \tag{12}$$

In this ELM kernel implementation, the hidden layer feature mappings need not be known to users and Gaussian kernel was used. Best values for positive regularization coefficient (ε) as 1 and Gaussian kernel parameter as 10 were found empirically after several experiments.

4. Experimental Results and Discussions

This section describes the average emotion recognition rates obtained for speaker-dependent and speaker-independent emotion recognition environments using proposed methods. In order to demonstrate the robustness of the proposed methods, 3 different emotional speech databases were used. Amongst, 2 of them were recorded using professional actors/actresses and 1 of them was recorded using university students. The average emotion recognition rates for the original raw and enhanced relative wavelet packet energy and entropy features and for best enhanced features were tabulated in Tables 3, 4, and 5. Table 3 shows the results for the BES database. kNN and ELM kernel classifiers were used for emotion recognition. From the results, ELM kernel always performs better compared to kNN classifier in terms of average emotion recognition rates irrespective of different orders of "db" wavelets. Under speaker-dependent experiment, maximum average emotion recognition rates of 69.99% and 98.98% were obtained with ELM kernel classifier using the raw and enhanced relative wavelet packet energy and entropy features, respectively. Under speaker-independent experiment, maximum average emotion recognition rates of 56.61% and 97.24% were attained with ELM kernel classifier using the raw and enhanced relative wavelet packet energy and entropy features, respectively. kNN classifier gives only maximum average recognition rates of 59.14% and 49.12% under speaker-dependent and speaker-independent experiment, respectively.

The average emotion recognition rates for SAVEE database are tabulated in Table 4. Only audio signals from SAVEE database were used for emotion recognition experiment. According to Table 4, ELM kernel has achieved better average emotion recognition of 58.33% than kNN classifier which gives only 50.31% using all the raw relative wavelet packet energy and entropy features under speaker-dependent experiment. Similarly, maximum emotion recognition rates of 31.46% and 28.75% were obtained under speaker-independent experiment using ELM kernel and kNN classifier, respectively.

After GMM based feature enhancement, average emotion recognition rate was improved to 97.60% and 94.27% using ELM kernel classifier and kNN classifier under speaker-dependent experiment. During speaker-independent experiment, maximum average emotion recognition rates of 77.92%

Table 2: Number of selected enhanced features using SWLDA.

Different order "db" wavelets	P value	BES				SAVEE				SES			
		Features from speech signals		Features from glottal signals		Features from speech signals		Features from glottal signals		Features from speech signals		Features from glottal signals	
		Number of selected RWPEGYs	Number of selected RWPEPYs	Number of selected RWPEGYs	Number of selected RWPEPYs	Number of selected RWPEGYs	Number of selected RWPEPYs	Number of selected RWPEPYs	Number of selected RWPEPYs	Number of selected RWPEGYs	Number of selected RWPEPYs	Number of selected RWPEPYs	Number of selected RWPEPYs
db3	0.05–0.1	21	26	12	12	18	21	10	13	19	25	16	19
	0.01–0.05	17	22	8	11	17	20	7	10	16	24	13	16
	0.001–0.01	16	18	7	8	14	17	6	5	15	22	5	8
	0.0001–0.001	7	13	5	7	13	13	5	5	8	18	5	6
db6	0.05–0.1	18	23	13	14	18	20	17	19	18	23	13	14
	0.01–0.05	13	22	7	12	16	21	17	17	13	22	7	12
	0.001–0.01	12	19	7	10	13	19	11	10	12	19	7	10
	0.0001–0.001	14	13	7	7	13	19	11	10	14	13	7	7
db10	0.05–0.1	22	22	12	12	5	9	4	4	20	23	12	13
	0.01–0.05	17	17	11	10	5	9	4	4	19	19	12	12
	0.001–0.01	14	15	6	7	5	9	4	4	16	18	9	9
	0.0001–0.001	14	15	6	5	5	9	4	3	12	11	6	8
db44	0.05–0.1	25	27	14	15	14	20	7	11	25	27	14	15
	0.01–0.05	20	24	9	9	14	20	7	10	20	24	9	9
	0.001–0.01	15	15	8	7	14	19	6	9	15	15	8	7
	0.0001–0.001	12	13	4	5	13	18	5	3	12	13	4	5

Table 3: Average emotion recognition rates for BES database.

