Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 3741 – 3757

doi:10.3906/elk-1903-121

Speech emotion recognition using semi-NMF feature optimization

Surekha Reddy BANDELA*, T. Kishore KUMAR
Department of Electronics and Communication Engineering, NIT Warangal, Telangana, India

Received: 20.03.2019 • **Accepted/Published Online:** 05.06.2019 • **Final Version:** 18.09.2019

Abstract: In recent times, much research is progressing forward in the field of speech emotion recognition (SER). Many SER systems have been developed by combining different speech features to improve their performances. As a result, the complexity of the classifier increases to train this huge feature set. Additionally, some of the features could be irrelevant in emotion detection and this leads to a decrease in the emotion recognition accuracy. To overcome this drawback, feature optimization can be performed on the feature sets to obtain the most desirable emotional feature set before classifying the features. In this paper, semi-nonnegative matrix factorization (semi-NMF) with singular value decomposition (SVD) initialization is used to optimize the speech features. The speech features considered in this work are mel-frequency cepstral coefficients, linear prediction cepstral coefficients, and Teager energy operator-autocorrelation (TEO-AutoCorr). This work uses k-nearest neighborhood and support vector machine (SVM) for the classification of emotions with a 5-fold cross-validation scheme. The datasets considered for the performance analysis are EMO-DB and IEMOCAP. The performance of the proposed SER system using semi-NMF is validated in terms of classification accuracy. The results emphasize that the accuracy of the proposed SER system is improved remarkably upon using the semi-NMF algorithm for optimizing the feature sets compared to the baseline SER system without optimization.

Key words: Speech emotion recognition, spectral, Teager energy operator, feature fusion, semi-nonnegative matrix factorization, k-nearest neighborhood, support vector machine

1. Introduction

Speech emotion recognition (SER) is the process of detecting the emotional state of a speaker from speech signals. The field of emotion recognition has gained a lot of interest in human–computer interaction these days, and much research is going on in this area using different feature extraction techniques and machine learning algorithms. SER is used in a wide range of applications like call-center services, in vehicles to know the psychological state of the person who is driving, as a diagnosing tool in medical services, in story-telling and in E-tutoring applications, and so on. Basically, there are six archetypal emotions: anger, neutrality, happiness, disgust, surprise, fear, and sadness [1,2]. In situations where only a person's speech signals are available, SER plays a prominent role.

A major challenge in SER is to identify the speech features that can effectively extract the emotional characteristics from a speech signal. Speech features can be classified as continuous, voice quality, spectral, and nonlinear Teager energy operator (TEO)-based features [1]. The categorical representation of a few of these speech features is shown in Figure 1. The continuous prosodic features are pitch, zero crossing rate, energy, formants, etc., which have an effect on the emotional variation of a speech signal. Among all these features,

*Correspondence: elektrik@tubitak.gov.tr

pitch has a huge variation for different emotions in humans and has been extensively used in the development of an SER system to characterize emotions [3–5]. Voice quality features have a strong relationship with perceived emotion [6]. These are categorized as voice pitch, voice level, temporal and feature boundary structures, jitter, and shimmer [7], glottal waveforms and its variants [8–10], etc., which are useful for speech emotion recognition. The spectral features are represented as the short-time representation of the speech signal. The spectral energy distribution of a speech signal varies with its emotional content. Based on this, emotions are classified as high-arousal and low-arousal emotions. High-arousal emotions have higher energies at high frequencies, like happiness or anger, whereas low-arousal emotions have less energy in the same range of frequencies, like sadness. Compared to other speech features, spectral features are able to characterize emotional contents more accurately [11–13]. It is well known that there is a nonlinear airflow during the speech production process in the vocal tract system [14]. Under stressful conditions, the flow of air in the vocal tract system is affected by the muscle tension of the speaker while producing sounds. These nonlinear speech features are highly affected when stressed emotional speech signals are produced. The nonlinear TEO was developed by Teager and Kaiser to enhance the quality of stressed speech signals [14,15]. Many TEO features were developed to characterize stressed emotions [16–18]. The TEO features are also combined with glottal features to further improve the stressed speech emotion recognition performance [19,20].

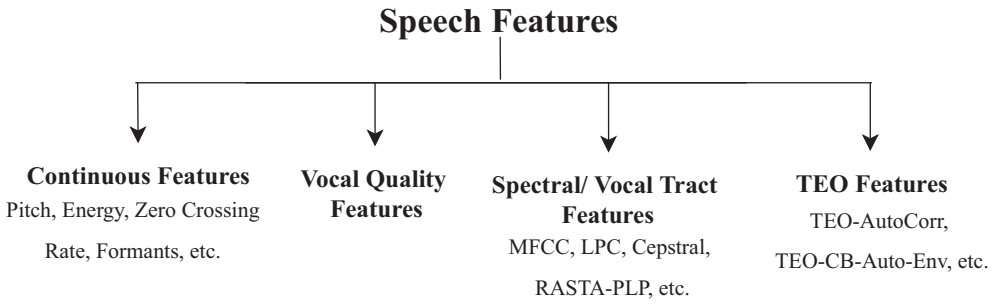

Figure 1. Categorization of speech features.

Most of the research related to the development of SER systems is mainly focused on identifying the speech features that can characterize emotions effectively. Mel-frequency cepstral coefficients (MFCCs) [11,21,22], linear prediction coefficients (LPCs) [23], relative spectral perceptual linear prediction (RASTA-PLP) [16], and variants of these features like modified MFCC (M-MFCC) [13], feature fusion of MFCC, and short-time energy features with velocity (Δ) and acceleration ($\Delta+\Delta$) [23] are some of the well-known spectral features that are used for speech emotion recognition. Apart from these, log frequency power coefficients (LFPCs) [24], Fourier parameter features [25], time-frequency features with AMS-GMM mask [26], modulation spectral features [27], and amplitude-based features [28] are some of the variants of spectral features that are now used in SER analysis. Among all these features, MFCC is the most widely used feature for SER that gave promising results. Hence, in most studies, the MFCC feature set is used as a benchmark feature set to analyze the performance of the rest of the SER systems [11,12,25,26,29]. It is evident from the literature that the combination of speech features, i.e. feature fusion, increases the classification accuracy of the SER system [6,23,28] and hence became the most common practice in this field.

1.1. Motivation

Despite the fact that feature fusion increases the performance of SER systems in terms of classification accuracy, it also increases the computational overhead on the classifier. This is because some of the features contribute

in a better way, while some of them might not be useful at all in the emotion recognition process. Using the irrelevant speech features leads to the curse of dimensionality and decreases the performance of the SER system as shown in Figure 2, where the performance of the system in terms of classification accuracy is decreased after a particular feature dimension threshold with an increase in the dimension of the feature set. By choosing an appropriate feature dimension, optimal performance can be achieved. Another disadvantage of increasing the number of speech features is that it will increase the computational complexity, and it also causes the overfitting problem, i.e. the model achieves better accuracy while training but fails when tested on new data [30,31]. These drawbacks can be overcome by adopting feature optimization techniques before emotion classification. Therefore, it is always preferable to perform feature selection or optimization of the feature sets before emotion classification. There are several feature selection and optimization techniques for dimension reduction of the feature set to overcome the disadvantages of having huge feature sets. In the feature selection, a subset of the original features is selected, which retains the desired feature set. In the case of feature optimization, the feature space is transformed into another domain and the discriminant feature information is concentrated in a particular part of the coefficients in the transformed domain. Several feature selection techniques are being used by researchers to select the most appropriate feature set [32,33].

