IMPROVED SPEAKER-INDEPENDENT EMOTION RECOGNITION FROM SPEECH USING TWO-STAGE FEATURE REDUCTION

[1]Hasrul Mohd Nazid, [2]Hariharan Muthusamy, [3]Vikneswaran Vijean, [4]Sazali Yaacob

[1,2, & 3] School of Mechatronic Engineering,
Universiti Malaysia Perlis, Malaysia
[4] Universiti Kuala Lumpur Malaysian Spanish Institute,
Kulim Hi-Tech Park, Malaysia

kids.hasrul@gmail.com;wavelet.hari@gmail.com;viky.86max@gmail.com;sazali22@yahoo.com

ABSTRACT

In the recent years, researchers are focusing to improve the accuracy of speech emotion recognition. Generally, high emotion recognition accuracies were obtained for two-class emotion recognition, but multi-class emotion recognition is still a challenging task . The main aim of this work is to propose a two-stage feature reduction using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for improving the accuracy of the speech emotion recognition (ER) system. Short-term speech features were extracted from the emotional speech signals. Experiments were carried out using four different supervised classifiers with two different emotional speech databases. From the experimental results, it can be inferred that the proposed method provides better accuracies of 87.48% for speaker dependent (SD) and gender dependent (GD) ER experiment, 85.15% for speaker independent (SI) ER experiment, and 87.09% for gender independent (GI) experiment.

Key words: Emotional speech, cepstral features, feature reduction, emotion recognition.

INTRODUCTION

To recognize human emotion, various modalities are used such as facial images and videos, speech and physiological signals. In the recent years, researchers have published several works on emotion recognition from spoken utterances (El Ayadi, Kamel, & Karray, 2011; Koolagudi & Rao, 2012). Spoken utterances of an individual can provide information about his/her health state, emotion, language used and gender. Speech is the one of the most natural form of communication between individuals. Understanding of an individual's emotion can be useful for applications like web movies, electronic tutoring applications, in-car board system, diagnostic tool for therapists and call-centre applications (El Ayadi et al., 2011; Koolagudi & Rao, 2012). Researchers have proposed several parameterization methods in the field of emotion recognition from speech, however it is not clear that which speech features are best in distinguishing between emotions. Researchers have used four primary emotions such as happiness, sadness, anger, fear, surprise and disgust. The recognition accuracy between high-activation emotions and low-activation emotions are always high, but recognition between different emotions is still challenging (Wang & Guan, 2004).

Emotional speech database contains three speech categories: simulated, elicited and natural. Simulated emotions tend to be more expressive than real ones and are most commonly used (El Ayadi et al., 2011). For 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. In natural category, all emotions may not be availabe and is difficult to model because they are completely naturally expressed. Table 1 presents the summary of the significant research works on ER from speech signals.

Table 1

Summary of previous research works on ER from speech

No.	Reference	Features	Emotion	Classifier	Highest Accuracy
1	(Dey, Rajan, Padmanabhan, & Murthy, 2011)	MFCC, LPCC, Modified group delay features (MODGDF), Mel-slope features	Anger, Anxiety, boredom, disgust, happy, sad, neutral	128-mixture Gaussian mixture models	MFCC -57.61% MODGDF-49.52% LPCC-54.76% MFCC+MODGDF-57.67% MFCC+LPCC-57.41%
2	(Pan, Shen, & Shen, 2005)	Energy, pitch, LPCC, MFCC, Mel-energy spectrum dynamic coefficients (MEDC)	Happy, sad, neutral	SVM	MFCC+MEDC+ Energy -95.1%
3	(Pan, Shen, & Shen, 2012)	Energy, pitch, MFCC, its 1st order difference, 2nd order difference, MEDC, its 1st order difference, 2nd order difference	Happy, sad, neutral	SVM	MFCC+MEDC+ Energy - 95.1%
4	(J. Huang, Yang, & Zhou, 2012)	MFCC	Happiness, Anger, Neutral, Sadness	-Variance-based Gaussian kernel fuzzy vector quantization, -Fuzzy weighted vector quantization error, -Fuzzy C-means Clustering Vector Quantization-SVM	-

(continued)

No.	Reference	Features	Emotion	Classifier	Highest Accuracy
5	Shen et al. (2011)	Energy, pitch, LPCC, MFCC, Linear Prediction coefficients and Mel ceptrum coefficients (LPCMCC)	Disgust, Boredom, Sadness, Neutral, Happiness	SVM	Energy & Pitch-66.02% LPCMCC – 70.7% Both – 82.5%
6	(Bozkurt, Erzin, Erdem, & Erdem, 2010)	Line Spectral Frequency, MFCC	Anger, Anxiety, boredom, disgust, happy, sad, neutral	Gaussian Mixture Model (GMM)	84.58%
7	(Iliou & Anagnostopoulos, 2010)	Pitch, MFCC, Energy, Formants	Anger, Anxiety, boredom, disgust, happy, sad, neutral	-Multilayer Percepton (MLP), -Probabilistic Neutral Networks (PNP), -SVM	SVM – 78% MLP – 53%
8.	(Kotti & Paternò, 2012)	Pitch, formants, energy contours, spectrum, cepstrum, perceptual and many others	Anger, Anxiety, boredom, disgust, happy, sad, neutral	k-nearest neighbor classifier and SVM	SVM-87.7% for speaker independent
9.	(Sezgin, Gunsel, & Kurt, 2012)	Perceptual audio features	Anger, Anxiety, boredom, disgust, happy, sad, neutral	SVM and GMM	Binary arousal and valence discrimination was performed.
10.	(SHAHZADI, AHMADYFARD, HARIMI, & YAGHMAIE, 2013)	Conventional+nonlinear features	Anger, Anxiety, boredom, disgust, happy, sad, neutral	kNN+GA based feature selection	82-86%