Different order "db" wavelets	Speaker dependent					
	Raw features	Enhanced features	SET1	SET2	SET3	SET4
			ELM kernel			
db3	69.99	98.24	98.52	98.15	97.50	98.15
db6	68.55	95.65	96.39	97.31	96.67	97.13
db10	68.77	98.52	97.78	97.69	96.85	96.39
db44	67.43	98.70	98.43	98.98	98.89	98.61
			*k*NN			
db3	59.14	95.19	96.57	96.11	95.83	96.30
db6	58.68	89.72	87.31	91.30	90.65	91.67
db10	57.93	97.04	96.57	96.57	95.46	95.28
db44	57.63	95.46	95.56	95.56	95.00	96.11
			Speaker independent			
			ELM kernel			
db3	56.61	96.04	97.07	97.24	96.40	96.40
db6	54.70	92.26	91.26	91.83	91.62	91.48
db10	52.08	92.12	94.13	93.37	93.00	92.93
db44	53.60	93.04	93.13	94.10	93.67	94.33
			*k*NN			
db3	49.12	91.75	93.32	92.01	92.79	93.90
db6	48.21	81.93	82.22	81.43	82.18	82.77
db10	45.17	90.64	89.81	89.46	90.23	90.15
db44	46.87	91.69	90.15	90.57	90.35	91.99

Table 4: Average emotion recognition rates for SAVEE database.

Different order "db" wavelets	Speaker dependent					
	Raw features	Enhanced features	SET1	SET2	SET3	SET4
			ELM kernel			
db3	55.63	96.35	96.77	95.21	95.83	96.04
db6	55.63	96.35	95.52	97.60	95.73	96.35
db10	58.23	96.77	91.35	92.60	93.33	91.67
db44	58.33	96.04	95.52	95.63	95.83	96.35
			*k*NN			
db3	47.08	90.63	89.90	90.52	91.56	91.77
db6	47.81	93.54	92.81	94.27	92.29	93.75
db10	46.56	93.23	93.96	93.33	92.60	94.17
db44	50.31	90.21	91.04	91.77	90.10	91.35
			Speaker independent			
			ELM kernel			
db3	27.92	70.00	69.17	68.54	70.21	70.00
db6	31.04	67.92	72.71	73.75	76.25	76.25
db10	31.04	77.92	76.88	76.88	76.88	76.67
db44	31.46	70.83	76.46	76.04	76.25	76.67
			*k*NN			
db3	28.75	63.96	62.50	62.50	63.13	64.38
db6	27.92	63.75	66.04	65.42	66.25	66.25
db10	27.92	69.17	67.50	67.50	67.50	67.71
db44	27.50	64.17	62.50	62.71	61.67	62.50

Table 5: Average emotion recognition rates for SES database.

Different order "db" wavelets	Speaker dependent					
	Raw features	Enhanced features	SET1	SET2	SET3	SET4
		ELM kernel				
db3	41.93	89.92	90.38	89.00	88.46	86.21
db6	42.14	92.79	91.21	92.04	90.63	91.21
db10	40.90	88.83	89.96	90.04	88.58	88.67
db44	42.38	90.00	89.79	88.83	84.67	82.13
		*k*NN				
db3	31.15	77.79	79.50	78.00	78.42	76.88
db6	32.80	81.79	82.17	82.96	81.46	81.63
db10	31.23	78.63	79.38	80.08	79.79	81.08
db44	32.08	78.79	79.38	78.25	74.63	73.71
		Speaker independent				
		ELM kernel				
db3	27.25	78.75	79.17	78.25	78.50	77.50
db6	26.00	83.67	83.75	84.58	83.58	83.42
db10	27.08	79.42	80.42	80.25	80.33	79.92
db44	26.00	80.50	80.92	78.67	76.33	74.92
		*k*NN				
db3	25.92	69.33	69.67	67.92	68.17	66.83
db6	25.50	71.92	73.67	74.00	72.17	73.50
db10	26.33	70.00	71.00	71.08	71.17	70.92
db44	24.67	67.58	69.50	68.08	66.08	67.83