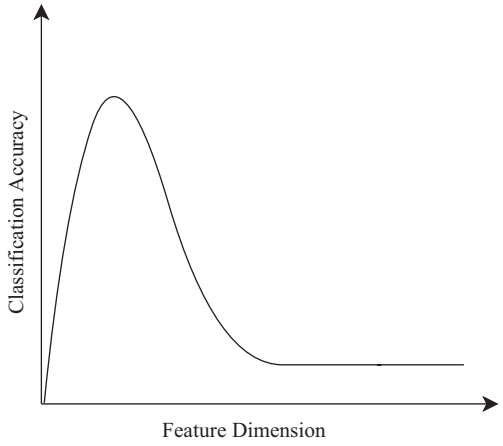

Figure 2. Curse of dimensionality.

Feature dimension reduction is the best way to solve the problem of high dimensionality, but the reduction of the number of feature vectors causes an uncertain loss in the information and subsequently leads to instability in the performance of the system. This problem can be overcome by using linear transformation techniques. If an n-dimensional input feature vector $x = [F_1, F_2, ..., F_n]^T$ is considered, then the transformed output vector will be $y = [b_1, b_2, ..., b_r]^T$, where r$<<$n and r is the reduced dimension. For this purpose, many optimization techniques are developed in machine learning to acquire the most optimal feature sets that improve the SER accuracy. These techniques can be classified based on feature set labeling as supervised or unsupervised. In supervised techniques, the feature sets are labeled, and in unsupervised techniques, feature sets are not labeled [30,34]. These techniques are further classified based on feature transformation as linear and nonlinear techniques, in which the high-dimensional feature sets are scaled down to a lower-dimensional space preserving the locality and geometric structures. The taxonomy of the feature optimization techniques is shown in Figure 3. In linear transformation, the structure of a given feature set is determined using Euclidean distance based on second-order statistics, whereas the nonlinear transformation techniques recover the useful and meaningful submanifolds from high-dimensional datasets [34].

In SER, principal component analysis (PCA) [3,32] is one of the important and most widely used feature optimization techniques, which is based on feature selection. Many other feature optimization techniques such as linear discriminant analysis (LDA), which is a supervised machine learning technique [3], and singular value decomposition (SVD) [35], locally linear embedding (LLE) [36], and nonnegative matrix factorization (NMF) [37] are unsupervised feature optimization techniques that are commonly used for speech emotion recognition. In SVD and NMF, the complete set of features transforms with matrix factorization to obtain a lower-dimensional feature set, acquiring an optimal feature set. In [38,39], variants of autoencoders, namely adversarial and variational autoencoders, were used to transform huge feature sets into lower dimensions and this reduced feature set was used for SER to acquire high performance.

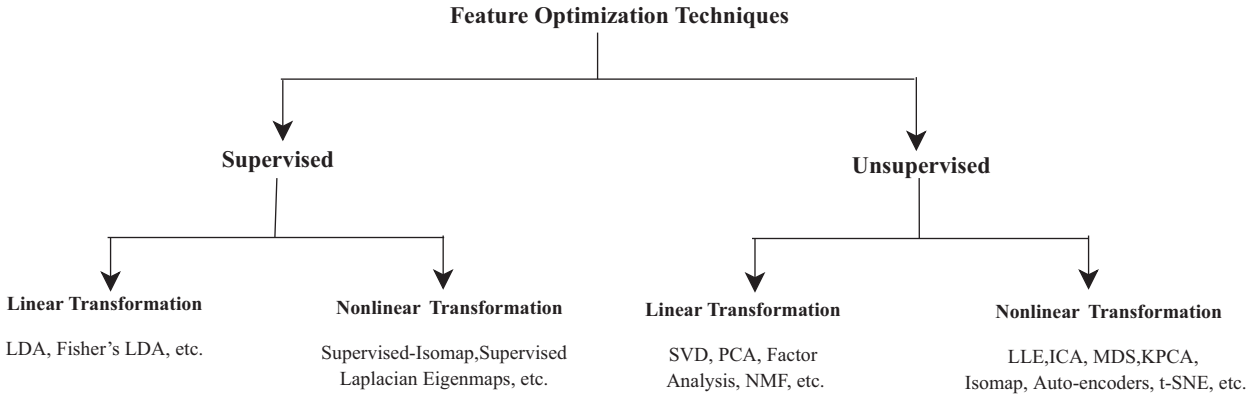

Figure 3. Taxonomy of feature optimization techniques.

The major contribution of this paper is to optimize the MFCC, LPCC, and TEO-AutoCorr features using the semi-NMF algorithm with SVD initialization so as to improve the performance of the proposed SER system. The rest of the paper is organized as follows: the proposed SER system using the semi-NMF optimization technique with SVD initialization is introduced in Section 2. The experimental analysis and simulation results of the proposed SER system compared with the existing ones are discussed in Section 3, and Section 4 concludes the paper.

2. Proposed speech emotion recognition system

A conventional SER system consists of only three stages: speech preprocessing, feature extraction, and classification [40]. Most of the existing SER systems use the entire set of speech features for emotion recognition so far. This increases the computational overhead on the classification model. In order to overcome this drawback, the semi-NMF optimization technique is incorporated in the development of the SER system before classifying the features for obtaining the emotions, as shown in Figure 4.

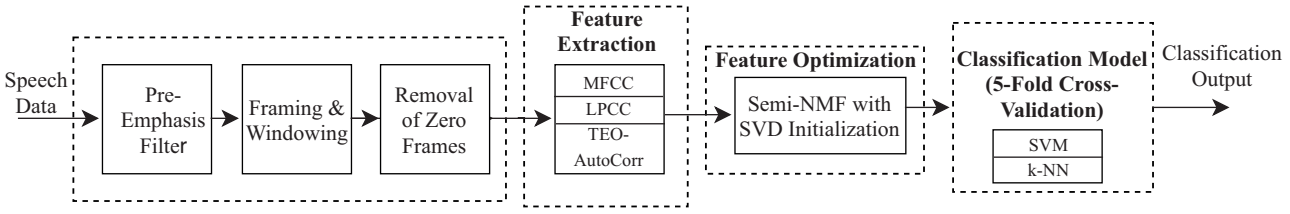

Figure 4. Proposed speech emotion recognition system.