Researchers have used different speech signal processing algorithms and classification algorithms and the recognition accuracies varies between 49.52% and 95.10%. They have used Berlin emotional speech database (EmoDB) and also their own emotional speech database. High emotion recognition accuracies were obtained for two-class emotion recognition (High arousal Vs Low arousal) but multi-class 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, (d) long-term and short-term emotional states (El Ayadi et al., 2011; Koolagudi & Rao, 2012; Ververidis & Kotropoulos, 2006). Although all the above works are novel contributions to the field of speech emotion recognition, it is difficult to compare them directly since the division of datasets are inconsistent, difference in number of emotions used, difference in number of datasets used, inconsistency in the usage of simulated or naturalistic speech emotion databases and lack of uniformity in the computation and presentation of the results. Hence, in this paper, the proposed algorithms were tested under different experimental conditions such as gender/speaker dependent, speaker independent, and gender independent using two different emotional speech databases. Short-term speech features such as Linear Predictive Coding (LPC), Linear Prediction Ceptral Coefficients (LPCCs), Weighted Linear Prediction Cepstral Coefficients (WLPCCs) and Mel-frequency Cepstral Coefficients (MFCCs) were extracted from the emotional speech signals. Two-stage feature reduction was proposed using PCA and LDA to improve the ER accuracy. Four different supervised classifiers such as *k*-Nearest Neighbor (KNN), Fuzzy *k*-Nearest Neighbor (FKNN), Multiclass Support Vector Machine (MCSVM) and Extreme Learning Machine (ELM) were employed for testing the effectiveness of the original and dimensionality reduced (DimRed) short term speech features for the recognition of emotions.

MATERIALS AND METHODS

Database

In this paper, two emotional speech databases were used and they were diffierent in terms of sampling frequency, number of emotions, number of subjects, number of words and utterances. Berlin emotional speech database (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005) was recorded at the Technical University, Berlin. Seven emotions such as neutral (N), anger (A), fear (F), happiness (H), sadness (Sa), disgust (D) and boredom (BD) were simulated by 10 actors (5 male and 5 Female). In EmoDB, 535 utterances were recorded for each emotion with 10 german sentences. Sahand emotional speech database (SESD) (Sedaaghi, 2008) comprises of utterances expressed by five male and five female students in five emotional states (A, N, H, Sa, and Surprise-Su). SESD consists of 1200 utterances which include twenty four words, short sentences and paragraphs spoken in Farsi language. The sampling frequency was 16 kHz for EmoDB and 8 kHz for SESD. In our analysis, we set 8 kHz as the sampling frequency and hence the speech samples of EmoDB were down-sampled to 8 kHz. Number of utterances for each category of emotion available in EmoDB and SESD is shown in Table 2.

Table 2

Number of Utterances

	Anger	**Happiness**	**Neutral**	**Sadness**	**Boredom/ Surprise**	**Disgust**	**Fear**
SES	120	120	120	120	120	NA	NA
EmoDB	127	71	79	62	81	46	69

The proposed system for ER from speech signals is shown in Fig 1.

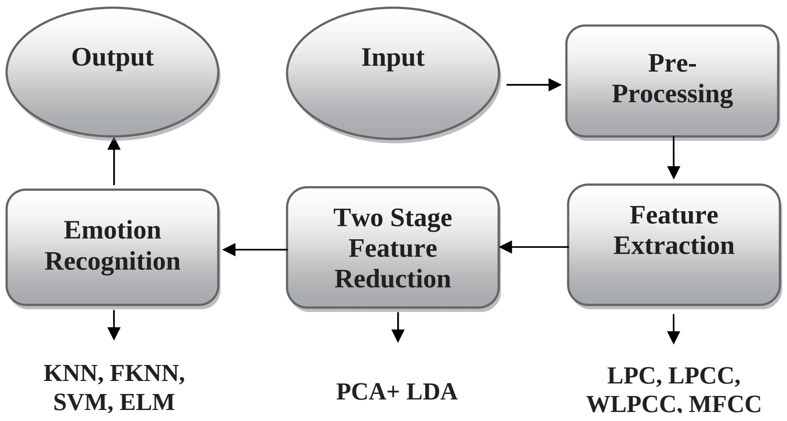

Figure 1. Block diagram of the proposed speech emotion recognition system

Feature Extraction

In this paper, short term cepstral features were extracted from the emotional speech signals of two different emotional speech databases. The following sections describe the derivation of short-term cepstral features.

Extraction of LPC based cepstral parameters

Linear prediction is used to extract the relavent information that represents the signal. All the speech samples were pre-emphasized using a first-order digital filter as shown in equation 1. In Equation 1, the value of ã = 15/16 = 0.9375 was used with the sampling rate of 8 kHz for the fixed-point implementations (Rabiner & Juang, 1993).