(ELM kernel) and 69.17% (*k*NN) were achieved using the enhanced relative wavelet packet energy and entropy features. Table 5 shows the average emotion recognition rates for SES database. As emotional speech signals were recorded from nonprofessional actors/actresses, the average emotion recognition rates were reduced to 42.14% and 27.25% under speaker-dependent and speaker-independent experiment, respectively. Using our proposed GMM based feature enhancement method, the average emotion recognition rates were increased to 92.79% and 84.58% under speaker-dependent and speaker-independent experiment, respectively. The superior performance of the proposed methods in all the experiments is mainly due to GMM based feature enhancement and ELM kernel classifier.

A paired t-test was performed on the emotion recognition rates obtained using the raw and enhanced relative wavelet packet energy and entropy features, respectively, with the significance level of 0.05. In almost all cases, emotion recognition rates obtained using enhanced features were significantly better than using raw features. The results of the proposed method cannot be compared directly to the literature presented in Table 1 since the division of datasets is inconsistent: the number of emotions used, the number of datasets used, inconsistency in the usage of simulated or naturalistic speech emotion databases, and lack of uniformity in computation and presentation of the results. Most of the researchers have widely used 10-fold cross validation and conventional validation (one training set + one testing set) and some of them have tested their methods under speaker-dependent, speaker-independent, gender-dependent, and gender-independent environments. However, in this work, the proposed algorithms have been tested with 3 different emotional speech corpora and also under speaker-dependent and speaker-independent environments. The proposed algorithms have yielded better emotion recognition rates under both speaker-dependent and speaker-independent environments compared to most of the significant works presented in Table 1.

5. Conclusions

This paper proposes a new feature enhancement method for improving the multiclass emotion recognition based on Gaussian mixture model. Three different emotional speech databases were used to test the robustness of the proposed methods. Both emotional speech signals and its glottal waveforms were used for emotion recognition experiments. They were decomposed using discrete wavelet packet transform and relative wavelet packet energy and entropy features were extracted. A new GMM based feature enhancement method was used to diminish the high within-class variance and to escalate the between-class variance. The significant enhanced features were found using stepwise linear discriminant analysis. The findings show that the GMM based feature enhancement method significantly enhances the discriminatory power of the relative wavelet packet energy and entropy features and therefore the performance of the speech emotion recognition system could be enhanced particularly in the recognition of multiclass emotions. In the future work, more low-level and high-level speech features will be derived and

tested by the proposed methods. Other filter, wrapper, and embedded based feature selection algorithms will be explored and the results will be compared. The proposed methods will be tested under noisy environment and also in multimodal emotion recognition experiments.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research is supported by a research grant under Fundamental Research Grant Scheme (FRGS), Ministry of Higher Education, Malaysia (Grant no. 9003-00297), and Journal Incentive Research Grants, UniMAP (Grant nos. 9007-00071 and 9007-00117). The authors would like to express the deepest appreciation to Professor Mohammad Hossein Sedaaghi from Sahand University of Technology, Tabriz, Iran, for providing Sahand Emotional Speech database (SES) for our analysis.

References

[1] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[2] R. Corive, E. Douglas-Cowie, N. Tsapatsoulis et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[3] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[5] D. Y. Wong, J. D. Markel, and A. H. Gray Jr., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 4, pp. 350–355, 1979.

[6] D. E. Veeneman and S. L. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 369–377, 1985.

[7] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109–118, 1992.

[8] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Computer Speech and Language*, vol. 26, no. 1, pp. 20–34, 2012.

[9] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 34–43, 2007.

[10] K. E. Cummings and M. A. Clements, "Improvements to and applications of analysis of stressed speech using glottal waveforms," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, vol. 2, pp. 25–28, San Francisco, Calif, USA, March 1992.

[11] K. E. Cummings and M. A. Clements, "Application of the analysis of glottal excitation of stressed speech to speaking style modification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '93)*, pp. 207–210, April 1993.

[12] K. E. Cummings and M. A. Clements, "Analysis of the glottal excitation of emotionally styled and stressed speech," *Journal of the Acoustical Society of America*, vol. 98, no. 1, pp. 88–98, 1995.