2.1. Preprocessing

The speech signal is preprocessed before feature extraction and the stages involved are filtering, framing, and windowing [40]. The preemphasis filter performs as a first-order high-pass filter to boost the energy of the speech signal in the higher frequencies, which are attenuated in the speech signal production process. If s[n] is the speech signal, the time domain and z-domain representations of a preemphasis filter are given as:

$$h[n] = s[n] - \alpha s[n-1] \; (or) \; H(z) = S(z)[1 - \alpha z^{-1}]. \tag{1}$$

Here α is the filter coefficient and its value must be between 0.9 and 1 [40]. It is well known that the speech signal is not stationary; hence, it is difficult to analyze speech signals. To overcome this problem, the preemphasized speech signal is framed into an equal number of samples so that each frame can be considered as stationary so that the signal processing techniques can be applied. Each frame consists of an equal number of samples, which is also called frame length. The number of frames varies from one speech signal to another depending on the length of the speech signal. When the signal is divided into frames, there exist some discontinuities at the edges of the frames of the input speech signal. In order to avoid these discontinuities, each frame is passed through a tapered window. There are different types of windows used in the speech preprocessing, like Hamming, Hanning, Barlett, etc. Among these, the Hamming window is chosen in this work, as it provides less spectral leakage at the edges of the frames. The window size is chosen based on the frame length ('N'). The Hamming window is given by [40]:

$$w[n] = 0.54 - 0.46\cos(2\pi\frac{n}{N}) \;\; where \;\; 0 \leq n \leq N. \tag{2}$$

Here, N is the window size and n is the speech signal length. An overlap between the frames is allowed so that there is no loss in the speech signal information. In this paper, the frame length is considered to be 256 and the overlap allowed between the frames is chosen as 80. Later, preemphasized speech signal '$h[n]$' is multiplied by this window function by allowing a frame overlap to obtain the resultant signal. After preprocessing, the speech frames are fed to the feature extraction block:

$$x[n] = h[n] \times w[n] \tag{3}$$

2.2. Feature extraction

Feature extraction in SER is the process of extracting the specific speech features that portray the emotion-relevant information. Instead of using the speech signal directly to classify emotions, a particular set of features can be extracted using various signal processing techniques with which the emotions can be classified using different classification techniques. MFCC, LPCC, and TEO-AutoCorr features are extracted in the development of the proposed SER system.

MFCC features contribute mostly in SER system development, as these features are designed based on the human ear speech perception. In MFCCs, initially the speech signal frames are transformed into the frequency domain using DFT and the transformed frames are fed to the mel-filter bank to convert the log frequency-scale to the mel-frequency scale, which mimics the perception of a human ear [40]:

$$mel(f) = 2595 \times log_{10}(1 + f/700) \;\; (or) \;\; mel(f) = 1127 * ln(1 + f/700) \tag{4}$$

Here, f is the frequency of the transformed speech signal. These transformed mel-frequency domain features are further converted to the cepstrum domain using the discrete cosine transform (DCT). A total of 12 MFCC

features are extracted in this process. It is well known that the difference between the consecutive MFCC features, which are termed as Δ (delta) features, contributes efficient emotion recognition. Hence, a total of 24 features with 12 MFCC and 12 Δ (delta) features are extracted.

In the extraction of LPCCs, initially, linear predictive analysis is performed on the speech signal. The basic idea behind the linear predictive analysis is that the nth speech sample can be estimated by a linear combination of its previous p samples as shown in the following equation:

$$x[n] \approx a_1 x[n-1] + a_2 x[n-2] + a_3 x[n-3] + ... + a_p x[n-p] \tag{5}$$

Here, a_1, a_2, a_3, . . . a_p are assumed to be constants over a speech analysis frame. These are known as predictor coefficients or linear predictive coefficients. These coefficients are used to predict the speech samples. The difference between actual and predicted speech samples is known as an error. It is given by:

$$e[n] = s[n] - \hat{s}[n] = s[n] \sum_{i=1}^{p} [a_i x[n-i]] \tag{6}$$

Here, e[n] is the error in prediction, s[n] is the original speech signal, $\hat{s[n]}$ is a predicted speech signal, and a_i for i=1,2 ,……p are the predictor coefficients. Later, the cepstral coefficients are derived from the LPCCs derived using the following recursion [41]:

$$\begin{aligned} C_0 &= log_e[p] \\ C_m &= a_m + \sum_{i=1}^{m-1} \frac{i}{m} C_i a_{m-i}, for\ 1 < m < p \\ C_m &= \sum_{i=m-p}^{m-1} \frac{i}{m} C_i a_{m-i}, for\ m > p \end{aligned} \tag{7}$$

The resultant $[C_0, C_1, ...C_m]$ are the LPCCs with k = m+1 features. The LPCC feature extraction is designed to obtain 21 features and is used in this analysis.

Even though MFCCs and LPCCs are widely used for SER, a few of the stressed emotions like anger or anxiety could not be analyzed properly. Therefore, TEO features are also used in this work. The TEO-AutoCorr feature extraction is shown in Figure 5.

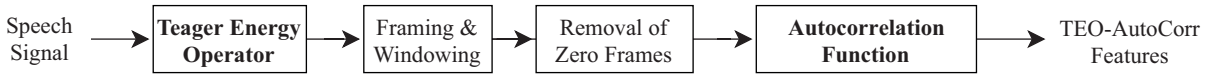

Figure 5. TEO-AutoCorr feature extraction.

Teager [14,15] proposed an energy operator, i.e. a measure of speech signal energy, based on his experiments known as the TEO. In the experiments, Teager showed that the flow of air in the vocal tract is separated and follows the vocal tract walls. Later, Teager conducted several experiments on the hearing process and came up with a measurement of the energy parameter to find proof of speech modulation patterns. The energy operator is as follows [15,16]:

$$\psi(x(t)) = \left(\frac{\mathrm{d}}{\mathrm{d}t}x(t)\right)^2 - x(t)\left(\frac{\mathrm{d}^2}{\mathrm{d}t^2}x(t)\right)\ or\ \psi(x[n]) = x^2[n] - x[n]x[n+1] \tag{8}$$

Here, x(t) and x[n] are the speech signals in the continuous and discrete domain.

The autocorrelation function is the correlation of a signal with a delayed copy of itself as a function of delay. Informally, it is the similarity between observations as a function of the time lag between them:

$$R_{xx}(k) = \sum_{n=k}^{M-1} s[n]s[n-1] \tag{9}$$

Here, x(t) is the input signal to the function and τ is the delay parameter. When the frames of the Teager energized signal are given to the autocorrelation function, the correlation between the adjacent frames is obtained. If the correlation is high, the energy of the speech signal is further increased, resulting in the TEO-AutoCorr features. All the extracted features, e.g., $[F_1, F_2, ..., F_k]$, are further fed to the feature optimization block.

