Speech Emotion Recognition Using Unsupervised Feature Selection Algorithms

Surekha Reddy BANDELA, T. Kishore KUMAR

Dept. of ECE, National Institute of Technology Warangal, India

{suri.reddyband, kishorefr}@gmail.com

Submitted December 28, 2019 / Accepted April 17, 2020

Abstract. *The use of the combination of different speech features is a common practice to improve the accuracy of Speech Emotion Recognition (SER). Sometimes, this leads to an abrupt increase in the processing time and some of these features contribute less to emotion recognition often resulting in an incorrect prediction of emotion due to which the accuracy of the SER system decreases substantially. Hence, there is a need to select the appropriate feature set that can contribute significantly to emotion recognition. This paper presents the use of Feature Selection with Adaptive Structure Learning (FSASL) and Unsupervised Feature Selection with Ordinal Locality (UFSOL) algorithms for feature dimension reduction to improve SER performance with reduced feature dimension. A novel Subset Feature Selection (SuFS) algorithm is proposed to reduce further the feature dimension and achieve a comparable better accuracy when used along with the FSASL and UFSOL algorithms. 1582 INTERSPEECH 2010 Paralinguistic, 20 Gammatone Cepstral Coefficients and Support Vector Machine classifier with 10-Fold Cross-Validation and Hold-Out Validation are considered in this work. The EMO-DB and IEMOCAP databases are used to evaluate the performance of the proposed SER system in terms of classification accuracy and computational time. From the result analysis, it is evident that the proposed SER system outperforms the existing ones.*

Keywords

Speech Emotion Recognition (SER), INTERSPEECH Paralinguistic Feature Set, GTCC, feature selection, feature optimization, FSASL, UFSOL, SuFS

1. Introduction

Speech Emotion Recognition (SER) is the method of detecting the emotional state of a speaker using a speech signal. The field of emotion recognition has gained a lot of interest in human-computer interaction these days, and intensive research is going on in this field using various feature extraction techniques and machine learning algorithms. SER is used in the applications viz., call-center services, in vehicles, as a diagnosing tool in medical services, story-telling and in E-tutoring applications etc.

There are six archetypal emotions: anger, neutral, happiness, disgust, surprise, fear and sadness. In situations where only a person's speech signals are available, SER plays a prominent role [1], [2]. Speech features can be classified as Continuous, Voice Quality, Spectral and Nonlinear Teager Energy Operator (TEO) features. Figure 1 shows the categorical representation of some of these speech features. A significant challenge in SER is the identification of useful speech features that holds the emotional characteristics from a speech signal, and most of the research related to SER is focused on identifying the effective feature set. It is evident from the literature that the feature fusion increased the classification accuracy of the SER system and became the most common practice.

Even though the classification accuracy of the SER system increases due to feature fusion, 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 for emotion recognition. The feature selection methods simplify the task of interpretation by the classification algorithms easier. These techniques majorly eradicate the loss caused due to the curse of dimensionality and also the problem of overfitting by improving the generalization in the model, i.e., the use of less redundant data that leads to incorrect predictions increasing the accuracy of the SER system and thus, enhancing the prediction performance by decreasing the computational time and memory by the SER system. Hence, feature dimension reduction is the best way to enhance the accuracy of the SER system. The reduction of the number of features causes an uncertain loss of information and subsequently leads to instability in the performance of the SER system. To overcome this drawback and to acquire the most optimal feature sets that improve SER accuracy, many feature selection techniques are developed in machine learning.

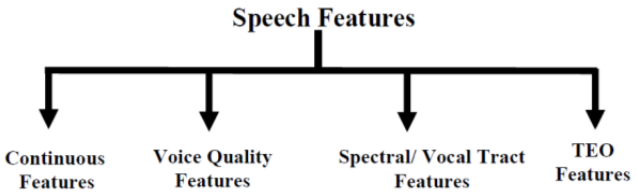

Fig. 1. Categorization of speech features.

In feature selection, from the original feature set, a subset of features is selected with respect to their relevance and redundancy. It improves the prediction performance and reduces the computational complexity and storage, providing faster and cost effective models [3]. In machine learning, a feature vector is the *n*-dimensional vector representing the features of all samples. The space-related to these vectors is the feature space. To decrease the dimensionality of feature space, the feature selection or feature transformation methods can be used. In feature transformation, the original feature space is transformed into a different space having a distinct set of axes to reduce the dimensionality of the data.

In contrast, feature selection reduces into a subspace from the original feature space without transformation. Some examples of feature selection methods are ReliefF, Fisher Score, Information Gain, Chi Squares, LASSO, etc. Feature selection techniques can be categorized based on labelling of the data as supervised, unsupervised and semi-supervised. In supervised feature selection, the data is labelled feature evaluation process. If the data is huge, labelling of the data is costly and a tedious task. Unsupervised feature selection can overcome these drawbacks of supervised approaches. But this is more difficult than supervised ones since it does not have labelled data and still its result can be good even without any prior knowledge. The evaluation of feature selection methods can be further classified into four types, i.e., filter, wrapper, embedded, hybrid and ensemble feature selection, as shown in Fig. 2.

Filter feature selection techniques use the statistical analysis to assign a distinctive feature with a score. Their score ranks the features, and later, these are retained or removed from the original feature vector set accordingly. These filter techniques mostly use a single variable in their analysis and features are considered independent of each other or dependent terms. The most commonly used filter methods are the Chi-squared test [4], variance threshold [5], information gain, etc. The fast feature selection method, i.e., Fisher feature selection is used in [6] with decision SVM for SER. The wrapper feature selection techniques consider a set of features with various combinations of the feature sub-sets. Later, these feature subsets are compared with one another as a search problem which is estimated and compared with other groups. Further, the prediction process is performed to assign the score onto each of the feature sets depending on the prediction accuracy. The process of search can be systematical such as searching the first best features, for example, hill-climbing algorithm or using heuristics. The search process can be systematic, stochastic or heuristic such as a best-first search, random hill-climbing algorithm, forward and backward passes to add and remove features. Genetic Algorithms, Recursive Feature Elimination (RFE), Sequential Feature Selection (SFS), etc. are some of the wrapper methods of feature selection. In [7], SFS and Sequential Floating Feature Selection (SFFS) are used for SER.

Embedded methods, while creating the model in the learning process, select the best features that can be used to

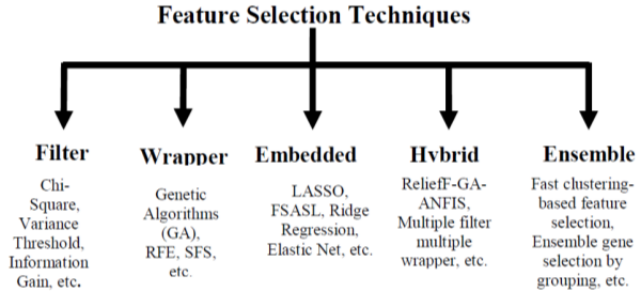

Fig. 2. Categorization of feature selection techniques.

enhance accuracy. Regularization techniques are the most commonly used embedded methods for feature selection: LASSO, FSASL, Ridge Regression, Elastic Net, etc. In [8], an L1-Norm with multiple kernel learning and embedded feature selection method are used for SER. The hybrid method is a combination of two or more feature selection methods (e.g., filter + wrapper). These methods try to acquire the benefits of both techniques by combining their corresponding strengths. It achieves improved efficiency, prediction performance and decreases computational complexity. The most widely used hybrid method is the combined feature selection with filter and wrapper approaches.

Ensemble method constructs a collection of feature subgroups and produces an aggregate result from the group. The primary goal of this method is to tackle the unpredictability problems in most of the feature selection algorithms. This method is based on various subsampling schemes in which one feature selection technique runs on many subsamples, and the resultant features are combined to attain a subset with more stability. With this, for high dimensional data, the feature selection performance is no longer dependent on any individual selected subset, thus attains more flexibility and robustness.

In [9], a sparse representation based sparse partial least squares regression (SPLSR) feature selection method is used for SER. Apart from these, feature selection techniques, feature transformation methods can also be used for feature dimension reduction in SER [10–12]. In [10], semi-NMF feature transformation technique with multiple kernel Gaussian process is used for feature dimension reduction. In [11], a supervised feature transformation based dimension reduction method i.e., modified supervised locally linear embedding (MSLLE) algorithm is adopted for SER. In [12], principal component analysis (PCA) is used for SER to transform the high dimensional feature space to a lower dimension.

In [13], unsupervised feature learning is carried out using k-means clustering, sparse Auto-Encoders (AE) and sparse restricted Boltzmann machines for feature mapping to obtain optimal feature set for SER. The adversarial AEs and variational AEs have the ability to encode the high dimensional feature vector to a lower dimension and also have the ability to reconstruct the original feature space. Therefore, in [14], [15], these are used as feature dimension reduction techniques for SER. In [16], a new variant of feature extraction technique i.e., deep neural network based heterogeneous model consisting of AE, denoising AE and

an improved shared hidden layer AE is used to extract the features from speech signal. These layers also provide feature optimization up to some extent. But to obtain better performance for SER with the high-dimension feature set, a fusion level network with support vector machine (SVM) classifier is used.