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

The overview of Linear Prediction process is shown in Fig 2. The pre-emphasized speech signals were segmented at the length of 20 ms with 50% overlap. Speech signal at time t, $\hat{r}(t)$, can be estimated as a linear combination of the past order of the LPC for speech samples.

$$\hat{r}(t) = \sum_{a=1}^{p} s_a r(t-a) \qquad (2)$$

where *p* represent the order of the LPC.

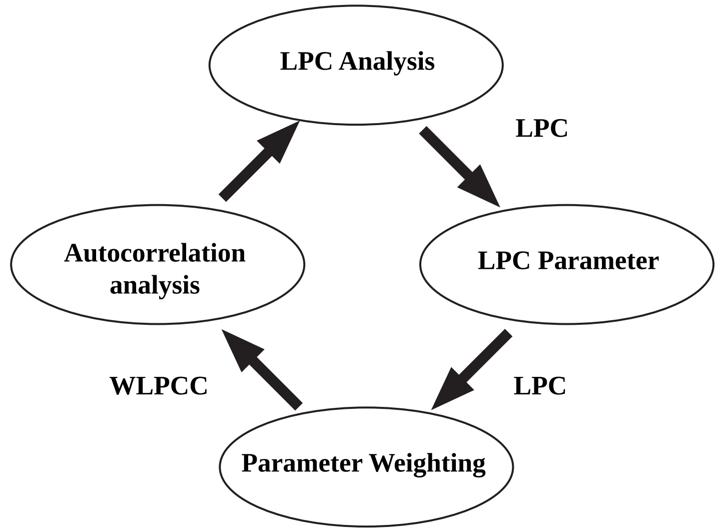

Figure 2. Linear prediction process

The differences between the actual and the estimated sample value is known as the prediction error, e(t) and defined as

$$e(t) = r(t) - \hat{r}(t) = r(t) - \sum_{a=1}^{p} s_a r(t-a) \tag{3}$$

where s_a is the LPC. By minimizing the mean squared error over a frame of the speech signal, the LPC's are calculated. Therefore, autocorrelation method is employed to each frame of the windowed signal as shown in Eqs. 4 and 5.

$$s(a) = \sum_{t=0}^{T-1-a} z(t)z(t+a), a = 0,1,..., p \tag{4}$$

where the autocorrelation function is symmetric, so that the LPC equations can be defined as

$$\sum_{a=1}^{p} s(|a-q|)s_a = s(a), 1 \le a \le p \tag{5}$$

In this paper, we set the order of *p* as 14. LPCC's are LPC's represented in the cepstrum domain and the coefficient of the Fourier transform representation of the log magnitude spectrum. The number of LPCC used to represent each frame is calculated by applying *Q* ≈ (3/2)*p* when *Q* > *p*. By deriving directly from LPC using recursion technique, LPCC's were obtained.

$$c_o = \ln \sigma^2 \tag{6}$$

$$c_a = s_a + \sum_{q=1}^{a=1} \left(\frac{q}{a} \right) c_k s_{a-k}, 1 \le a \le p \tag{7}$$

$$c_a = \sum_{q=1}^{a=1} \left(\frac{q}{a} \right) c_k s_{a-k}, m \triangleright p \tag{8}$$

σ^2 gain term in the LPC model

c_a LPCC

s_a LPC

P order p

WLPCC generated by multiplying LPCC with the wighted formula (9). Weighted function as bandpass filter in cepstral domain to de-emphasizes c_a around a = 1 and a = Q.

$$w_a = \left[1 + \frac{Q}{2} \sin \frac{\pi a}{Q} \right], 1 \le a \le Q \tag{9}$$

WLPCC, $\overset{\wedge}{c_a}$ are determined by using the formula

$$\overset{\wedge}{c_a} = w_a * c_a, 1 \le a \le Q \tag{10}$$

Mel frequency cepstral coefficients

Mel frequency cepstral coefficients are the most commonly used features for speech/speaker recognition. MFCCs take human perception sensitivity with respect to frequencies into consideration and are best for speech recognition (Jang, 2011). The block diagram of the MFCC extraction is shown in Fig 3.

The speech signals were pre-emphasized with a first-order digital filter and segmented into short overlapping frames as in LPC feature extraction (Chee, Ai, Hariharan, & Yaacob, 2009).The frame size for the study was fixed at 160 samples and 50% of data overlapping was used. Each frame is multiplied by Hamming window to minimize the spectral distortion and the signal discontinuities (Chee, et al., 2009) .