2.3. Semi-NMF using SVD initialization

The semi-NMF technique is a variant of the NMF algorithm and can be used for speech feature optimization. In this paper, the semi-NMF algorithm using SVD initialization is employed to optimize the speech features. Semi-NMF has been widely used in many data processing applications like data analysis and clustering [42]. The data matrix, i.e. a feature matrix M = $[F_1, F_2, \ldots\ldots\ldots F_k]$ with k as the feature vectors that are unconstrained (i.e. it may have mixed signs), is considered. A factorization that is referred to as semi-NMF in [42], in which V is restricted to be nonnegative while placing no restriction on the signs of U, is proposed. Semi-NMF can be defined as follows: Given a matrix $M \epsilon R^{m \times k}$ and a factorization rank r, solve

$$min_{U \epsilon R^{m \times r}, V \epsilon R^{r \times k}} \|M - UV\|_F^2 \ \ such\ that\ V \geq 0 \tag{10}$$

where $\|.\|_F$ is the Frobenius norm and $V \geq 0$ means that V is component-wise nonnegative. The concept of semi-NMF is motivated from the perspective of k-means clustering that can be applied to an input feature vector M to obtain cluster centroids, $U = u_1, u_2, ..., u_r$, where V is the cluster indicator [42]:

$$V = \begin{cases} 1 & if F_i \, \epsilon \, clusterc_r \\ 0 & otherwise \end{cases} \tag{11}$$

However, there are convergence issues in this method, due to which a different initialization technique rather than k-means can be used, since the initialization of U and V matrices is important to obtain an optimal solution for the factorization problem. The semi-nonnegative rank of matrix M can be denoted by M = UV with $U \epsilon R^{m \times r}$, $V \epsilon R^{r \times k}$, and $V \geq 0$. To summarize,

$$NMF \rightarrow X_+ \approx U_+ V_+^T \ \ \& \ \ Semi - NMF \rightarrow X_+ \approx U_\pm V_+^T \tag{12}$$

In other words, NMF has both U and V with nonnegative values, whereas semi-NMF has U consisting of both positive and negative values without any restriction and V with only nonnegative values. Accordingly, a singular value decomposition (SVD) and linear programming-based method are proposed to overcome the drawbacks of the basic semi-NMF for finding the optimal solution [43].

The semi-NMF technique with SVD initialization is discussed in Algorithm. In step 1, we apply SVD on the data or feature matrix to obtain the left singular matrix (A), diagonal matrix (S), and right singular matrix

Algorithm Semi-NMF using SVD initialization.

Input: A matrix $M \epsilon R^{m \times k}$, a factorization rank r.
Output: A rank-r semi-NMF (U,V) of $M \approx UV$ with $V \geq 0$

1: [A,S,B^T]=svds(M,r); 'svds' is a MATLAB function
2: For each $1 \leq i \leq r$: multiply B(i,:) by –1 if min_j B(i,j)$\leq min_j(-B(i,j))$;
3: Let (y^*, ε^*) be the optimal solution of the following optimization problem:
$\min_{y \epsilon R^r, \varepsilon \epsilon R^+} \varepsilon$ such that $(B(:,j) + \varepsilon e)^T y \geq 1 \; \forall \; j$
% if ϵ^*=0 ($\Leftrightarrow$ B is semi-nonnegative) then the heuristic is optimal
4: $x = (B + \varepsilon^* 1_{r \times k})^T y^* \geq 1$;
5: $\alpha_i = max(0, max_j \frac{-B(i,j)}{x(j)})$ for all $1 \leq i \leq r$
6: V=B+αx^T;
7: U$\leftarrow argmin_{x \epsilon R^{m \times r}} \|M - XV\|_F^2$

(B) considering rank-r approximation. Among these, matrix B is considered for further analysis. In step 2, the rows of matrix B are flipped. Further, in step 3, the actual optimization takes place, i.e. a heuristic for finding the optimal solution (y^*, ε^*) with r,

$$min_{y \epsilon \mathbb{R}^r \varepsilon \epsilon \mathbb{R}_+} \varepsilon \; such \; that \; (B(:,j) + \varepsilon e)^T y \geq \; 1 \forall j \; such \; that \; (B(:,j) + \varepsilon e) \neq 0 \tag{13}$$

Here e is the vector of all ones. If the value of ϵ^* is too small the probability of B being a semi-nonnegative matrix is high. Eq. (13) is solved using a bisection method on variable ϵ^*. In this method, if $\epsilon^* = 0$ initially, then an optimal semi-NMF can be obtained. Once the optimal solution (y^*, ε^*) is obtained, the matrices V and further the desired optimal solution U can be obtained in the consecutive steps.

In this work, the rank-r of the semi-NMF is chosen so as to acquire optimum performance. The MFCC, LPCC, and TEO-AutoCorr features will be scaled down to r number of features each. Using semi-NMF, the $m \times 24$ feature matrix is factorized into U_{mel} and V_{mel} matrices with $m \times r_1$ and $r_1 \times 24$. Likewise, the LPCC feature vector with $m \times 21$ dimensions, using semi-NMF, is factorized into U_{lp} and V_{lp} feature and coefficient matrices with $m \times r_2$ and $r_2 \times 21$. Similarly, in the case of TEO-AutoCorr features, using semi-NMF, the $m \times 20$ feature vector matrix is factorized into U_{teo} and V_{teo} feature and coefficient matrices with $m \times r_3$ and $r_3 \times 20$. Here, the matrices U_{mel}, U_{lp}, and U_{teo} are the desired optimal MFCC, LPCC, and TEO-AutoCorr feature vectors consisting of both the positive and negative data. The ranks r_1, r_2, and r_3 are chosen based on the type of the classifier (SVM or k-NN) and the feature set chosen (i.e. MFCC or LPCC or TEO-AutoCorr) by validating the performance of the SER system.

2.4. Classification

There are many pattern recognition algorithms that are used for emotion classification [1,2]. In this paper, k-nearest neighborhood (k-NN) with k = 6 folds and support vector machine (SVM) classification techniques with Gaussian kernel are used to classify the emotions. The optimized feature set $[O_1, O_2, ..., O_r]$ is given to the classification model for classifying the emotions. The classification techniques considered are supervised and hence the optimized feature sets are labeled with their corresponding emotions. The k-fold cross-validation is a resampling method employed to evaluate machine learning models on a limited dataset. The dataset is randomly divided into k groups or folds of nearly equal size. The first fold is used as a validation set, and the model is fit on the remaining k–1 folds. Cross-validation is basically applied to the machine learning algorithms

in order to estimate the skill of the developed model on unseen data. It is the most commonly used method because it is easy to understand and results in a less biased or less optimistic estimate of the model's skill than other techniques, such as a simple train/test split. In this work, a 5-fold cross-validation schema is used to train the classifiers, in which the training and testing are carried out in 5 folds.

3. Results

The database considered in the development of a SER system is one of the most important challenges faced by researchers. This is because of the variation of classification accuracy of the SER system with different language datasets because of different speaking styles. The EMO-DB and IEMOCAP datasets are considered in this work to analyze the performance of the proposed SER system. EMO-DB, a German database [44], is widely used in SER analysis by many researchers. The recording for emotional data was done in an anechoic chamber with 5 male and 5 female actors between the ages of 25 and 35. A total of 535 speech signals were recorded at 48 kHz with anger, boredom, disgust, anxiety/Fear, happiness, sadness, and neutrality. Later these were downsampled to 16 kHz. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is in English [45]. It is an acted, multimodal, and multispeaker database comprising 12 h of audiovisual data that include video, speech, text transcriptions, and motion capture of the face. The speech data with emotions of anger, excitement, frustration, happiness, neutrality, and sadness are considered in this work, with a total of 7112 utterances.