In this paper, an SER system is proposed with unsupervised feature selection algorithms with the Support Vector Machine (SVM) classifier using Linear and Radial Basis Function (RBF) kernels. The significant contributions of this work are:

i) Using the UFSOL and FSASL unsupervised feature selection algorithms for feature selection which have not yet been explored for SER.
ii) To propose a Subset Feature Selection (SuFS) algorithm to improve the performance of the proposed SER system further by selecting the subset of features after UFSOL and FSASL feature selection, based on the 10-fold validation accuracy obtained by using UFSOL and FSASL algorithms, as the decisive factor for feature selection.

The rest of the paper is structured as follows: Section 2 describes the proposed SER system with UFSOL, FSASL algorithms along with a novel Subset Feature Selection (SuFS) algorithm and Section 3 discusses the performance analysis of the proposed SER system followed by Section 4 with the conclusion and future scope of the proposed work.

2. Proposed Speech Emotion Recognition System using Unsupervised Feature Selection Algorithms

In the proposed SER system, after the feature extraction, the unsupervised feature selection algorithms, i.e., UFSOL and FSASL are used individually to select the most prominent from the original feature set as shown in Fig. 3.

2.1 Database

In the proposed work, EMO-DB and IEMOCAP datasets are considered for the SER analysis. EMO-DB, the German database [18] is widely used in SER analysis by many of the researchers. The recording for emotional data was done in an anechoic chamber by five male and five female actors between the age group of 25–35. Totally 535 speech signals were recorded at 48 kHz with Anger, Boredom, Disgust, Anxiety/Fear, Happiness, Sad and Neutral. Later these are down-sampled to 16 kHz. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [19] is an acted, multimodal and multi-speaker database. Twelve hours of audio-visual data that include video, speech, text transcriptions and motion capture of the face. In this work, the speech data with emotions, anger, happiness, neutral and sadness are considered as in most of the SER works, with a total of 4490 utterances.

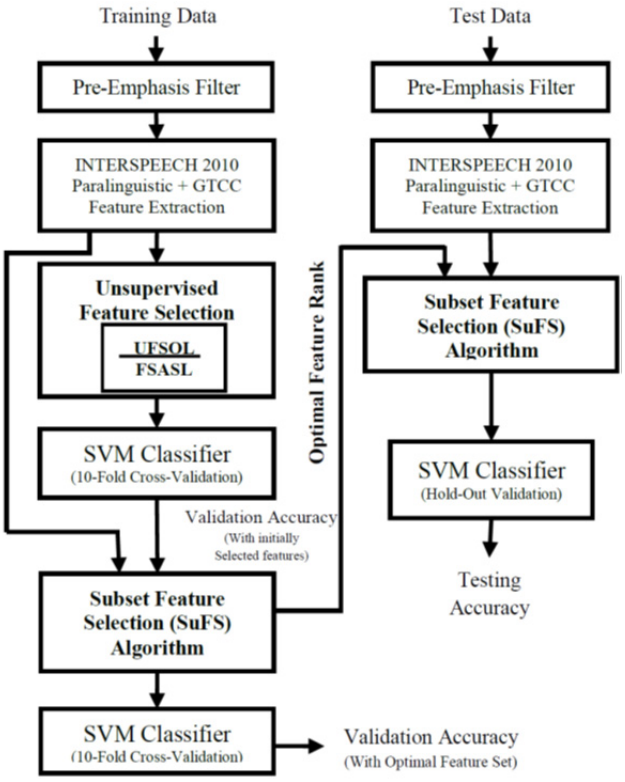

Fig. 3. Proposed SER system using unsupervised feature selection.

2.2 Pre-Processing

The speech signal is initially passed through a pre-emphasis filter to boost the energy in their higher frequencies which are attenuated during the speech signal production from vocal tract [20].

2.3 Feature Extraction

Feature Extraction in speech emotion recognition is the process of extracting the speech specific features that have the emotion relevant information [1]. In order to obtain the emotional contents from a speech signal, a particular set of features can be extracted by applying various signal processing techniques. In this work, INTERSPEECH 2010 paralinguistic challenge feature set and Gammatone Cepstral Coefficients (GTCC) are used as features. The INTERSPEECH 2010 paralinguistic challenge set consists of 1582 features with a four-set of features combined [21].

Descriptors	Functionals
PCM Loudness	Position – max./ min.
MFCC [0-14]	Arithmetic mean, Standard Deviation
Log Mel Freq. Band [0-7]	Skewness, Kurtosis
LSP Frequency [0-7]	Linear regression coefficient
F0 by Sub-Harmonic Sum.	Linear regression error
F0 Envelope	Quartile
Voicing Probability	Quartile range
Jitter Local	Percentile
Jitter DDP	Percentile range
Shimmer Local	Up-level time

Tab. 1. INTERSPEECH 2010 paralinguistic feature set.

The Munich open Speech and Music Interpretation by Large Space Extraction (openSMILE) toolkit [22] is utilized to extract the 1582 features for the individual speech signal. The configuration file 'IS10paraling:conf" is used to obtain these features and the features, along with the description are as shown in Tab. 1.

The gammatone filter takes its name from the impulse response, which is the product of a Gamma distribution function and a sinusoidal tone centered at the frequency, being computed as [23]:

$$g(t) = Kt^{(n-1)}\, \mathrm{e}^{-2\pi Bt} \cos\left(2\pi f_{\mathrm{c}} t + \phi\right) \tag{1}$$

where $g(t)$ is the impulse response of gammatone filter; K is the amplitude factor; n is the filter order; f_{c} is the central frequency in Hz; ϕ is the phase shift; B is the duration of impulse response ($B = 1.019 \times ERB(f_{\mathrm{c}})$).

ERB is the equivalent rectangular bandwidth i.e., $ERB(f) = 24.7 + 0.108f$. The center frequency f_{c} of each gammatone filter is equally spaced on *ERB* scale, i.e., $f_{\mathrm{c}} = ERBS^{-1}\left(ERBS(f_{\mathrm{low}}) + \dfrac{ERBS(f_{\mathrm{high}} - f_{\mathrm{low}})}{N}\right)$, where $ERBS(f) = 21.4 \log_{10}(1 + 0.00437 f)$.

The fourth order gammatone filter is similar to human auditory model, therefore $n = 4$. Here, $f_{\mathrm{low}} = 62.5$ Hz, $f_{\mathrm{high}} = 3400$ Hz and N is the number of gammatone filters i.e., 20. After obtaining the gammatone filter coefficients the cepstral analysis is applied to these, obtaining a total of 20 gammatone cepstral coefficients using the gammatone filter.

2.4 Unsupervised Feature Selection

The unsupervised feature selection algorithms, i.e., UFSOL and FSASL, which are not yet explored for SER so far, are used in this work. Apart from this, a novel Subset Feature Selection algorithm is modelled by the results obtained after using UFSOL and FSASL algorithms to improve the performance of the SER system further. The entire set of 1602 features is given to the feature selection algorithms to select the most prominent features, as shown in Fig. 3. The UFSOL and FSASL algorithms are discussed as below:

2.4.1 Unsupervised Feature Selection with Ordinal Locality (UFSOL):

Consider $\mathbf{X} = [\mathbf{x}_1,\ldots,\mathbf{x}_d] \in \mathbb{R}^{m \times d}$ as the initial feature matrix with d speech signals and m number of features. Generally the regularized regression, feature selection is formulated as [24]:

$$\min_W \mathbf{W}^{\mathrm{T}}\mathbf{X} - \mathbf{H}_{\mathrm{F}}^2 + \delta \mathbf{W}_{2,q} \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{m \times d_2}$ ($m < d_2$) is a projection matrix/ feature selection matrix; $l_{2,q}$-norm (q is typically set to 0 or 1) assures the sparseness in rows of $\mathbf{W}$; $\mathbf{H} = [h_1,\ldots,h_d] \in \mathbb{R}^{d_2 \times d}$ is a target matrix in this unsupervised feature selection algorithm.

Whereas, $\mathbf{H}$ is a label matrix in case of supervised multi-class data. In this work, the bi-orthogonal semi Non-negative Matrix Factorization (NMF) is used to decompose $\mathbf{H}$ into two new matrices i.e., $\mathbf{H} \cong \mathbf{UV}$ with $\mathbf{V} \geq 0$, $\mathbf{VV}^{\mathrm{T}} = I$ and $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$.