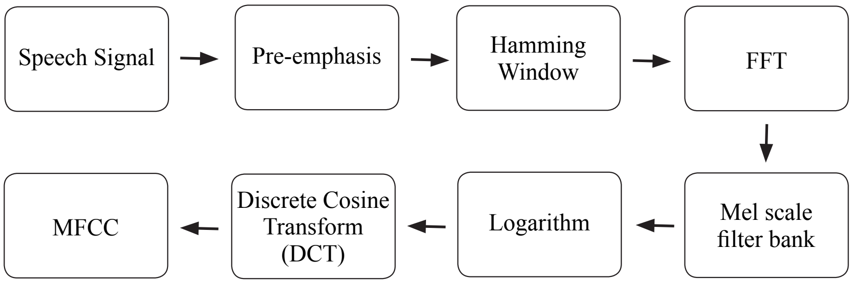

Figure 3. Block diagram of MFCC extraction

Fast Fourier Transform (FFT) was applied to convert time domain into the frequency domain. The spectrum of each frame was filtered by a set of filter after the FFT block and then, the power of each band was calculated (Chee, et al., 2009). To simulate the subjective spectrum, a filter band spaced uniformly on the Mel-scale was used. Mel-scale is defined as a logarithmic scale of frequency based on human pitch perception. Equation (11) shows the mapping from linear frequency to Mel-frequency (Chee, et al., 2009).

$$Mel(f) = 2595\log_{10}\left(1+\frac{f}{100}\right) \tag{11}$$

Lastly, the log Mel spectrum was converted to time using Discrete Cosine Transform (DCT) and the output is called as Mel Frequency Cepstrum Coefficients.

The emotional speech signals were subjected to feature extraction and feature database was formed by using 40-MFCCs, 14-LPCs, 21-LPCCs and 14-WLPCCs and totally there were 89 short-term cepstral features. Next section discusses about dimension reduction of the 89 features into fewer dimensions using the two-stage feature reduction with PCA and LDA.

Two stage feature reduction

Feature selection/reduction is a important step in all pattern recognition problems since the large feature space (curse of dimensionality) may reduce the classification performance. Dimensionality of feature set can be reduced by using statistical methods to minimize the *k*, relevant information (Haq & Jackson, 2009; Haq, Jackson, & Edge, 2008). PCA (Shlens, 2005) was used to extract the essential charateristics from high dimensional data set and to discard noise. The advantages of PCA is when we know the patterns in the data, we can compress the data by reducing the number of dimensions, without loss of information. First step in PCA analysis is the subtraction of the mean from each of the data dimensions. Next step is the estimation of covariance matrix and then determine the Eigenvectors and Eigenvalues of the covariance matrix. A linear transformation mapping was done to map the original *h*-dimensional feature space into an *l*-dimensional feature subspace (*l*<*h*). The *h*-dimensional short term cepstral feature vector is represented by considering a set of *N* sample features $\{a_1, a_2, .., a_N\}$. Next, the new vector $b_i \in R^l$ is defined by

$$b_i = M^T{}_{pca} a_i (i = 1,2,...,N) \tag{12}$$

where M_{pca} is the linear transformations matrix, and *i* is the number of features. The columns of M_{pca} are the *l* eigenvectors associated with the *l* largest eigenvalues of scatter matrix U_T, defined as

$$U_T = \sum_{i=1}^{N} (a_i - \mu)(a_i - \mu)^T \tag{13}$$

where $\mu \in R^h$ is the mean features of all samples (Deng, Jin, Zhen, & Huang, 2005).

Next, LDA is commonly used technique for dimensionality reduction (Duda, Hart, & Stork, 2012; Yusuf, Mahat, Siraj, & Yaacob, 2012). LDA maximizes the ratio of between-class variance to whithin-class variance to optimize seperability between classes. The within class scatter matrix U_m and between class scatter matrix U_e are defined as

$$U_m = \sum_{j=1}^{p} \sum_{i=1}^{N_j} (a_i^j - \mu)(a_i^j - \mu)^T \tag{14}$$

$$U_e = \sum_{j=1}^{p} (\mu_j - \mu)(\mu_j - \mu)^T \tag{15}$$

where a_i^j is the i^{th} sample of class *j*, μ_j is the mean of class *j*, μ is the mean features of all classes, *p* is the number of classes, and N_j is the number of samples of class *j*. To select M_{lda} is to maximize the ratio $\det|U_e|/\det|U_m|$.

PCA was applied on the feature database and the number of principal components were selected according to containment of 99% of the total variability and the number of features was reduced from the original 89 features. Next, LDA was applied on the reduced feature database obtained from PCA, to reduce the dimensionality of features further. PCA combined with LDA was applied as a feature reduction method in order to seek a projection that best represent the original data and best seperates the data in a least-squares sense. The PCA maps the original *h*-dimensional feature a_i to *l*-dimensional feature b_i as an intermediate space. Then, LDA projects the PCA output into a new *g*-dimensional feature vector c_i (Deng, et al., 2005).

$$c_i = M^T{}_{lda} M^T{}_{pca} a_i (i = 1,2,..., N) \tag{16}$$

In this study, recognition of three, five and seven classes of emotions were performed, so as to reduce the dimension of features into 2, 4 and 6 (i.e., number of class-1) respectively.

Classifiers

ER from speech signals is a typical pattern recognition application. The original and dimensionality reduced features were used to recognize the emotions. In this study, seven emotions of EmoDB database and five emotions of SESD database were considered. Recognition of three (3E) (N, H, Sa– (EmoDB and SESD)), five (5E) (N, H, Sa, A, BD – EmoDB and N, A, H, Sa, Su – SESD) and seven (7E) (N, A, F, H, Sa, D, BD – EmoDB) classes of emotions were done. The effect of gender and speaker dependency on recognition of emotions was also investigated using four different classifiers. The classification process was repeated for 10 times and the average emotion recognition accuracy was reported in all the experiments. The following sections give the basics of classifiers used.