Table 1. Simulation parameters of the proposed SER system.

Parameters	Specifications
Preemphasis filter	Coefficients, a = 0.97
Frame size/length	256 samples
Frame overlap	80 samples
Type of window	Hamming
Mel-filter banks	20
Semi-NMF	SVD Initialization
k-NN	k-Folds = 6 Euclidean distance measure
SVM	Gaussian kernel

As discussed in Section 2, the 24 MFCC, 21 LPCC, and 20 TEO-AutoCorr features are extracted. The simulation parameters used in the development of the proposed speech emotion recognition system, i.e. for speech preprocessing, feature extraction, optimization, and classification, are shown in Table 1. In the proposed SER system, the 24 MFCC, 21 LPCC, and 20 TEO-AutoCorr features are optimized using the semi-NMF algorithm using SVD initialization and the k-NN and SVM classifiers are used for emotion classification. The performance of the proposed SER system is evaluated using the machine learning performance metric, i.e. classification accuracy. All the simulations are carried out on a computer with Intel Xeon CPU E3-1220 v3 of a 3.10 GHz 64-bit processor with 16 GB RAM.

The entire set of emotional data of both the datasets is considered in this analysis and the corresponding results are shown in Figures 6 and 7 and Tables 2 and 3. In this work, the 5-fold cross-validation schema is used to train and test the accuracy of the proposed SER system. Hence, the entire dataset is randomly split into

5 parts, among which 4 parts are used for training the classifier (k-NN or SVM) and testing is carried out on the remaining or test data, i.e. the fifth part. This process is repeated in 5 folds, i.e. 5 times, until the entire dataset is completely trained. The evaluated score (i.e. the classification accuracy) at each fold is retained and, finally, the mean of these scores is calculated to obtain the overall classification accuracy of the proposed system.

The number of features into which the features get optimized depends on the rank of the semi-NMF. The choice of choosing the rank of the semi-NMF algorithm for optimizing the features in order to achieve high performance is very important. In order to decide the optimal rank, the optimization is performed individually on MFCC, LPCC, and TEO-AutoCorr features using different ranks of semi-NMF and these optimized features are classified using the classification models.

Figures 6 and 7 show the variation of classification accuracy of the proposed SER system with the EMO-DB and IEMOCAP databases for different ranks of semi-NMF with which the MFCC, LPCC, and TEO-AutoCorr features are optimized using SVM and k-NN classifiers. From the results, it is clearly understood that the performance of the SER system not only varies with the type of database used but also the classification model considered. From these figures, it can also be observed that the SER classification accuracy is increased with the rank of the semi-NMF considered, i.e. with an increase in the number of features, whereas after a particular rank of semi-NMF, the classification accuracy of the SER system is decreased, thus implying the curse of dimensionality. The rank at which utmost accuracy is obtained is considered to be the optimal rank. From Figures 6a and 6b, for the EMO-DB database the highest accuracy is achieved for MFCCs, when optimized with Rank-20 and Rank-22 for SVM and k-NN classifiers, as 67.76% and 85.6%, respectively. In the case of optimized LPCC and TEO-AutoCorr features using the SVM classifier, the highest accuracy is achieved with Rank-18 as 73.65% and 68.56%, respectively. Using the k-NN classifier, 88.2% and 83.7% classification accuracies are obtained for LPCC and TEO-AutoCorr features optimized at Rank-19. Therefore, the optimal ranks for MFCCs are 20 and 22 for SVM and k-NN, respectively. Similarly, for LPCCs and TEO, the optimal ranks are 19 and 18 using SVM and k-NN. From Figures 7a and 7b, for the IEMOCAP database the highest accuracy is achieved for MFCCs, when optimized with Rank-19, as 72% and 74.1% for SVM and k-NN classifiers, respectively. In the case of optimized LPCCs, the highest accuracy is achieved with Rank-17 as 79.6% for the SVM classifier and with Rank-12 as 75% for the k-NN classifier. Likewise, for TEO-AutoCorr features, the highest accuracy is achieved with Rank-17 as 67.7% using the SVM classifier and with Rank-19 as 71% using the k-NN classifier. The optimal rank for MFCC is 19 for both SVM and k-NN. Likewise, using SVM and k-NN classifiers, for LPCCs the optimal ranks are 17 and 12, whereas in the case of TEO features the optimal ranks are 17 and 19.

Tables 2 and 3 show the results of the performance comparison of MFCCs, LPCCs, TEO-AutoCorr, and their combinations without optimization and with the semi-NMF optimization technique validated using SVM and k-NN classifiers. From these results, it is clearly evident that by combining the features the performance is improved. This is the reason behind the extensive usage of huge feature sets for SER development.

From Table 2 for the EMO-DB database, it is observed that the highest accuracy is achieved with the feature fusion of the optimized MFCC, LPCC, and TEO-AutoCorr features with 56 features obtaining 90.12% accuracy using SVM and 89.3% accuracy using the k-NN classifier with 60 features. The minimum number of features at which the highest accuracy is obtained for the proposed system is 73.65% for SVM and 88.2% for k-NN, with LPCC features optimized at Rank-18 and Rank-19, respectively. Similarly, from Table 3 for the IEMOCAP database, the highest accuracy is achieved with the feature fusion of the optimized MFCC, LPCC, and TEO-AutoCorr features with 53 features obtaining 83.2% accuracy using SVM and 78% accuracy using the