If the feature set selected for original sample x_i is supposed to be $\mathbf{y}_i = \mathbf{W}^{\mathrm{T}}\mathbf{x}_i$, then $\mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{X}$. According to the principle of "*ordinal locality preserving*", given a triplet ($\mathbf{x}_i$, $\mathbf{x}_u$, $\mathbf{x}_v$) comprised of x_i and its neighbors $\mathbf{x}_u$ and $\mathbf{x}_v$, their corresponding feature groups also form a triplet ($\mathbf{y}_i$, $\mathbf{y}_u$, $\mathbf{y}_v$). Let the distance metric be denoted by dist(.,.). The feature selection holds *ordinal locality preserving* if the following condition is preserved: i.e., if $\mathrm{dist}(\mathbf{x}_i,\mathbf{x}_u) \leq \mathrm{dist}(\mathbf{x}_i,\mathbf{x}_v)$ then $\mathrm{dist}(\mathbf{y}_i,\mathbf{y}_u) \leq \mathrm{dist}(\mathbf{y}_i,\mathbf{y}_v)$.

Based on this, the appropriate feature group for each data point is identical to optimizing the following ordinal locality preserving loss function over a collection of triplets as below:

$$\max_{\mathbf{Y}} \sum_{i=1}^{d} \sum_{u \in \mathbf{N}_i} \sum_{v \in \mathbf{N}_i} \mathbf{A}_{uv}^{i} \left[\mathrm{dist}(\mathbf{y}_i,\mathbf{y}_u) - \mathrm{dist}(\mathbf{y}_i,\mathbf{y}_v)\right] \tag{3}$$

where $\mathbf{N}_i$ is a set of sequence numbers indicating the k nearest neighbors of $\mathbf{x}_i$; $\mathbf{A}^i$ denotes an antisymmetric matrix with $(u,v)^{\mathrm{th}}$ element, the $\mathrm{dist}(\mathbf{x}_i,\mathbf{x}_u) - \mathrm{dist}(\mathbf{x}_i,\mathbf{x}_v)$. If the weighting matrix is denoted as $\mathbf{C} \in \mathbb{R}^{m \times d}$ then

$$\mathbf{C}_{i,j} = \begin{cases} \sum_{u \in \mathbf{N}_i} \mathbf{A}_{uj}^{i} & , j \in \mathbf{N}_i \\ 0 & , j \notin \mathbf{N}_i \end{cases} \tag{4}$$

From (4), equation (3) is equivalent to

$$\min_{\mathbf{Y}} \sum_{i=1}^{d} \sum_{j=1}^{d} \mathbf{C}_{ij}\, \mathrm{dist}(y_i,\, y_j). \tag{5}$$

The squared Euclidean distance is used to establish each pairwise distance. The loss function of ordinal locality preserving can be written accordingly as $\min_{\mathbf{Y}} \sum_{i=1}^{d} \sum_{j=1}^{d} \mathbf{C}_{ij} \left\| \boldsymbol{y}_i - \boldsymbol{y}_j \right\|_2^2$, which has an equivalent compact matrix form: $\min_{\mathbf{Y}} \mathrm{Tr}\left(\mathbf{YLY}^{\mathrm{T}}\right)$ as well as $\min_{\mathbf{W}} \mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{XLX}^{\mathrm{T}}\mathbf{W}\right)$ by substituting $\mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{X}$. From these considerations, (2) can be formulated as

$$\min_{\mathbf{W,U,V}} \boldsymbol{F} = \left\{ \left\| \mathbf{W}^{\mathrm{T}}\mathbf{X} - \mathbf{UV} \right\|_F^2 + \delta \left\| \mathbf{W} \right\|_{2,1} + \rho\, \mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{XLX}^{\mathrm{T}}\mathbf{W}\right) \right\},$$
$$\text{s.t. } \mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathbf{I},\ \mathbf{V} \geq 0,\ \mathbf{VV}^{\mathrm{T}} = \boldsymbol{I} \tag{6}$$

where δ and ρ are scalar constants that controls the relativeness of corresponding terms.

According to half-quadratic theory, for a fixed t, there

is a conjugate function $\psi(.)$, with $\sqrt{t^2+\varepsilon} = \inf_{r\in R}\left\{\frac{r}{2}t^2 + \psi(r)\right\}$. The infimum could be reached at $r = 1/\sqrt{t^2+\varepsilon}$. With this, (4) can be optimized by minimizing its augmented function $\hat{F}$ as below:

$$\min_{\boldsymbol{W},\boldsymbol{U},\boldsymbol{V},\boldsymbol{R}} \hat{\mathrm{F}} = \left\{\begin{array}{l}\left\|\mathbf{W}^{\mathrm{T}}\mathbf{X}-\mathbf{H}\right\|_{\boldsymbol{F}}^2 + \delta\sum_{i=1}^{m}\left\{\frac{\mathbf{R}_{ii}}{2}\left\|\mathbf{W}_i\right\|_2^2 + \boldsymbol{\psi}_i\left(\mathbf{R}_{ii}\right)\right\}\\ +\rho\,\mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathrm{T}}\mathbf{W}\right)\end{array}\right\},$$
$$\text{s. t. } \mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}, \mathbf{V}\geq 0,\ \mathbf{V}\mathbf{V}^{\mathrm{T}}=\mathbf{I} \tag{7}$$

where $\mathbf{R}$ is a $m \times d_2$ diagonal matrix storing the auxiliary variables and $\{\psi_i\}_{i=1}^m$ are conjugate functions. i.e.,

$$\min_{\mathbf{W},\mathbf{U},\mathbf{V}} \mathrm{F}(\mathbf{W},\mathbf{U},\mathbf{V}) = \min_{\mathbf{W},\mathbf{U},\mathbf{V},\mathbf{R}} \hat{\mathrm{F}}(\mathbf{W},\mathbf{U},\mathbf{V},\mathbf{R}). \tag{8}$$

The minimization of $\hat{\mathrm{F}}$(**W**,**U**,**V**,**R**) is as shown below:

i) The diagonal elements of **R** are updated in parallel:

$$\mathbf{R}_{ii} = 1/\sqrt{\left\|\mathbf{W}_i\right\|_2^2 + \varepsilon} \tag{9}$$

The algorithm to solve (7) is as below:

Algorithm 1: The algorithm to solve (7)

Input: Data matrix $\mathbf{X} = [x_1,\ldots,x_d] \in \mathbb{R}^{m\times d}$; Number of each sample's nearest neighbors k; Parameters d_2, c, δ, and ρ.

Solution:

1: Compute **C** via (6) and its corresponding Laplacian matrix **L**;
2: Initialize $\mathbf{W}^{(0)}$ with d_2 different columns randomly selected $d_1 \times d_1$ identity matrix, $t = 0$;
3: **while** not convergence **do**
4: $t \leftarrow t + 1$;
5: Update $\mathbf{R}^{(t)}$ via (9);
6: Update $\mathbf{U}^{(t)}$ and $\mathbf{V}^{(t)}$ by K-means;
7: Update $\mathbf{W}^{(t)}$ by Eigen decomposition;
8: **end while**

Output:
W→Feature Selection matrix; **V**→ cluster indicator matrix.

ii) To solve (7), (**U, V**) is updated for fixed **W** by applying orthogonal Semi-NMF on projected data i.e., feature selection matrix $\mathbf{Y} = \mathbf{W}^{\mathrm{T}}\mathbf{X}$. The orthogonal semi-NMF problem, $\min_{\mathbf{U},\mathbf{V}}\left\|\mathbf{Y}-\mathbf{U}\mathbf{V}\right\|_F^2$, s.t. $\mathbf{V}\geq 0, \mathbf{V}\mathbf{V}^{\mathrm{T}}$ is equivalent to relaxed K-means clustering. The zero gradient condition $\mathbf{U} = \mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{V}^{\mathrm{T}}$ is attained by updating (**U**, **V**) using k-means clustering.

iii) **W** is updated with (**U**, **V**) fixed, substitute $\mathbf{U} = \mathbf{W}^{\mathrm{T}}\mathbf{X}\mathbf{V}^{\mathrm{T}}$ in f(**W**,**U**,**V**) and the objective function $\min_{\mathbf{W}^{\mathrm{T}}\mathbf{W}=\mathbf{I}} \mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{G}\mathbf{W}\right)$ is solved by applying Eigen decomposition on $\mathbf{G} = \frac{\delta}{2}\mathbf{R} + \mathbf{X}\left(\rho\mathbf{L} + \mathbf{I} - \mathbf{V}^{\mathrm{T}}\mathbf{V}\right)\mathbf{X}^{\mathrm{T}}$.

The optimal **W** comprises d_2 Eigen vectors corresponding to the smallest Eigen values of d_2.

All the above steps are updated until convergence as summarized in Algorithm 1. **W** is the resultant feature selection matrix.