K-nearest neighbor

k-NN is the elementary classification model that apply lazy learning. The *k*-NN prediciton of the query instance is determined by the majority voting of the nearest neighbour category. To locate the *k*-NN category of the training data set, the minimum distance from the test speech signal to the each of the training speech signal in the training test was calculated (Chia Ai, Hariharan, Yaacob, & Sin Chee, 2012; Hariharan, Chee, Ai, & Yaacob, 2012;Yusuf, Mahat, Siraj, & Yaacob, 2012). Class label of the test speech signal was determined by using majority voting between the *k* nearest training speech samples from the *k*-NN category. Hence, the *k* values show an important role in *k*-NN classification (Chia Ai, et al., 2012; Hariharan, et al., 2012; Liu, Lee, & Lin, 2010). Therefore, in this study, the best *k* value was found between 1 and 10.

Fuzzy K-nearest neighbor

FKNN is a classification technique that provides the simplicity and the practicability of classical *k*-NN using fuzzy logic concept. The FKNN algorithm assigns class membership to a sample vector rather than assigning the vector to a particular class. The basic of the algorithm is to assign membership as a function of the patterns distance from its *k*-NN and those neighbors membership in the possible classes. It is similar to the traditional set theory in the sense that it must also search the labelled sample set for the *k*-NN. The FKNN keeps the main idea of *k*-NN, in which the class decision is made by the nearest neighbor class information. The advantages of using fuzzy set theory is that no arbitrary assignments are made (Keller, Gray, & Givens, 1985; Kim & Han, 1995) which are the residues that are assigned with a membership value in each class rather than binary decision of 'belongs to' or 'does not belong to'. The advantage of such assignment is that these membership values act as strength or confidence with which the current residue belongs to a particular class (Bondugula, Duzlevski, & Xu, 2005).

Multiclass support vector machine

SVM is the one of most popular supervised classifiers for binary classification problems and it is insensitive to high dimensionality of the feature space. However, SVM can also be used for multi-class classification problems using multiple binary SVM classifiers with either one-against-all or one-against-one approach. In binary classification, the class labels can take only two values (1 and -1). The idea of multiclass is to use the one-against-all approach where it constructs *E* two-class rules, where the *m*th function $w_m^T \phi(x) + b$ separates training vectors. Hence there are *E* decision functions but all are obtained by solving one problem. The formulation is as follows:

$$\min \frac{1}{2} \sum_{m=1}^{k} w_m^T w_m + C \sum_{i=1}^{l} \sum_{m \neq y_i} \xi_i^m \qquad (17)$$

$$w_{y_i}^T \phi(x_i) + b_{y_i} \geq w_m^T \phi(x_i) + b_m + 2 - \xi_i^m \qquad (18)$$

$$\xi_i^m \geq 0, i = 1, \ldots, l, m \in \{1, \ldots, k\} \qquad (19)$$

x is in class which has the largest value of the decision function $w_m^T \phi(x) + b$.

In this study, we used one-against-all MCSVM from Kernel Methods MATLAB Toolbox(Canu, Grandvalet, Guigue, & Rakotomamonjy, 2005) to classify the features. In our study, we fixed the value of hyper parameter C (C = 1000).

Extreme learning machine

The ELM has two main advantages such as it requires less training time compared to conventional neural network based classifiers and need to tune the parameter *L* (hidden layer nodes) to get better accuracy. ELM has higher generalization capability and suitable for many nonlinear activation function and kernel functions. ELM is developed for generalized single hidden layer feedback networks (SLFNs) with a wide variety of hidden nodes. ELM randomly selects all the hidden note parameters, after which the network can be represented as a linear system and the output of weights can be computed analytically (G.-B. Huang, Zhou, Ding, & Zhang, 2012; G.-B. Huang, Zhu, & Siew, 2006; J. Huang, et al., 2012). In ELM, the input data is mapped from the input space to *L*-dimensional hidden layer feature space. The output of ELM is

$$f_L(x) = \sum_{i=1}^{L} \beta_i h_i(x) = h(x)\beta \qquad (20)$$

where $\beta = [\beta_1, ..., \beta_L]^T$ is the output weight vector from hidden nodes to the output node. $h(x) = [h_1(x), ..., h_L(x)]$ is the row vector presenting the output of the *L* hidden nodes with respect to the input *x*. In other words, *h(x)* maps the data from the *d*-dimensional input space to the *L*-dimensional hidden layer feature space *H*. In our study, we used the ELM code developed by (G.-B. Huang, et al., 2012; G.-B. Huang, et al., 2006; J. Huang, et al., 2012). The number of hidden neurons was set to 20 after several experiments.

Experimental Results

One-way analysis of variance (ANOVA) was performed by using statistical package for the social science (SPSS) to validate the discerning abilities of the features between the groups. Table 3 shows the ANOVA results and F-ratio of dimensionality reduced (DimRed) features which were greater than the original features.