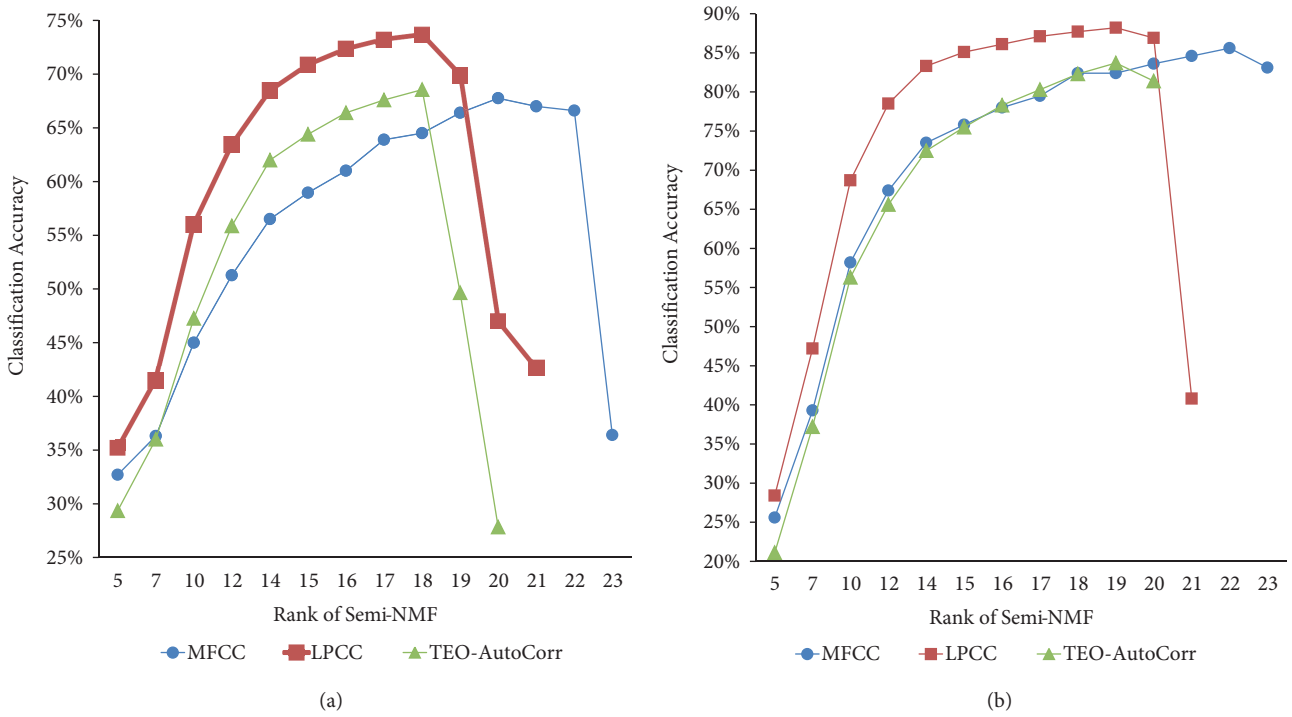

Figure 6. Performance variation of the proposed SER system with different ranks of semi-NMF for MFCC, LPCC, and TEO-AutoCorr features using a) SVM and b) k-NN for EMO-DB database.

Table 2. Performance Comparison of Baseline and Proposed SER system with Semi-NMF for EMO-DB Database using SVM & k-NN Classifiers for Different Feature Sets.

Optimization techniques	Features	SVM		k-NN	
		No. of features	Classification accuracy	No. of features	Classification accuracy
Baseline	MFCC	24	47%	24	53%
	LPCC	21	45%	21	40.8%
	TEO-AutoCorr (TEO)	20	31.8%	20	27.3%
	MFCC+LPCC	45	64.2%	45	70.5%
	MFCC+TEO	44	62.5%	44	74.6%
	LPCC+TEO	42	46.47%	42	51.7%
	MFCC+LPCC+TEO	65	55.36%	65	69.9%
Semi-NMF with SVD	MFCC	20	67.76%	22	85.6%
	LPCC	18	73.65%	19	88.2%
	TEO-AutoCorr (TEO)	18	68.56%	19	83.7%
	MFCC+LPCC	38	85%	41	89%
	MFCC+TEO	38	81.54%	41	87.8%
	LPCC+TEO	36	84.13%	38	88.7%
	MFCC+LPCC+TEO	56	90.12%	60	89.3%

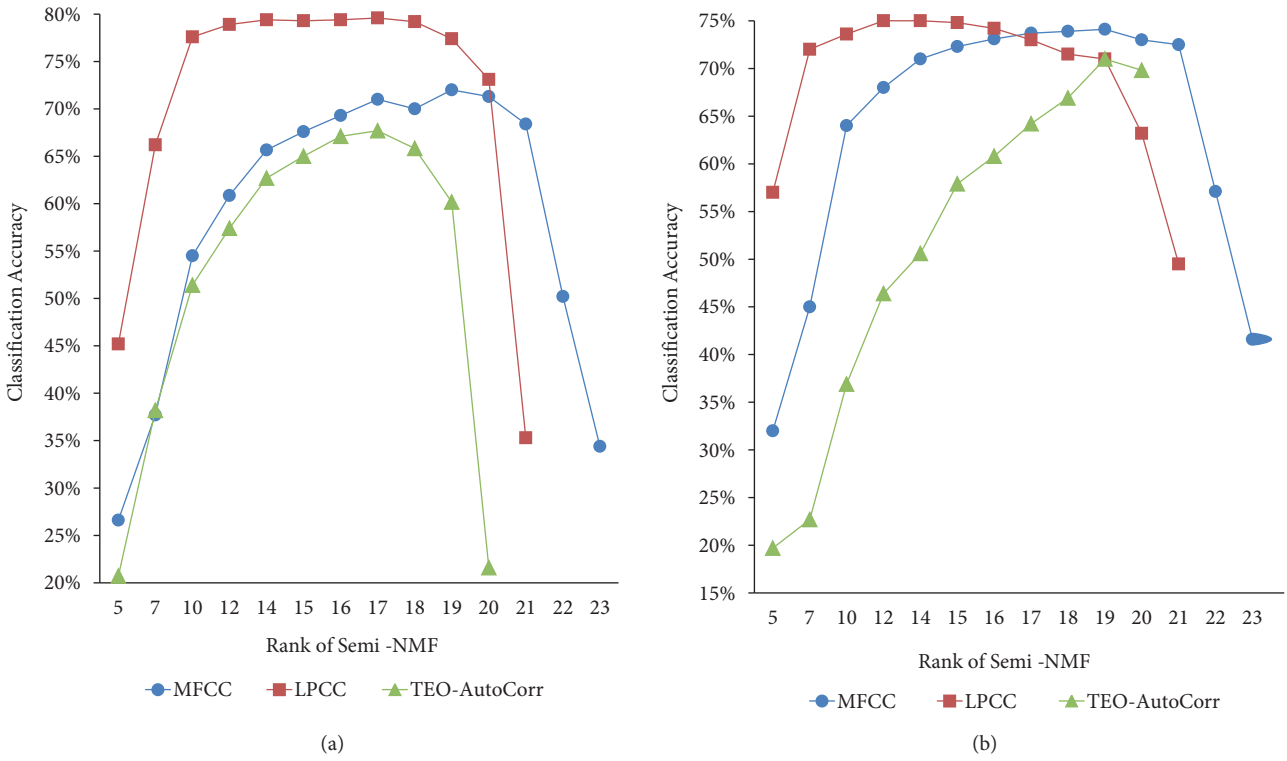

Figure 7. Performance variation of the proposed SER system with different ranks of semi-NMF for MFCC, LPCC, and TEO-AutoCorr features using a) SVM and b) k-NN for IEMOCAP database.

Table 3. Performance comparison of baseline and proposed SER system with semi-NMF for IEMOCAP database using SVM and k-NN classifiers for different feature sets.