[13] E. Moore II, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2849–2852, September 2003.

[14] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 1, pp. 96–107, 2008.

[15] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 9, pp. 1530–1540, 2004.

[16] A. I. Iliev and M. S. Scordilis, "Spoken emotion recognition using glottal symmetry," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, Article ID 624575, pp. 1–11, 2011.

[17] L. He, M. Lech, J. Zhang, X. Ren, and L. Deng, "Study of wavelet packet energy entropy for emotion classification in speech and glottal signals," in *5th International Conference on Digital Image Processing (ICDIP '13)*, vol. 8878 of *Proceedings of SPIE*, Beijing, China, April 2013.

[18] P. Giannoulis and G. Potamianos, "A hierarchical approach with feature selection for emotion recognition from speech," in *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1203–1206, European Language Resources Association, Istanbul, Turkey, 2012.

[19] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 1517–1520, Lisbon, Portugal, September 2005.

[20] S. Haq, P. J. B. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP '08)*, pp. 185–190, Tangalooma, Australia, 2008.

[21] M. Sedaaghi, "Documentation of the sahand emotional speech database (SES)," Tech. Rep., Department of Electrical Engineering, Sahand University of Technology, Tabriz, Iran, 2008.

[22] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, vol. 14, PTR Prentice Hall, Englewood Cliffs, NJ, USA, 1993.

[23] M. Hariharan, C. Y. Fook, R. Sindhu, B. Ilias, and S. Yaacob, "A comparative study of wavelet families for classification of wrist motions," *Computers & Electrical Engineering*, vol. 38, no. 6, pp. 1798–1807, 2012.

[24] M. Hariharan, K. Polat, R. Sindhu, and S. Yaacob, "A hybrid expert system approach for telemonitoring of vocal fold pathology," *Applied Soft Computing Journal*, vol. 13, no. 10, pp. 4148–4161, 2013.

[25] M. Hariharan, K. Polat, and S. Yaacob, "A new feature constituting approach to detection of vocal fold pathology," *International Journal of Systems Science*, vol. 45, no. 8, pp. 1622–1634, 2014.

[26] M. Hariharan, S. Yaacob, and S. A. Awang, "Pathological infant cry analysis using wavelet packet transform and probabilistic neural network," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15377–15382, 2011.

[27] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *International Journal of Speech Technology*, vol. 16, no. 2, pp. 143–160, 2013.

[28] Y. Li, G. Zhang, and Y. Huang, "Adaptive wavelet packet filterbank based acoustic feature for speech emotion recognition," in *Proceedings of 2013 Chinese Intelligent Automation Conference: Intelligent Information Processing*, vol. 256 of *Lecture Notes in Electrical Engineering*, pp. 359–366, Springer, Berlin, Germany, 2013.

[29] M. Hariharan, K. Polat, and R. Sindhu, "A new hybrid intelligent system for accurate detection of Parkinson's disease," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 3, pp. 904–913, 2014.

[30] S. R. Krothapalli and S. G. Koolagudi, "Characterization and recognition of emotions from speech using excitation source information," *International Journal of Speech Technology*, vol. 16, no. 2, pp. 181–201, 2013.

[31] R. Farnoosh and B. Zarpak, "Image segmentation using Gaussian mixture model," *IUST International Journal of Engineering Science*, vol. 19, pp. 29–32, 2008.

[32] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, pp. 28–31, August 2004.

[33] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive GMMRF model," in *Computer Vision—ECCV 2004*, vol. 3021 of *Lecture Notes in Computer Science*, pp. 428–441, Springer, Berlin, Germany, 2004.

[34] V. Majidnezhad and I. Kheidorov, "A novel GMM-based feature reduction for vocal fold pathology diagnosis," *Research Journal of Applied Sciences*, vol. 5, no. 6, pp. 2245–2254, 2013.

[35] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing: A Review Journal*, vol. 10, no. 1, pp. 19–41, 2000.

[36] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falcão, "Spoken emotion recognition through optimum-path forest classification using glottal features," *Computer Speech and Language*, vol. 24, no. 3, pp. 445–460, 2010.