Table 3

ANOVA results

Classes	Original features				DimRed features			
	EmoDB		SESD		EmoDB		SESD	
	F	Sig	F	Sig	F	Sig	F	Sig
3E	173.599	0.000	8.331	0.000	670.680	0.000	163.108	0.000
5E	94.407	0.000	6.092	0.000	195.294	0.000	67.602	0.000
7E	78.011	0.000	NA	NA	118.197	0.000	NA	NA

From Table 3, it was also noticed that all the *p*-value was used for testing the hypothesis and it was equal to 0.000. Since the *p*-value of 0.000 is less than significance level of 0.05, we reject H_0. The statistical analysis provides sufficient evidence to conclude that mean weights of features from 3E, 5E and 7E were different for both the databases. Different experiments of ER were performed such as speaker dependent (SD), speaker independent (SI), gender dependent (GD), gender independent (GI), recognition of 3E, 5E and 7E. The training and testing sets were prepared as shown in Table 4 for SI ER experiment.

Table 4

Details of training and testing sets used in SI ER

Exp.	EmoDB (Subject)		SESD (Subject)	
	training	testing	training	testing
1	03,08	10,11,12,15,09,13,14,16	Abdl_N, Bari_A	Fthi_F, Jeda_A, Mona_M, Tavk_B, Dara_E, Emam_M, Khri_R, Seif_M
2	10,09	03,11,12,15,08,13,14,16	Fthi_F, Dara_E	Abdl_N, Jeda_A, Mona_M, Tavk_B, Bari_A, Emam_M, Khri_R, Seif_M
3	11,13	03,10,12,15,08,09,14,16	Jeda_A, Emam_M	Abdl_N, Fthi_F, Mona_M, Tavk_B, Bari_A, Dara_E, Khri_R, Seif_M
4	12,14	03,10,11,15,08,09,13,16	Mona_M, Khri_R	Abdl_N, Fthi_F, Jeda_A, Tavk_B, Bari_A, Dara_E, Emam_M, Seif_M
5	15,16	03,10,11,12,08,09,13,14	Tavk_B, Seif_M	Abdl_N, Fthi_F, Jeda_A, Mona_M, Bari_A, Dara_E, Emam_M, Khri_R

Gender and Speaker Dependent

Training and testing sets were prepared using the original features and DimRed features extracted from all the utterances (535 utterances in EmoDB database and 1200 utterances in SESD). Conventional validation method (80% training + 20% testing) was used in GD and SD ER experiment. Out of the total utterances, 80% of the utterances were used as training set and the remaining 20% of utterances were used as testing set. Experiments were repeated for 10 times and the average of 10 repetitions was reported as the ER accuracy. Table 5 presents the experimental results of GD and SD ER by using the four different classifiers for original and DimRed features. From the Table 5, it can be seen that the MCSVM performed well in recognizing 3E, 5E and 7E using DimRed features for EmoDB and SESD compared to other classifiers (KNN, FKNN and ELM).

Table 5

ER Results for GD and SD

Classes	Features	Accuracy							
		EmoDB				SESD			
		FKNN	KNN	MCSVM	ELM	FKNN	KNN	MCSVM	ELM
3E	Original	92.62	91.43	92.92	69.78	59.24	59.10	64.31	53.65
	DimRed	100	100	100	97.08	60.49	64.38	68.75	66.02
5E	Original	75.52	73.28	77.58	58.54	40.50	41.38	47.50	37.38
	DimRed	93.13	92.54	91.74	86.99	45.04	46.17	46.58	47.50
7E	Original	72.83	72.74	75.14	62.19	–	–	–	–
	DimRed	86.79	85.75	87.48	83.49	–	–	–	–

Speaker Independent

SI ER in both database are evaluated in five separate experiments. In each experiment, training set was formed using the original and DimRed features extracted from two speakers. The duo was selected in order to get one male and one female speaker at a time (Table 4).

From Table 6, it can be observed that we obtained highest recognition accuracies of 100% and 65.28% by using DimRed features with MCSVM classifier in classifying 3E for EmoDB and SESD respectively. The DimRed features will convey more discerning information about the different emotional speech which results in highest ER accuracy compared to original features in all the experiments. While the ER accuracies for 5E were 93.10% and 48.39 % using *k*-NN and ELM classifier respectively. In recognition of 7E, both MCSVM and *k*-NN classifiers provided highest accuracy of 85.15% for EmoDB.

Gender dependent

Table 7 shows the ER results for GI experiment for original and DimRed features. Here, training and testing sets were prepared using features extracted from the male speakers and female speakers respectively. From Table 7, inferences unfold that MCSVM and FKNN were performed equally better in providing highest accuracy for both the databases. MCSVM and FKNN provided highest accuracy of 87.09% and 96.04% for 7E and 5E using DimRed features for EmoDB. MCSVM and ELM provided maximum accuracy of 71.11% and 52.05% for 3E and 5E using DimRed features for SESD.