2.4.2 Feature Selection with Adaptive Structure Learning (FSASL):

In this algorithm, consider the feature set as $\mathbf{X} \in \mathbb{R}^{d\times m}$, where d is the dimension of the speech files and m is the number of features. The parameters α, β, γ, μ are considered as the regularization parameters used to balance sparsity and reconstruction error of global as well as local structure learning. Also, considering the optimized data dimension as c, the resultant optimized feature set $\in \mathbb{R}^{d\times c}$. Using the general equation that guides the FSASL method [25]:

$$\begin{aligned}&\min_{\mathbf{Z},\mathbf{S},\mathbf{P}}\left(\left\|\mathbf{Z}^{\mathrm{T}}\mathbf{X}-\mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{S}\right\|^2 + \alpha\left\|\mathbf{S}\right\|_1\right) +\\ &\beta\sum_{q,r}^{m}\left(\left\|\mathbf{Z}^{\mathrm{T}}x_q - \mathbf{Z}^{\mathrm{T}}x_r\right\|^2\mathbf{P}_{qr} + \mu\mathbf{P}_{qr}{}^2\right) + \gamma\left\|\mathbf{Z}\right\|_{21},\end{aligned} \tag{10}$$

subject to $\mathbf{S}_{qr} = 0$, $\bar{\mathbf{P}}\mathbf{1}_m = \mathbf{1}_m$, $\bar{\mathbf{P}} \geq 0$, $\mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Z} = \bar{\mathbf{I}}$; where, **X** = Input Feature set; x = a particular row of data matrix;

Algorithm 2: FSASL Algorithm

Input: Input feature set as $\mathbf{X} \in \mathbb{R}^{m\times d}$; d is the dimension of the speech files; m is the number of features.

Solution:

For each data sample x_q, all the data points $\{x_r\}^m{}_{r=1}$ are considered as the neighborhood of x_q with probability $P(q,r)$.

S = Weight matrix of the data matrix;
s = a particular row of the Weight matrix;
Z = feature selection and transformation matrix.

The optimization problem in (10) derives different variables (**S**, **P** and **Z**(t)) into a set of sub-problems with only single variable involved and is solved as follows:

1) Solving for **S** by keeping **P** and **Z** as constant. For each q, update the q^{th} column of **S** by solving the problem:

$$\min_{s_q}\left(\left\|x_q^{'} - X^q s_q\right\|^2 + \alpha\left|s_q\right|\right), \text{ s.t. } \mathbf{S}_{qq} = 0 \tag{11}$$

where **X'** and **x'** are the transpose matrices of **X** and **x**.

2) Solving for **P** by keeping **S** and **Z** as constant. For each q, update the q^{th} column of **P** by solving the problem

$$\min_{\mathbf{W},\mathbf{S},\mathbf{P}}\sum_{q,r}^{m}\left(\left\|x_q^{'} - x_r^{'}\right\|^2\mathbf{P}_{qr} + \mu\mathbf{P}_{qr}{}^2\right) \tag{12}$$

s.t. $\mathbf{1}^{\mathrm{T}}{}_m p_q = 1$, $\mathbf{P}_{qr} \geq 0$. Denote $\mathbf{A} \in \mathbb{R}^{m \times m}$ be a square matrix with $\mathbf{A}_{qr} = -\frac{1}{2\mu}\left\| x_q^{'} - x_r^{'} \right\|^2$, then the above problem can be written as:

$$\min_{\mathbf{p}_q^{\mathrm{T}}} \frac{1}{2}\left\| \mathbf{p}_q^{\mathrm{T}} - \mathbf{a}_q^{\mathrm{T}} \right\|^2, \text{ s.t. } \mathbf{p}^{\mathrm{T}}{}_q \mathbf{1}_m = 1,\ 0 \leq \mathbf{p}^{\mathrm{T}}{}_{qr} \leq 1 \qquad (13)$$

where $p'(t)$ is the q^{th} row of $\mathbf{P}$.

3) Compute the overall graph laplacian by $\mathbf{L} = \mathbf{L}_{\mathrm{S}} + \beta(\mathbf{L}_{\mathrm{S}})$, then

$$\mathbf{L}_{\mathrm{P}} = \mathbf{D}_{\mathrm{P}} - \left(\mathbf{P} + \mathbf{P}^{\mathrm{T}}\right)/2 \qquad (14)$$

where, $\mathbf{D}_{\mathrm{P}}$ is a diagonal matrix whose i^{th} diagonal element is $\sum_r \left(\mathbf{P}_{qr} + \mathbf{P}_{rq}\right)/2$

$$\mathbf{L}_{\mathrm{S}} = \left(\mathbf{I} - \mathbf{S}\right)\left(\mathbf{I} - \mathbf{S}\right)^{\mathrm{T}}. \qquad (15)$$

4) Now computing the feature selection or transformation matrix $\mathbf{Z}$ by keeping $\mathbf{P}$ and $\mathbf{S}$ as constant and using the equation below:

$$\min_{\mathbf{Z}} \ \mathrm{Tr}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{L}\mathbf{X}^{\mathrm{T}}\mathbf{Z}^{\mathrm{T}}\right) + \gamma\left\|\mathbf{Z}\right\|_{21}, \text{ s.t. } \mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Z} = \mathbf{I}. \ (16)$$

Given the i^{th} estimation $\mathbf{Z}^t$ and $\mathbf{D}_{\mathbf{Z}^t}$ denoting the diagonal matrix with the i^{th} diagonal element as $1/\left(2\left\|z_q^t\right\|^2\right)$, (16) can be rewritten as:

$$\min_{\mathbf{W}} \ \mathrm{Tr}\left(\mathbf{Z}^{\mathrm{T}}\mathbf{X}(\mathbf{L} + \gamma\mathbf{D}_{\mathbf{Z}^t})\mathbf{X}^{\mathrm{T}}\mathbf{Z}\right, \text{ s.t. } \mathbf{Z}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Z} = \mathbf{I}. \ (17)$$

The optimal solution of $\boldsymbol{Z}$ gives the Eigen vectors corresponding to the c smallest Eigen values of generalized Eigen-problem:

$$\mathbf{X}\left(\mathbf{L} + \gamma\mathbf{D}_{\mathbf{Z}^t}\right)\mathbf{X}^{\mathrm{T}}\mathbf{Z} = \mathbf{\Lambda}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Z} \qquad (18)$$

where $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal elements are Eigen values.

Output: Sort all the d features according to $||z_q||_2$ (q = 1, ..., d) in descending order and select the top k ranked features.

The resultant is $\mathbf{Z}$ as the feature selection matrix. Both the FSASL and UFSOL algorithms rearrange the original feature set accordingly, as per their prominence with the ranks of the corresponding algorithms. Later, the rear-ranged feature sets are fed to the classifiers to perform emotion classification.

2.4.3 Subset Feature Selection (SuFS):

After the unsupervised feature selection, a novel Subset Feature Selection algorithm is introduced upon the UFSOL and FSASL algorithms.

Algorithm 3: Subset Feature Selection (SuFS)

Input: Ranking vector **r** based on Unsupervised Feature Selection; Original Feature Vector **F** (1602 features); Accuracy vector **a** with accuracies based on ranking of various features using Feature Selection algorithm; l = number of features at which highest accuracy is obtained using UFSOL or FSASL.

Solution:

1: Initialize sub-rank (**sr**) with $a(1)$ (since, first accuracy value is always > 0)
2: Initialize $h = 2$
  for g = 0:1:l
    if **a**(g + 1) > **a**(g)
      sr(h) = **r**(g + 1)
    update $h \leftarrow h + 1$
  end
3: **for** i = 0:1:len(**sr**)
  sf(g) = **F**(:, sr(g))
  end

Output: Subset of original feature vector (**sf**)

To further reduce the dimension of the feature set without effecting the accuracy of the SER system, i.e., to obtain a better accuracy with a reduced feature set. The proposed SuFS depends on the ranking vector (i.e., prominence of the features) and the validation accuracy obtained from the features selected from UFSOL and FSASL algorithms. The ranking vector is according to d_2 smallest Eigen values of UFSOL algorithm and d smallest Eigen values of FSASL algorithm. The SuFS algorithm is discussed in Algorithm 3. The SuFS algorithm is applied to the features selected by UFSOL and FSASL to obtain **sf** feature vector. Further, the subset of features, i.e., features obtained from UFSOL-SuFS and FSASL-SuFS are given to the SVM classifier for both validation and testing.

3. Simulation Results and Discussion

In the proposed SER system, the 1602 INTER-SPEECH Paralinguistic and GTCC features are extracted from the speech signal. This huge set of features is fed to the UFSOL and FSASL algorithms for feature selection. In this paper, the support vector machine (SVM) classification technique with Linear and Radial Basis Function (RBF) kernels using Hold-Out and 10-fold Cross-Validation are used for emotion classification. Initially, the speech signal database is divided into training and testing datasets. The 80% of the dataset is considered for training and 20% for testing for hold-out validation. The k-fold cross-validation (here, k = 10) 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. In this work, the 10-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 10 parts, among that 9 parts are used for training the classifier (SVM), and testing is carried out on the hold-out or test data, i.e., the tenth part. This process is repeated in 10 folds, i.e., 10 times, until the entire dataset is completely trained.