Optimization techniques	Features	**SVM**		**k-NN**	
		No. of features	Classification accuracy	No. of features	Classification accuracy
Baseline	MFCC	24	44%	24	50.57%
	LPCC	21	41.1%	21	39.95%
	TEO-AutoCorr (TEO)	20	31.2%	20	29.2%
	MFCC+LPCC	45	45.2%	45	52.36%
	MFCC+TEO	44	43.4%	44	47.24%
	LPCC+TEO	42	42.3%	42	35.4%
	MFCC+LPCC+TEO	65	50.34%	65	55.63%
Semi-NMF with SVD	MFCC	19	72%	19	74.1%
	LPCC	17	79.6%	12	75%
	TEO-AutoCorr (TEO)	17	67.7%	19	71%
	MFCC+LPCC	36	82.88%	31	74.98%
	MFCC+TEO	36	79.23%	38	75%
	LPCC+TEO	34	82.6%	31	74.32%
	MFCC+LPCC+TEO	53	83.2%	50	78%

k-NN classifier with 50 features. The minimum number of features at which the highest accuracy is obtained for the proposed system is 79.6% for SVM and 75% for k-NN, with LPCC features optimized at Rank-17 and Rank-12, respectively.

Table 4. Comparison of the proposed SER system with the existing methods for EMO-DB database.

Approaches		No. of optimized features	Classification accuracy
Chen et al. [46]		72	77.74%
Zhang et al. [47]		9	80.85%
Zhang et al. [48]		11	73.9%
Yan et al. [49]		13	79.23%
Kuchibhotla et al. [50]		12	88.1%
Daneshfar et al. [51]		20	79.22%
Gudmalwar et al.[52]		36	75.32%
Özseven [53]		304	84.07%
Sun et al. [54]		500	86.86%
Proposed	SVM	18	73.65%
	k-NN	56	90.12%
	SVM	19	88.2%
	k-NN	60	89.3%

Furthermore, the performance of the proposed SER system is compared with different works in the Table 4 for EMO-DB and Table 5 for IEMOCAP in terms of the number of optimized features and classification accuracy performance measures.

Table 5. Comparison of the proposed SER system with the existing methods for IEMOCAP database.

Approaches		No. of optimized features	Classification accuracy
Sahu et al. [38]		100	58.38%
Latif et al. [39]		128	56.42%
Proposed	SVM	17	79.6%
		53	83.2%
	k-NN	12	75%
		60	78%

In [46], semi-NMF with k-means clustering initialization was used to transform feature sets, which were further combined with the original dataset to obtain a total of 72 features for SER obtaining 77.74% accuracy. In [47–52], different optimizing and feature selection techniques, namely enhanced kernel isometric mapping, the modified supervised locally linear embedding algorithm, sparse partial least squares regression, sequential floating forward selection, the scaled conjugate gradient, and principal component analysis, were used for improving the classification accuracy by reducing the feature set dimension. However, the classification accuracy obtained with the proposed SER system is higher than the other methods with 90% (approx.) using both classification techniques for the EMO-DB database. In [38,39], a new statistical feature selection and Fisher feature selection were used to select the most optimal feature sets, but still these techniques [38,39] have

lower performance both in terms of complexity, i.e. number of features, and classification accuracy compared to the proposed SER system. Likewise, in [44,45], variational and adversarial autoencoders were used for feature optimization and the performance was lower than that of the proposed SER system for the IEMOCAP database with classification accuracy of 79.6% for 17 features and 83.2% for 53 features using the SVM classifier and 75% for 12 features and 78% for 60 features using the k-NN classifier.

4. Conclusion

In the proposed SER system, the semi-NMF feature optimization technique with SVD initialization is employed to optimize the MFCC, LPCC, and TEO-AutoCorr features. The performance of the proposed SER system is analyzed with the EMO-DB and IEMOCAP databases using k-NN and SVM classifiers. A five-fold cross-validation scheme is used to train the feature sets so as to consider the entire dataset for both training and testing to avoid overfitting problems. The optimal rank is chosen for semi-NMF depending on the database and classification technique used for MFCC, LPCC, and TEO-AutoCorr features. The combination of these optimized feature sets is used in the proposed SER system to achieve highest classification accuracies using SVM and k-NN classifiers with 90.12% and 89.3% for the EMO-DB database and 83.2% and 78% for the IEMOCAP database, respectively. It is clearly evident from the results that the proposed SER system outperforms the baseline, i.e. the SER system without optimization, and also the existing literature works. The proposed SER system is language-dependent and it can be further improved to be language-independent with cross-corpus analysis.

Acknowledgment

The authors would like to acknowledge the Ministry of Electronics and Information Technology (MeitY), Government of India, for the financial support rendered for this research work through the Visvesvaraya PhD Scheme for Electronics and IT.

References

[1] El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 2011; 44 (3): 572-587.

[2] Ververidis D, Kotropoulos C. Emotional speech recognition: resources, features, and methods. Speech Communication 2006; 48 (9): 1162-1181.

[3] Seng KP, Ang LM, Ooi CS. A combined rule-based & machine learning audio-visual emotion recognition approach. IEEE Transitions on Affective Computing 2018; 9 (1): 3-13.

[4] Busso C, Lee S, Narayanan S. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Transactions on Audio, Speech, and Language Processing 2009; 17 (4): 582-596.

[5] Yang B, Lugger M, Emotion recognition from speech signals using new harmony features. Signal Processing 2010; 90 (5): 1415-1423. doi: 10.1016/j.sigpro.2009.09.009

[6] Lugger M, Yang B. The relevance of voice quality features in speaker independent emotion recognition. In: ICASSP 2007; Honolulu, HI, USA; 2007. pp. IV18-20.

[7] Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden Markov models. Speech Communication 2003; 41 (4): 603-623.

[8] Nordstrom KI, Tzanetakis G, Driessen PF. Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction. IEEE Transactions on Audio, Speech, and Language Processing 2008; 16 (6): 1087-1096.

[9] Kim MJ, Yoo J, Kim Y, Kim H. Speech emotion classification using tree structured sparse logistic regression. In: INTERSPEECH 2015; Dresden, Germany; 2015. pp. 1541-1545.

[10] Ying S, Xue-Ying Z. Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition. Future Generation Computer Systems 2018; 81: 291-296.

[11] Neiberg D, Elenius K, Laskowski K. Emotion recognition in spontaneous speech using GMMs. In: INTERSPEECH 2006; Pittsburgh, PA, USA; 2006. pp. 809-812.

[12] Hu H, Xu MX, Wu W. GMM supervector based SVM with spectral features for speech emotion recognition. In: ICASSP 2007; Honolulu, HI, USA; 2007. pp. IV 413-416.

[13] Wang Y, Hu W. Speech emotion recognition based on improved MFCC. In: Proceedings of the 2nd International Conference on Computer Science and Application Engineering; New York, NY, USA; 2018. doi: 10.1145/3207677.3278037

[14] Teager HM. Some observations on oral air flow during phonation. IEEE Transactions on Acoustics, Speech, and Signal Processing 1980; 28 (5): 599-601.

[15] Kaiser JF. Some useful properties of Teager's energy operators. In: ICASSP 1993; Minneapolis, MN, USA, USA; 1993. pp. 149-152.

[16] Zhou G, Hansen JHL, Kaiser JF. Classification of speech under stress based on features derived from the nonlinear Teager energy operator. In: ICASSP 1998; Seattle, WA, USA; 1998. pp. 549-552.