Table 6

ER Results for SI

Classes	Features	Accuracy							
		EmoDB				SESD			
		FKNN	KNN	MCSVM	ELM	FKNN	KNN	MCSVM	ELM
3E	Original	78.05	78.66	80.49	66.10	35.24	35.59	32.81	33.17
	DimRed	99.39	99.39	100	96.56	59.20	63.02	65.28	59.83
5E	Original	40.34	40.00	36.55	32.83	24.58	24.27	22.40	22.20
	DimRed	92.07	93.10	92.41	85.37	44.58	45.31	43.85	48.39
7E	Original	47.80	47.80	41.53	40.39	–	–	–	–
	DimRed	84.45	85.15	85.15	83.71	–	–	–	–

Table 7

ER Results for GI

Classes	Features	Accuracy							
		EmoDB				SESD			
		FKNN	KNN	MCSVM	ELM	FKNN	KNN	MCSVM	ELM
3E	Original	68.60	68.60	77.69	62.98	52.50	51.94	43.33	45.11
	DimRed	100	100	100	98.93	62.78	66.67	71.11	67.75
5E	Original	38.12	37.62	38.12	37.45	28.67	28.33	23.50	24.89
	DimRed	96.04	95.05	94.06	94.53	46.67	48.33	49.83	52.05
7E	Original	32.78	33.77	36.75	33.32	–	–	–	–
	DimRed	86.42	86.75	87.09	86.69	–	–	–	–

DISCUSSION

In the area of ER from speech, several feature extraction and classification techniques were proposed. Most of the current studies focus on developing new feature extraction and classification algorithms. ER accuracy depends on the relevant features, quality of the database and experimental setups, and classification techniques. The robustness of the proposed algorithms should be tested with more than one emotional speech database. In this paper, the proposed algorithms were tested using two different emotional speech databases and also we have conducted different experiments like SD, GD, SI and GI. The combination of PCA and LDA reduced the high dimension features into fewer dimensions and increased the discrimination ability of the features and hence, we obtained very promising ER accuracy in all the experiments. In (Pan, et al., 2012), energy, pitch, MFCCs and LPCCs were used as features and SVM as a classifier to classify 3E (N,H,Sa) from EmoDB. The highest ER accuracy was 95.1% only. However, in our work, we have achieved 100% by applying two-stage feature reduction (SI). In (Shen, Changjun, & Chen, 2011), LPCs and MFCCs were used as features and SVM as a classifier to recognize 5E from EmoDB and the ER accuracy was only 70.70%, but in our analysis, the highest ER accuracy was 96.04% using DimRed features (GI). In recognition of 7E, our analysis showed the highest accuracy of 87.48% with DimRed features (SI). Bozkurt et al. have obtained 84.58% accuracy using Line Spectral Frequency and MFCCs as their features and GMM as classifier (Bozkurt, et al., 2010). In (Giannoulis & Potamianos, 2012), prosodic features, spectral features, glottal flow features, AM-FM features were utilized and two-stage feature reduction was proposed for speech emotion recognition. The overall emotion recognition rates of 85.18% for gender dependent and 80.09% for gender independent was achieved using SVM classifier. In this work, we obtained 87.48% for gender dependent and 87.09% for gender independent. Ali Shahzadi et.al have proposed non-linear dynamics features (NLDs) for speech emotion recognition (Shahzadi, et al., 2013). They have achieved overall recognition rates between 82% and 86% using NLDs + prosodic + spectral features with 10-fold cross validation. Margarita Kotti and Fabio Paterno (Kotti & Paternò, 2012) have proposed a psychologically-inspired binary cascade classification scheme for speech based emotion recognition using low level audio descriptors and high level perceptual descriptors with Linear SVM. The best emotion recognition accuracy of 87.7% was obtained using SVM with linear kernel. In (Sezgin et al., 2012), a new set of acoustic features based on the perceptual quality metrics are proposed for the binary arousal and valence discrimination which include partial loudness of the emotional difference, emotional difference-to-perceptual mask ratio, measures of alterations of temporal envelopes, measures of harmonics of the emotional difference etc. They had not reported the results for seven classes of emotions discrimination. From the above results and discussions, it can be observed that the proposed method provides better ER accuracy compared to some of the significant works in the literature. Also, for the second database, the proposed algorithms also provided better ER accuracy under different experiments.

CONCLUSION

Generally speaker/gender dependent ER is easier and provides higher ER accuracy. The performance of the speaker/gender independent ER is low compared to speaker/gender dependent ER. In this work, two-stage feature reduction using PCA and LDA was proposed for gender/speaker independent ER from speech. Short-term (MFCCs, LPCs, LPCCs, WLPCCs) cepstral features were extracted from the emotional speech signals. The extracted short-term cepstral features were reduced to fewer dimensions using PCA followed by LDA. Four different classifiers such as *k*-NN, FKNN, MCSVM, and ELM were used to gauge the DimRed features in speaker/gender independent ER. From the simulation results, MCSVM showed good performance in ER for both databases and proposed methods provides very encouraging ER accuracy compared to existing work in the literature.

ACKNOWLEDGMENTS

This project work is supported by Fundamental Research Grant Scheme (FRGS), Malaysia. Grant Number: 9003-00297. All the authors would like to express the deepest appreciation to Prof. Mohammad Hossein Sedaaghi from Sahand University of Technology, Tabriz, Iran for providing me Sahand Emotional Speech database (SESD) for our analysis.

REFERENCES

Bondugula, R., Duzlevski, O., & Xu, D. (2005). Profiles and fuzzy K-nearest neighbor algorithm for protein secondary structure prediction. In Y. P. P. Chen and L. Wong (Ed.), *Proceedings of the 3rd Asia Bioinformatics Conference* (pp. 237-288). London: Imperial College Press.

Bozkurt, E., Erzin, E., Erdem, C. E., & Erdem, A. T. (2010). *Use of line spectral frequencies for emotion recognition from speech.* Paper presented at the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., & Weiss, B. (2005). *A database of German emotional speech.* Paper presented at the Interspeech 2005, Lisbon, Portugal.