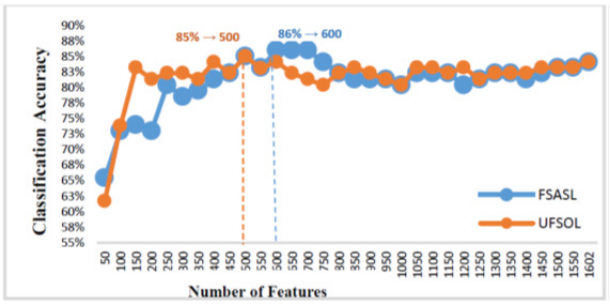

Fig. 4. Performance variation of the proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with EMO-DB database.

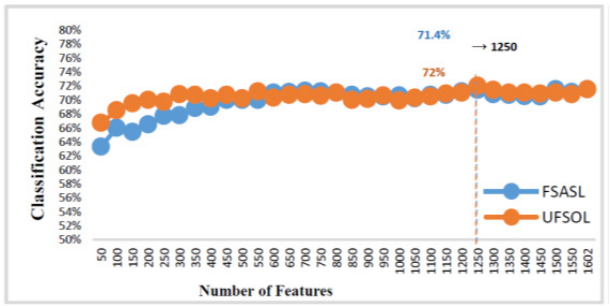

Fig. 5. Performance variation of the proposed SER system with FSASL and UFSOL feature selection using SVM classifier (10-fold cross-validation) with IEMOCAP database.

The performance of the proposed SER system is evaluated using the machine learning performance metric, i.e., the Classification Accuracy. In this work, the 10-fold cross-validation and Hold-Out Validation are used to train and test the accuracy of the proposed SER system. All the simulations are carried out in a Computer with Intel(R) Xeon(R) CPU E3-1220 v3 of 3.10 GHz 64-bit processor with 16 GB RAM. To select the first prominent features which give the highest accuracy, to select initial feature set, the feature selection matrix of both UFSOL and FSASL algorithms are given to the SVM classifier as shown in Fig. 3. Figures 4 and 5 show the variation of classification accuracy with the number of features using FSASL and UFSOL feature selection for EMO-DB and IEMOCAP.

For EMO-DB, using FSASL the highest validation accuracy of 86% is obtained for 600 features and 85% validation accuracy for 500 features with UFSOL. For IEMOCAP, for 1250 features the highest accuracy of 71.4% using FSASL and 72% using UFSOL is obtained.

It is evident from Figs. 4 and 5, even with initially selected features using UFSOL and FSASL algorithms, the SER accuracy is not increasing. Therefore, still, the feature selection is possible from initially chosen features. Hence, the SuFS algorithm is applied after UFSOL and FSASL feature selection to acquire better accuracy with less number of features. The initially selected features are fed to the SuFS algorithm to reduce further the number of features acquiring the best performance. Later, the highest prominent features selected by SuFS are fed to the SVM classifier with Linear and RBF kernels for emotion classification.

The best GTCC features selected for EMO-DB are GTCC [1] using FSASL and FSASL-SuFS, GTCC [2] using UFSOL and UFSOL-SuFS. While, for IEMOCAP, GTCC [1–20] i.e., the entire GTCC feature set is selected using FSASL, GTCC [1, 2, 4–7, 11] using FSASL-SuFS, GTCC [1–19] using UFSOL and GTCC [3–5, 7, 9, 10, 12–19] using UFSOL-SuFS. The best INTERSPEECH Paralinguistic 2010 features selected by each of the feature selection algorithm that are considered in the proposed SER are shown in Tab. 2.

The performance of the proposed SER system with different feature selection algorithms is compared with the baseline SER system (without feature selection) using SVM classifier with Linear and RBF kernels using hold-out validation and 10-fold cross-validation are as shown in Tab. 3 and 4 in terms of classification accuracy and validation (or) testing time. Tables 3 and 4 show the simulation results of the proposed SER system for EMO-DB and IEMOCAP databases with hold-out validation and 10-fold cross-validation using SVM classifier. From the results, it can be clearly understood that for EMO-DB database, better performance is achieved upon using the SVM classifier with Linear Kernel, and RBF kernel for IEMOCAP database.

From the results shown in Tab. 3 and 4, it is clear that the SVM with Linear kernel gives better classification for EMO-DB data and with RBF kernel in case of IEMOCAP data. Table 3 shows the hold-out validation results for EMO-DB and IEMOCAP database. For EMO-DB, the highest testing accuracy of 86% with the lowest computational time for training and testing, i.e., 0.165 and 0.032 seconds using FSASL-SuFS algorithm. Similarly, for IEMOCAP database, the highest testing accuracy and at lowest computational time of 14 and 2.9 seconds for training and testing is 77.5% using UFSOL-SuFS algorithm.

In Tab. 4, for EMO-DB database, using SVM with Linear kernel the 10-fold cross-validation accuracy of baseline SER system without feature selection is 85(±0.8) % with 1602 features. After applying the feature selection algorithms, the dimension of the feature set is reduced. The proposed SER system achieves an accuracy of 86% using UFSOL with selected 500 features and 85(±1.3)% using FSASL with 600 selected features. The SuFS algorithm is applied on these selected features of UFSOL and FSASL, thus reducing the number of features and acquiring the accuracy of 85(±1.5)% for UFSOL-SuFS with 450 features and 85(±0.8)% for FSASL-SuFS with 350 features. Similarly, for IEMOCAP database using SVM with RBF kernel, the 10-fold cross-validation accuracy of baseline SER system without feature selection is 69(±0.4)% with 1602 features. After feature selection, the proposed SER system achieves an accuracy of 69(±0.4)% using UFSOL and FSASL with selected 1250 selected features. The accuracy with UFSOL-SuFS is 77(±0.4)% with 800 features and 69(±0.4)% for FSASL-SuFS with 650 features. The confusion matrices with individual accuracy of each emotion of EMO-DB and IEMOCAP database using the proposed SER system with baseline, FSASL, UFSOL, FSASL-SuFS and UFSOL-SuFS are shown in Tab. 5 to 14.