[37] M. H. Siddiqi, R. Ali, M. S. Rana, E.-K. Hong, E. S. Kim, and S. Lee, "Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis," *Sensors*, vol. 14, no. 4, pp. 6370–6392, 2014.

[38] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Information Processing & Management*, vol. 45, no. 3, pp. 315–328, 2009.

[39] J. Jiang, Z. Wu, M. Xu, J. Jia, and L. Cai, "Comparing feature dimension reduction algorithms for GMM-SVM based speech emotion recognition," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA '13)*, pp. 1–4, Kaohsiung, Taiwan, October-November 2013.

[40] P. Fewzee and F. Karray, "Dimensionality reduction for emotional speech recognition," in *Proceedings of the International Conference on Privacy, Security, Risk and Trust (PASSAT '12) and International Confernece on Social Computing (SocialCom '12)*, pp. 532–537, Amsterdam, The Netherlands, September 2012.

[41] S. Zhang and X. Zhao, "Dimensionality reduction-based spoken emotion recognition," *Multimedia Tools and Applications*, vol. 63, no. 3, pp. 615–646, 2013.

[42] B.-C. Chiou and C.-P. Chen, "Feature space dimension reduction in speech emotion recognition using support vector machine," in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA '13)*, pp. 1–6, Kaohsiung, Taiwan, November 2013.

[43] S. Zhang, B. Lei, A. Chen, C. Chen, and Y. Chen, "Spoken emotion recognition using local fisher discriminant analysis," in *Proceedings of the IEEE 10th International Conference on Signal Processing (ICSP '10)*, pp. 538–540, October 2010.

[44] D. J. Krusienski, E. W. Sellers, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "Toward enhanced P300 speller performance," *Journal of Neuroscience Methods*, vol. 167, no. 1, pp. 15–21, 2008.

[45] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[46] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[47] W. Huang, N. Li, Z. Lin et al., "Liver tumor detection and segmentation using kernel-based extreme learning machine," in *Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '13)*, pp. 3662–3665, Osaka, Japan, July 2013.

[48] S. Ding, Y. Zhang, X. Xu, and L. Bao, "A novel extreme learning machine based on hybrid kernel function," *Journal of Computers*, vol. 8, no. 8, pp. 2110–2117, 2013.

[49] A. Shahzadi, A. Ahmadyfard, A. Harimi, and K. Yaghmaie, "Speech emotion recognition using non-linear dynamics features," *Turkish Journal of Electrical Engineering & Computer Sciences*, 2013.

[50] P. Henríquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, and J. R. Orozco-Arroyave, "Application of nonlinear dynamics characterization to emotional speech," in *Advances in Nonlinear Speech Processing*, vol. 7015 of *Lecture Notes in Computer Science*, pp. 127–136, Springer, Berlin, Germany, 2011.

[51] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[52] S. R. Krothapalli and S. G. Koolagudi, "Emotion recognition using vocal tract information," in *Emotion Recognition Using Speech Features*, pp. 67–78, Springer, Berlin, Germany, 2013.

[53] M. Kotti and F. Paternò, "Speaker-independent emotion recognition exploiting a psychologically- inspired binary cascade classification schema," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 131–150, 2012.

[54] A. S. Lampropoulos and G. A. Tsihrintzis, "Evaluation of MPEG-7 descriptors for speech emotional recognition," in *Proceedings of the 8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP '12)*, pp. 98–101, 2012.

[55] N. Banda and P. Robinson, "Noise analysis in audio-visual emotion recognition," in *Proceedings of the International Conference on Multimodal Interaction*, pp. 1–4, Alicante, Spain, November 2011.

[56] N. S. Fulmare, P. Chakrabarti, and D. Yadav, "Understanding and estimation of emotional expression using acoustic analysis of natural speech," *International Journal on Natural Language Computing*, vol. 2, no. 4, pp. 37–46, 2013.

[57] S. Haq and P. Jackson, "Speaker-dependent audio-visual emotion recognition," in *Proceedings of the International Conference on Audio-Visual Speech Processing*, pp. 53–58, 2009.