Canu, S., Grandvalet, Y., Guigue, V., & Rakotomamonjy, A. (2005). SVM and kernel methods MATLAB toolbox. *Perception Systmes et Information, INSA de Rouen, Rouen, France, 2*, 21.

Chee, L. S., Ai, O. C., Hariharan, M., & Yaacob, S. (2009). *MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA.* Paper presented at the 2009 IEEE Student Conference on Research and Development (SCOReD), Serdang, Malaysia.

Chia Ai, O., Hariharan, M., Yaacob, S., & Sin Chee, L. (2012). Classification of speech dysfluencies with MFCC and LPCC features. *Expert Systems with Applications, 39*(2), 2157-2165.

Deng, H.-B., Jin, L.-W., Zhen, L.-X., & Huang, J.-C. (2005). A new facial expression recognition method based on local gabor filter bank and PCA plus LDA. *International Journal of Information Technology, 11*(11), 86-96.

Dey, S., Rajan, R., Padmanabhan, R., & Murthy, H. A. (2011). *Feature diversity for emotion, language and speaker verification.* Paper presented at the 2011 National Conference on Communications (NCC), Bangalore, India.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: John Wiley & Sons.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572-587.

Giannoulis, P., & Potamianos, G. (2012). *A hierarchical approach with feature selection for emotion recognition from speech.* Paper presented at the Language Resources and Evaluation (LREC), Istanbul, Turkey.

Haq, S., & Jackson, P. (2009). *Speaker-dependent audio-visual emotion recognition.* Paper presented at the International Conference on Audio-Visual Speech Processing, Norwich, UK.

Haq, S., Jackson, P. J., & Edge, J. (2008). *Audio-visual feature selection and reduction for emotion classification.* Paper presented at the Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia.

Hariharan, M., Chee, L. S., Ai, O. C., & Yaacob, S. (2012). Classification of speech dysfluencies using LPC based parameterization techniques. *Journal of Medical Systems, 36*(3), 1821-1830.

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 42*(2), 513-529.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing, 70*(1), 489-501.

Huang, J., Yang, W., & Zhou, D. (2012). *Variance-Based Gaussian Kernel Fuzzy Vector Quantization for Emotion Recognition with Short Speech.* Paper presented at the 2012 IEEE 12th International Conference on Computer and Information Technology (CIT), Chengdu, China.

Iliou, T., & Anagnostopoulos, C.-N. (2010). *SVM-MLP-PNN classifiers on speech emotion recognition field-A comparative study.* Paper presented at the 2010 5th International Conference on Digital Telecommunications (ICDT), Athens, Greece.

Jang, R. (2011). *Audio signal processing and recognition*. Retrieved from http://neural.cs.nthu.edu.tw/jang/books/audiosignal processing/

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man and Cybernetics,*(4), 580-585.

Kim, Y. K., & Han, J. H. (1995). *Fuzzy K-NN algorithm using modified K-selection.* Paper presented at the International Joint Conference of the 4th IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium, Yokohama, Japan.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. *International Journal of Speech Technology, 15*(2), 99-117.

Kotti, M., & Paternò, F. (2012). Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. *International Journal of Speech Technology, 15*(2), 131-150.

Liu, C.-L., Lee, C.-H., & Lin, P.-M. (2010). A fall detection system using k-nearest neighbor classifier. *Expert Systems with Applications, 37*(10), 7174-7181.

Pan, Y., Shen, P., & Shen, L. (2005). *Feature Extraction and Selection in Speech Emotion Recognition.* Paper presented at the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2005), Como, Italy.

Pan, Y., Shen, P., & Shen, L. (2012). Speech Emotion Recognition Using Support Vector Machine. *International Journal of Smart Home, 6*(2).

Rabiner, L. R., & Juang, B.-H. (1993). *Fundamentals of Speech Recognition* (1st ed.): New Jersey: Prentice Hall Englewood Cliffs.

Sedaaghi, M. (2008). Documentation of the *Sahand emotional speech database* (SES). Technical Report, Department of Electrical Engineering, Sahand University of Technology, Iran, .

Sezgin, M. C., Gunsel, B., & Kurt, G. K. (2012). Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing, 2012*(1), 1-21.

Shahzadi, A., Ahmadyfard, A., Harimi, A., & Yaghmaie, K. (2013). Speech emotion recognition using non-linear dynamics features. *Turkish Journal of Electrical Engineering & Computer Sciences*. doi: 10.3906/elk-1302-90

Shen, P., Changjun, Z., & Chen, X. (2011). *Automatic speech emotion recognition using support vector machine.* Paper presented at the 2011 International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT), Heilongjiang, China.

Shlens, J. (2005). A tutorial on principal component analysis: University of California at San Diego.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication, 48*(9), 1162-1181.

Wang, Y., & Guan, L. (2004). *An investigation of speech-based human emotion recognition.* Paper presented at the 2004 IEEE 6th Workshop on Multimedia Signal Processing, Siena, Italy .

Yusuf, S. A. M., Mahat, N. I., Siraj, F., & Yaacob, S. (2012). Noise robustness of first formant bandwidth (f1bw) features in malay vowel recognition. *Journal of Information and Communication Technology, 11*, 147-162.