Method			EMO-DB			IEMOCAP
FSASL	Position – max.		For all functionals and their deltas	Position – max.		For all functionals and their deltas
	Position – min.		For all functionals and their deltas except F0 Env	Position – min.		For all functionals and there deltas
	Arithmetic mean		*F0 Sub,F0 Env, MFCC[1,3-14]*	Arithmetic mean		*F0Sub+Δ,F0Env+Δ,Voicing Prob,JitterLocal,JitterDDP,ShimmerLocal, PCM,MFCC[0-14],MFCCΔ[0,1,3,5-7,9-14],LogMel[0-7],LSP [1-7]*
	Standard Deviation		*F0 Sub, F0 Env+Δ,MFCC[1-14], MFCCΔ[0,1,3,4,7,9-14]*	Standard Deviation		*F0 Sub+Δ,F0 Env+Δ,Jitter Local,Jitter DDP, Shimmer Local, PCM, MFCC+Δ[0-14], Log Mel [0-5,7], Log Mel Δ [0-7], LSP [1-7]*
	Skewness		*F0 Sub Δ, F0 Env+Δ, Voicing Prob, Jitter Local Δ, Jitter DDP+Δ,Shimmer Local+Δ,LSP[6],LSPΔ[6,7],MFCCΔ[1]*	Skewness		For all functionals and their deltas
	Kurtosis		*F0 Sub+Δ, F0 Env+Δ,Voicing Prob+Δ, Jitter Local+ Δ Jitter DDP+ Δ,Shimmer Local+ Δ, PCM Δ,MFCC[0,3-5,10,12,14], MFCCΔ[2,4-10,12,13], Log Mel [0-3,6], Log Mel Δ [1-7], LSP [0,5-7], LSPΔ [0-3,6]*	Kurtosis		For all functionals and their deltas
	Linear regression coefficient	c1	*F0 Sub*	Linear regression coefficient	c1	*F0 Sub+Δ,F0 Env+Δ, MFCC[0-2,4-14]*
		c2	*F0 Sub, F0 Env+Δ, MFCC[0-14]*		c2	*F0 Sub+Δ,F0 Env+Δ,Voicing Prob, Shimmer Local, PCM, MFCC+Δ[0-14],Log Mel [0-7], Log Mel Δ [0-6], LSP [0-7]*
	Linear regression error	A	*F0 by Sub, F0 Env, MFCC[0-13], MFCCΔ[1,7,9,10]*	Linear regression error	A	*F0 Sub+Δ,F0 Env+Δ, Jitter Local, Jitter DDP, Shimmer Local,PCM, MFCC+Δ[0-14], Log Mel+Δ [0-7]*
		Q	*MFCC[0-14], MFCCΔ[0-7,9-13], F0 Sub + Δ, F0 Env + Δ, Log Mel [0-6]*		Q	*F0 Sub+Δ,F0 Env+Δ, Shimmer Local, PCM, MFCC [0-14], MFCCΔ[0-13],Log Mel [1-7], Log Mel Δ [0-7]*
	Quartile	1	*F0 Sub, F0 Env, Log Mel [0,1,4], MFCC[0-12,14], MFCCΔ[1,10,14]*	Quartile	1	*F0 Sub+Δ,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP [0-7]*
		2	*F0 Sub, F0 Env, MFCC[0-14]*		2	*F0 SubΔ,F0 Env+Δ, Voicing Prob, Shimmer Local, MFCC+Δ[0-14], Log Mel [0-7], Log Mel Δ [0,1,4-6], LSP [1-7]*
		3	*F0 Sub, F0 Env+ Δ, MFCC[0-11,13,14], MFCCΔ[0,3,7,10,11,14], Log Mel [3-5]*		3	*F0 Sub+Δ,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM, Jitter Local, Jitter DDP, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP [1-7]*
	Quartile range	2-1	*F0 Sub, F0 Env, MFCC[0-4,6-9,11-14], MFCCΔ[7,10], Log Mel [0,5]*	Quartile range	2-1	*F0 Sub+Δ,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM, MFCC+Δ[0-14], Log Mel [0-7], Log Mel Δ [0,2-7], LSP [0,2,3,5]*
		3-1	*F0 Sub, F0 Env+ Δ, MFCC[0-14], MFCCΔ[0,1,3,4,6,7,9-12,14], Log Mel [0-7]*		3-1	*F0 Sub+Δ,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM+Δ, Jitter Local, Jitter DDP, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP [0,3,5]*
		3-2	*F0 Sub, F0 Env+ Δ, MFCC[0-4,6-14], MFCCΔ[0,1,3,6,9,10,14], Log Mel [3-5]*		3-2	*F0Sub+Δ,F0Env+Δ,VoicingProb, ShimLocal,PCM,JitteLocal,Jitter DDP, MFCC+Δ[0-14],LogMel[0-7],LogMelΔ [0,1,3-6], LSP [3,4]*
	Percentile	99.0	*F0 Sub, F0 Env+ Δ, MFCC[0-14], MFCCΔ[1-14], Log Mel [1-7], PCM*	Percentile	99.0	*F0Sub+Δ,F0Env+Δ,VoicingProb,ShimLocal, PCM+Δ, Jitter Local, JitterDDP,MFCC[0,2-14],MFCCΔ[0-14],LogMel+Δ[0-7],LSP[0-7]*
		1.0	*F0 Env Δ, MFCC+Δ[0-14], Log Mel [3-7]*		1.0	*F0Env+Δ,VoicingProb,PCMΔ,MFCC+Δ[0-14],LogMel+Δ[0-7], LSP[0-7]*
	Percentile range		*F0 Env, MFCC+Δ[0-14], Log Mel [1,2,5,6], Log Mel Δ[0,1,7], PCM*	Percentile range		*F0 Env+Δ, VoicingProb, PCM+Δ, MFCC+Δ[0-14], LogMel[0-7], LogMelΔ [0-2,4-7], LSP [0-7], LSPΔ[2]*
	Up-level time 90		*Jitter Local Δ*	Up-level time	75	*F0 Sub+Δ,F0 Env+Δ,Voicing Prob,Jitter Local+Δ, Jitter DDP, Shimmer Local+Δ,PCMΔ, MFCC[0-3,5-10,12-14],MFCCΔ[0-14], LogMel[0-6], LogMelΔ [0,4,5,7], LSP [0-7], LSPΔ[0-2,4,6,7]*
					90	*F0Sub+Δ,F0Env,VoicingProb,JitterLocal+Δ,JitterDDP+Δ, ShimmerLocal+Δ, MFCC[0,1,5-7,9,13],MFCCΔ[2,4,7,12], LogMel[0-2,4-6], LSP [0,2,3,5-7], LSPΔ[2]*
UFSOL	Position – max.		For all functionals and their deltas except LogMelΔ[6]	Position – max.		For all functionals and there deltas
	Position – min.		For all functionals and their deltas except F0 Env	Position – min.		For all functionals and there deltas
	Arithmetic mean		*F0 Sub,F0 Env, PCM, MFCC[0-14]*	Arithmetic mean		*F0SubΔ,F0Env+Δ,VoicingProb,,JitterDDP+Δ,ShimmerLocal+Δ, PCM, MFCC[0-14],MFCCΔ[3-6,8,12,14],LogMel[0-7],LSP[0-7]*
	Standard Deviation		*F0 Sub, F0 Env, MFCC[0,2,5,8-12]*	Standard Deviation		*F0 Sub+Δ,F0 Env+Δ, Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local+Δ,PCM, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP[0-2,5]*
	Skewness		*F0 Sub Δ, Jitter Local Δ, Jitter DDP+Δ,Shimmer LocalΔ,*	Skewness		For all functionals and their deltas
	Kurtosis		*F0 Sub+Δ,F0 Env+Δ,Voicing Prob,Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local+Δ,PCM, Log Mel [6], Log Mel Δ [2,4,6,7], LSP[6,7], LSPΔ [4,6]*	Kurtosis		For all functionals and their deltas
	Linear regression coefficient	c1	*PCM*	Linear regression coefficient	c1	*F0 Sub+Δ,F0 Env+Δ, PCM, MFCC[0-11,13,14]*
		c2	*F0 Sub,F0 Env+Δ,PCM,MFCC[0-14],Log Mel[1,2,4]*		c2	*F0Sub+Δ,F0Env+Δ,JitterLocal,JitterDDP,ShimmerLocal+Δ,PCM, VoicingProb,MFCC+Δ[0-14],LogMel[0-7],LSP[0-7],LogMelΔ[0-2]*
	Linear regression error	A	*F0 Sub,F0 Env, PCM, MFCC[0-4,7,8,10]*	Linear regression error	A	*F0 Sub+Δ,F0 Env+Δ, PCM, Shimmer Local+Δ, MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP[0]*
		Q	*F0 Sub,F0 Env+Δ,PCM,MFCC+Δ[0-14],Log Mel[0-7]*		Q	*F0Sub+Δ,F0Env+Δ,PCM,JitterDDPΔ,MFCC+Δ[0-14],LogMel[0-7], LogMelΔ[1-7]*
	Quartile	1	*F0 Sub,F0 Env,MFCC[0-14],Log Mel[0,2,3,5,6]*	Quartile	1	*F0Sub+Δ,F0Env+Δ,Jitter DDP,Shimmer Local+Δ, PCM,Voicing Prob,MFCC+Δ[0-14], Log Mel+Δ [0-7], LSP [0-7]*
		2	*F0 Sub,F0 Env,MFCC[0-14],Log Mel[6,7]*		2	*F0Sub+Δ,F0Env+Δ,VoicingProb, Jitter DDP,Shimmer Local+Δ, PCM,JitterLocal, MFCC+Δ[0-14],Log Mel [0-7], LSP [0-7]*
		3	*F0 Sub,F0 Env,MFCC[0-14]*		3	*F0Sub+Δ,F0Env+Δ,VoicingProb, Jitter DDP+Δ,Shimmer Local, PCM,JitterLocal, MFCC+Δ[0-14],Log Mel +Δ [0-7], LSP [0-7]*
	Quartile range	2-1	*F0 Sub,F0 Env,MFCC[0-7,9,10,12-14]*	Quartile range	2-1	*F0Sub+Δ,F0Env,ShimmLocal+Δ,PCM,MFCC+Δ[0-14],LogMel+Δ[0-7],LSP[6]*
		3-1	*F0 Sub,F0 Env+Δ,MFCC[0-14], MFCCΔ[0,1,2,6,7], Log Mel[1]*		3-1	*F0Sub+Δ,F0Env+Δ,Shimmer Local+Δ, PCM, VoicingProb Δ, Jitter DDPΔ, MFCC+Δ[0-14],Log Mel +Δ [0-7], LSP [0-3,5]*
		3-2	*F0 Sub,F0 Env, MFCC[0-3,5,6,8,9,13,14]*		3-2	*F0Sub+Δ,F0Env+Δ,Shimmer Local+Δ, PCM, MFCC+Δ[0-14], Log Mel [0-7], Log Mel Δ [1-7], LSP [1,2]*
	Percentile	99.0	*F0 Sub,F0 Env+Δ,MFCC[0-14], MFCCΔ[0-5,7-12],Log Mel[2,3]*	Percentile	99.0	*F0Sub+Δ,F0Env+Δ,Shimmer Local+Δ, Jitter DDP+Δ, PCMΔ, LSPΔ[1], , LSP [0-7], VoicingProb, Jitter Local, MFCC+Δ[0-14],Log Mel +Δ [0-7]*
		1.0	*F0 Env+Δ,MFCC[0-14], MFCCΔ[0-4,6-12],Log Mel[0,2,3,5-7]*		1.0	*F0Env+Δ, PCM+Δ, VoicingProb, MFCC+Δ[0-14],Log Mel +Δ [0-7], LSP [0-7], LSPΔ[0,1]*
	Percentile range		*F0Env+Δ,MFCC+Δ[0-14],LogMel[0,2,3,5-7], LogMelΔ[2]*	Percentile range		*F0Env+Δ, PCM+Δ, VoicingProb+Δ, MFCC+Δ[0-14],Log Mel +Δ [0-7], LSP [0-7], LSPΔ[1-6]*