[17] Cairns DA, Hansen JHL. Nonlinear analysis and classification of speech under stressed conditions. Journal of the Acoustical Society of America 1994; 96 (6): 3392-3400.

[18] Zhou G, Hansen JHL, Kaiser JF. Nonlinear feature based classification of speech under stress. IEEE Transactions on Speech and Audio Processing 2001; 9 (3): 201-216.

[19] Sun R, Moore E. Investigating glottal parameters and teager energy operators in emotion recognition. Lecture Notes in Computer Science 2011; 6975: 425-434.

[20] Wu K, Zhang D, Lu G. GMAT: Glottal closure instants detection based on the multiresolution absolute Teager–Kaiser energy operator. Digital Signal Processing 2017; 69: 286-299.

[21] Attabi Y, Alam MJ, Dumouchel P, Kenny P, O'Shaughnessy D. Multiple windowed spectral features for emotion recognition. In: ICASSP 2013; Vancouver, Canada; 2013. pp. 7527-7531.

[22] Bou-Ghazale SE, Hansen JHL. A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Transactions on Speech and Audio Processing 2000; 8 (4): 429-442.

[23] Liu ZT, Xie Q, Wu X, Cao W, Mei Y et al. Speech emotion recognition based on an improved brain emotion learning model. Neurocomputing 2018; 309: 145-156.

[24] Pohjalainen J, Saeidi R, Kinnunen T, Alku P. Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions. In: INTERSPEECH 2010; Makuhari, Japan; 2010. pp. 1477-1480.

[25] Wang K, An N, Li BN, Zhang Y, Li L. Speech emotion recognition using Fourier parameters. IEEE Transactions on Speech and Audio Processing 2015; 6 (1): 69-75.

[26] Zao L, Cavalcante D, Coelho R. Time-frequency feature and AMS-GMM mask for acoustic emotion classification. IEEE Signal Processing Letters 2014; 21 (5): 620-624.

[27] Wu S, Falk TH, Chan WY. Automatic speech emotion recognition using modulation spectral features. Speech Communication 2011; 53 (5): 768-785.

[28] Deb S, Dandapat S. Multiscale amplitude feature and significance of enhanced vocal tract information for emotion classification. IEEE Transactions on Cybernetics 2018; 49 (3): 802-815.

[29] Caponetti L, Buscicchio C, Castellano G. Biologically inspired emotion recognition from speech. EURASIP Journal on Advances in Signal Processing 2011; 24: 1-6.

[30] Domingos P. A few useful things to know about machine learning. Communications of the ACM 2012; 55 (10): 78-87.

[31] Mitchell TM. Machine Learning. New York, NY, USA: McGraw-Hill Education, 1997.

[32] Lee CM, Narayanan SS. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing 2005; 13 (2): 293-303.

[33] Wu D, Parsons TD, Narayanan SS. Acoustic feature analysis in speech emotion primitives estimation. In: INTERSPEECH 2010; Makuhari, Japan; 2010. pp. 785-788.

[34] Bartenhagen C, Klein HU, Ruckert C, Jiang X, Dugas M. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. BMC Bioinformatics 2010; 11: 567.

[35] Wall M, Rechtsteiner A, Rocha L. Singular value decomposition and principal component analysis. In: Berrar DP, Dubitzky W, Granzow M (editors). A Practical Approach to Microarray Data Analysis. Berlin, German: Springer, 2003, pp. 91-109.

[36] Zhang S, Li L, Zhao Z. Spoken emotion recognition using kernel discriminant locally linear embedding. Electronics Letters 2010; 46 (19): 1344-1346.

[37] Song P, Ou S, Zheng W, Jin Y, Zhao L. Speech emotion recognition using transfer non-negative matrix factorization. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing; Shanghai, China; 2016. pp. 5180-5184.

[38] Sahu S, Gupta R, Sivaraman G, Abd Almageed W, Espy-Wilson C. Adversarial Auto-encoders for speech based emotion recognition. In: INTERSPEECH 2017; Stockholm, Sweden; 2017. pp. 1243-1247.

[39] Latif S, Rana R, Qadir J, Epps J. Variational autoencoders to learn latent representations of speech emotion. In: Interspeech 2018; Hyderabad, India; 2018. pp. 3107-3111.

[40] Koolagudi SG, Rao KS. Emotion recognition from speech using source, system, and prosodic features. International Journal of Speech Technology 2012; 15 (2): 265-289.

[41] Rao KS, Reddy VR, Maity S. Language Identification Using Spectral and Prosodic Features. Cham, Switzerland: Springer International Publishing, 2015.

[42] Ding C, Li T, Jordan MI. Convex and semi-nonnegative matrix factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence 2010; 32 (1): 45-55.

[43] Gillis N, Kumar A. Exact and heuristic algorithms for semi-nonnegative matrix factorization. SIAM Journal of Matrix Analytics and Application 2014; 36 (4): 1404-1424.

[44] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: Interspeech 2005; Lisbon, Portugal; 2005. pp. 1-4.

[45] Busso C, Bulut M, Lee C, Kazemzadeh A, Mower E et al. IEMOCAP: Interactive emotional dyadic motion capture database. Journal of Language Resources and Evaluation 2008; 42 (4): 335-359.

[46] Chen S, Wang J, Hsieh W, Chin Y, Ho C et al. Speech emotion classification using multiple kernel Gaussian process. In: APSIPA 2016; Jeju, South Korea; 2016. pp. 1-4.

[47] Zhang S, Zhao X, Lei B. Speech emotion recognition using an enhanced kernel isomap for human-robot interaction. International Journal of Advanced Robotic Systems 2017; 10 (2): 1-7.

[48] Zhang S, Zhao X. Dimensionality reduction-based spoken emotion recognition. Multimedia Tools and Applications 2013; 63 (3): 615–646.

[49] Yan J, Wang X, Gu W, Ma L. Speech emotion recognition based on sparse representation. Archives of Acoustics 2013; 38 (4): 465–470.

[50] Kuchibhotla S, Vankayalapati HD, Anne KR. An optimal two stage feature selection for speech emotion recognition using acoustic features. International Journal of Speech Technology 2016; 19 (4): 657-667.

[51] Daneshfar F, Kabudian SJ. Speech emotion recognition using optimal dimension reduction and non-isotropic Gaussian radial basis function network trained with scaled conjugate gradient descent. Iranian Journal of Electrical and Electronic Engineering of Iran University of Science and Technology (in press).

[52] Gudmalwar AP, Rama Rao CV, Dutta A. Improving the performance of the speaker emotion recognition based on low dimension prosody features vector. International Journal of Speech Technology (in press).

[53] Özseven T. A novel feature selection method for speech emotion recognition. Applied Acoustics 2019; 146: 320–326.

[54] Sun L, Fu S, Wang F. Decision tree SVM model with Fisher feature selection for speech emotion recognition. EURASIP Journal on Audio, Speech, and Music Processing 2019; 2019: 2. doi: 10.1186/s13636-018-0145-5