	Up-level time 75		*Shimmer Local* Δ	Up-level time	75	*F0Sub+Δ,F0Env,Shimmer Local+Δ, Jitter DDP+Δ, VoicingProb, Jitter Local+Δ, MFCC+Δ[0-14], Log Mel[0-7],Log MelΔ[2,3,6, 7], LSP[0-7], LSPΔ[0-3,5-7]*
					90	*F0Sub+Δ,F0Env,ShimmerLocal+Δ,JitterDDP+Δ,VoicingProb,JitterLocal +Δ,MFCC[1,6,13,14],MFCCΔ[2],LogMel[0,1,6,7],LSP[1,3-7], LSPΔ[5]*
FSASL -SuFS	Position – max.		*F0 Sub+Δ,F0 Env,Voicing Prob+Δ,Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local,PCM+Δ, MFCC[0,2,3,5-8,10,11], MFCCΔ[0-3,9,11,14],Log Mel[0-2,6,7],Log MelΔ[0-2,4,6,7], LSP[1,2,5,7], LSPΔ[0,3-7]*	Position – max.		*F0SubΔ,F0Env,Shimmer Local+Δ, Jitter DDP+Δ, VoicingProb,PCM+Δ, MFCC[0,3-14], MFCCΔ[0,2-6,8-14], Log Mel[0-3,5-7], Log MelΔ[0-3,5,7], LSP[0,4-7], LSPΔ[0-3,5-7]*
	Position – min.		*F0 Sub+Δ,F0 Env Δ,Voicing Prob+Δ, Jitter DDP+Δ, MFCC[1-3,5,7,11,12], MFCCΔ[0-5,7,9,10,12-14],Log Mel[0,1,6,7],LogMelΔ[1,3-7], LSP[2,3,5-7], LSPΔ[2-7]*	Position – min.		*F0Sub+Δ,F0Env+Δ,Shimmer Local, Jitter DDP+Δ, VoicingProb+Δ, Jitter LocalΔ,PCM, MFCC[3-5,7,9,10,12-14], MFCCΔ[0,1,5-14], Log Mel +Δ [0-7], LSP[0,2-7], LSPΔ[0-7]*
	Arithmetic mean		*MFCC[1,3,4,8,10-13]*	Arithmetic mean		*F0SubΔ,Voicing Prob, JitterLocal, JitterDDP, PCM,MFCC[0,2-9,11-13], MFCCΔ[0,1,3,6,7,11-14],LogMel[1,4],LSP [2]*
	Standard Deviation		*F0 Env Δ, MFCC[2-4,6,7,9,10,12,13], MFCCΔ[7,10]*	Standard Deviation		*Jitter Local, PCM, MFCC[2,3,9,11], MFCCΔ[3,11,14], Log Mel [1,2,6], Log Mel Δ [2,5], LSP [0,2,3]*
	Skewness		*F0SubΔ,F0EnvΔ,Voicing Prob,Jitter Local Δ,Shimmer LocalΔ*	Skewness		*Shimmer Local+Δ,Jitter Local,PCM, MFCC[10,12], MFCCΔ[1,3,5,7 Log Mel [3,6,7], Log Mel Δ [0,4], LSP [1,6],LSPΔ[7]*
	Kurtosis		*F0SubΔ,F0Env,VoicingProb+Δ,JitterLocal,Shimmer LocalΔ, Jitter DDP+Δ, MFCC[0,3,12,14], MFCCΔ[2,9,10,12], Log Mel[0,1,2],Log MelΔ[4,6,7], LSP[0,6,7], LSPΔ[7]*	Kurtosis		*F0SubΔ,F0Env,Shimmer LocalΔ, Jitter DDP+Δ, VoicingProbΔ, Jitter Local, PCM, MFCC[5,9,10,13], MFCCΔ[1,3,7-9,11], Log Mel [0-3,6,7], Log Mel Δ [0,3,5-7], LSP [3-5],LSPΔ[0-2,5-7]*
	Linear regression coefficient	c1	*F0 Sub*	Linear regression coefficient	c1	*F0 SubΔ,F0 Env+Δ, MFCC[1,2,5-8,10-13]*
		c2	*F0 Sub, F0 Env+Δ, MFCC[0,2-4,6,8,10,12]*		c2	*F0 SubΔ, Voicing Prob, MFCC[0,2-8,11,12-14],MFCCΔ[1,2,12,13],Log Mel [0,1,4,5], Log Mel Δ [0,2,4-6], LSP [1,7]*
	Linear regression error	A	*F0 Sub, F0 Env, MFCC[1-8,10,11,13]*	Linear regression error	A	*F0 EnvΔ, Jitter Local, Jitter DDP, PCM, MFCC[0-2,4,5,7-9,12-14], MFCCΔ[1,3,6,12], Log Mel[1,2,5],Log MelΔ [1,2,5,7]*
		Q	*MFCC[0,2,5,7-14], MFCCΔ[0-5,9,12-13], F0 Sub, F0 Env, Log Mel [0-6]*		Q	*F0 EnvΔ, Shimmer Local, PCM, MFCC [1,3,6,7,11,13,14], MFCCΔ[0-2,4-8,10,12],Log Mel [0-5], Log Mel Δ [0,2,6,7]*
	Quartile	1	*F0 Sub, Log Mel [0], MFCC[0,1,4,6,9-11,14], MFCCΔ[10]*	Quartile	1	*F0 Sub,Voicing Prob,MFCC[1,3-7,10-14],MFCCΔ[1,3,4,6,10,11], Log Mel[1,6],LogMelΔ [0,2-4], LSP [0-2]*
		2	*MFCC[0-4,6-9,11-13]*		2	*F0 Sub,F0 EnvΔ, Voicing Prob, Shimmer Local, MFCC[1,2,4-11,13,14], MFCCΔ[0,2,3,5,7,8,11], Log Mel [0,1,3,4], Log Mel Δ [0,1,4,5], LSP [2]*
		3	*F0 Env+ Δ, MFCC[0-2,4,6-11,13], Log Mel [4]*		3	*F0 EnvΔ, Jitter Local, Jitter DDP, MFCC[0,2-4,8-10,13,14], MFCCΔ[3,8,12,13] Log Mel[0-3,6,7]LogMelΔ [1-3], LSP [1,2,7]*
	Quartile range	2-1	*F0 Sub, F0 Env, MFCC[2,4,6,8,9,11,14], MFCCΔ[10], Log Mel [0]*	Quartile range	2-1	*F0 Sub,F0 Env+Δ, Voicing Prob, Shimmer Local, PCM, MFCC[0,1,5-9,12], MFCCΔ[0-2,7,13], Log Mel[1,2],LogMelΔ [0,2,3,5], LSP [0,2,3,5]*
		3-1	*MFCC[3,6,10,11,13,14], MFCCΔ[0,1,3,4,6,7,9,10,11,14], Log Mel [0,3-6]*		3-1	*F0 SubΔ, Voicing Prob, PCM,Jitter Local, Jitter DDP, MFCC[0-4,6-8,10-12,14],MFCCΔ[2,5,10,15], Log Mel[0-3,6],LogMelΔ [0,3,7], LSP [3,5]*
		3-2	*F0 Env, MFCC[0,2,4,7,8,10,11-14], Log Mel [4]*		3-2	*F0SubΔ,F0EnvΔ,VoicingProb, PCM, JitterLocal, Jitter DDP, MFCC[2-13], LogMel[0,2], LogMelΔ [1,4,5], LSP [3,4]*
	Percentile	99.0	*MFCC[1-6,9,11], MFCCΔ[2-7,9-13], Log Mel [1-3,6,7]*	Percentile	99.0	*F0SubΔ, VoicingProb,ShimLocal, PCM+Δ, MFCC[4-6,8-11,13], MFCCΔ[0,1,3,5,8,9,13], LogMel[2],LogMelΔ[0,2,6,7],LSP[0,2,6,7]*
		1.0	*F0 Env Δ, MFCC+Δ[0-14], Log Mel [3-7]*		1.0	*F0 Env+Δ, VoicingProb, PCMΔ, MFCC[1-3,6-8,10-14],MFCCΔ[1-5,7-9,11,13,14], LogMel[4,7], LogMel Δ[1,3,7], LSP[1,4]*
	Percentile range		*MFCC[3,7-9],MFCCΔ[0-2,4,6-9,12,13],Log Mel [1,2,5], Log Mel Δ[0]*	Percentile range		*VoicingProb, MFCC[2,4-6,8,9,12,13],MFCCΔ[0,2,6-9,11,13], LogMel[3,5], LogMelΔ [2,5,6], LSP [0,1,3,7], LSPΔ[2]*
	Up-level time		___	Up-level time	75	*F0 EnvΔ, Jitter DDP, MFCC[3,5,8,12,14],MFCCΔ[0,2,8,9,11,13,14], LogMel[1,2,4,5], LogMelΔ [0,4,5,7], LSP [0-3,5], LSPΔ[0,4,7]*
					90	*Voicing Prob, Jitter DDP, Shimmer LocalΔ, MFCC[5-7,9,13], MFCCΔ[4,7,12], LogMel[0,2,4-6], LSP [0,2,3,5-7], LSPΔ[2]*
UFSOL -SuFS	Position – max.		*F0 Sub+Δ,F0 Env+Δ,Voicing Prob+Δ,Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local, PCM, MFCC[0-7,10-14], MFCCΔ[0-3,5-13], Log Mel [0,5-7], Log Mel Δ[4-7], LSP [0-6], LSPΔ [0,2,3-7]*	Position – max.		*F0 SubΔ,F0 Env,Voicing Prob+Δ,Jitter Local,,Shimmer LocalΔ, PCMΔ, MFCC[3-9,12], MFCCΔ[1,3,5,6,8-10], Log Mel [0,1,3,4,7], Log Mel Δ[0,3,5], LSP [0,2,7], LSPΔ [1,3-7]*
	Position – min.		*F0 Sub+Δ,Voicing Prob+Δ, Jitter DDP+Δ,Shimmer Local+Δ, PCM, MFCC[0-3,7-10,14], MFCCΔ[3-13], LogMel[0,1,4-7],LogMel Δ[3-7],LSP[3-5,7],LSPΔ[3-7]*	Position – min.		*F0 SubΔ,F0 EnvΔ,Voicing Prob,Jitter LocalΔ, Jitter DDPΔ,Shimmer LocalΔ, PCM+Δ, MFCC[1-3,6,7,12,13], MFCCΔ[0,3,5-8,10,12,13], Log Mel [4,7], Log Mel Δ[3,5], LSP [1,3,4,6,7], LSPΔ [2,4-6]*
	Arithmetic mean		*F0 Sub,F0 Env, PCM, MFCC[0-14]*	Arithmetic mean,		*F0SubΔ,F0Env+Δ,VoicingProb,,JitterDDPΔ,ShimmerLocalΔ,MFCC[1,3,5,6,8, 11,12], MFCCΔ[3-6,8,12,14],LogMel[2,5,6],LSP[0,1,4-7]*
	Standard Deviation		*F0 Sub, F0 Env, MFCC[0,2,5,8-12]*	Standard Deviation		*F0Sub+Δ,F0EnvΔ,JitterLocal,JitterDDP+Δ,ShimmerLocal+Δ,PCM,MFCC[1,2,6,7,11-14],MFCCΔ[0,3,4,8,12],LogMel[1-3,6],LogMelΔ[0-7],LSP[0-2,5]*
	Skewness		*F0 Sub Δ, Jitter Local Δ, Jitter DDP+Δ,Shimmer LocalΔ*	Skewness		*F0 Sub+Δ,F0 Env+Δ,Voicing ProbΔ,Jitter Local+Δ, Jitter DDP+Δ, Shimmer Local+Δ, PCM+Δ, MFCC[0,,3-5,8,10,11,13,14], MFCCΔ[0,5,7-9,14], Log Mel [0-3,5,7], Log Mel Δ[0,3-7], LSP [0,1,3-7], LSPΔ [2,3,6]*
	Kurtosis		*F0 Sub+Δ,F0 Env+Δ,Voicing Prob,Jitter Local+Δ, Jitter DDP+Δ,Shimmer Local+Δ,PCM, Log Mel [6], Log Mel Δ [2,4,6,7], LSP[6,7], LSPΔ [4,6]*	Kurtosis		*F0SubΔ,F0 Env+Δ, Jitter LocalΔ,Shimmer LocalΔ, PCM+Δ, MFCC[1,3-5,8,12-14], MFCCΔ[0,3,7,9,10,13,14], Log Mel [1-5], Log Mel Δ[0-3,6], LSP [0,1,3,4,6,7], LSPΔ [0,5-7]*
	Linear regression coefficient	c1	*PCM*	Linear regression coefficient	c1	*F0 Sub,F0 EnvΔ,MFCC[0-11,13,14]*
		c2	*F0 Sub,F0 Env+Δ,PCM,MFCC[0-14],Log Mel[1,2,4]*		c2	*F0Env,JitterLocal,JitterDDP,ShimmerLocal+Δ,VoicingProb,MFCC[3,4,6,13,14],MFCCΔ[2,4-6,8],LogMel[2,4,5],LSP[0,3-7],LogMelΔ[0-2]*
	Linear regression error	A	*F0 Sub,F0 Env, PCM, MFCC[0-4,7,8,10]*	Linear regression error	A	*F0 Sub+Δ,F0 Env+Δ, Shimmer Local+Δ, MFCC[0,2-5,7,8,10,12-14], MFCCΔ[1,5,6,7,9,10,13,14], LogMel[0,3,4,7],LogMelΔ [0-7], LSP[0]*
		Q	*F0 Sub,F0 Env+Δ,PCM,MFCC[0,3-7,12,13],MFCCΔ[0,3-12],Log Mel[0-5,7]*		Q	*F0 SubΔ,F0 Env+Δ,Jitter DDP Δ, MFCC[0-14],MFCCΔ[0,3-9,11], Log Mel [1-4,6], Log Mel Δ [1,3-7]*
	Quartile	1	*F0 Sub,F0 Env,MFCC[0-14],Log Mel[0,2,3,5,6]*	Quartile	1	*F0SubΔ,JitterDDP,ShimmerLocal,PCM,VoicingProb,MFCC[0,6,8,11,12,14], MFCCΔ[0,1,6,12-14], LogMel[2,4,5,7],LogMelΔ [0-7], LSP [0,1,4-7]*
		2	*F0 Sub,F0 Env,MFCC[0-14],Log Mel[6,7]*		2	*F0EnvΔ,VoicingProb, Jitter DDP,Shimmer Local,JitterLocal, MFCC[0,1,4-6,8,10,12,13],MFCCΔ[0-9,11-14],Log Mel [2,6], LSP [0,1,4-7]*

		3	*F0 Sub,F0 Env,MFCC[0-14]*		3	*F0Sub+Δ,VoicingProb,JitterDDP+Δ,ShimmLocal,PCM,JitterLocal, MFCC[0,4,6-8,10-12],MFCCΔ[0-2,7-9,11-14],LogMel[1,6,7], LogMelΔ[0-7],LSP[0,1,3-7]*
	Quartile range	2-1	*F0 Sub,F0 Env,MFCC[0-7,9,10,12-14]*	Quartile range	2-1	*F0SubΔ,Shimmer Local+Δ,PCM,MFCC[1-5,8,10,12],MFCCΔ[0,4,7-9,11-14], LogMel[0,1,5,6], LogMelΔ[0-7],LSP [6]*
		3-1	*F0 Sub,F0 Env+Δ,MFCC[0-14], MFCCΔ[0,1,2,6,7],Log Mel[1]*		3-1	*F0SubΔ,Shimmer LocalΔ, VoicingProb Δ, Jitter DDPΔ, MFCC[1,4,5,8-14],MFCCΔ[0,1,10-14],LogMel[1,4,6], LogMel [0-7], LSP [0-3,5]*
		3-2	*F0 Sub,F0 Env,MFCC[0-3,5,6,8,9,13,14]*		3-2	*F0Sub,F0EnvΔ,Shimmer LocalΔ, PCM, MFCC[1,2,8,9,11,12,14], MFCCΔ[1,2,7-9,11-14], LogMel [2-7], Log Mel Δ [1-7], LSP [1,2]*
	Percentile	99.0	*F0 Sub,F0 Env+Δ,MFCC[0-14], MFCCΔ[0-5,7-12],Log Mel[2,3]*	Percentile	99.0	*F0Sub+Δ,ShimmerLocalΔ,JitterDDPΔ,PCMΔ,VoicingProb,MFCC[0,2,5-9,11, 12,14],MFCCΔ[1-8,10-14],LogMel[3,5],LogMelΔ[1,3-7], LSP[0,2-7], LSPΔ[1]*
		1.0	*F0 Env+Δ,MFCC[0-14], MFCCΔ[0-4,6-12],Log Mel[0,2,3,5-7]*		1.0	*F0EnvΔ,PCM+Δ,VoicingProb, LogMel[0,2,4,7], LogMelΔ [1,5-7], LSP[0,1,3-7], LSPΔ[0,1],MFCC[6,7,9,12,14],MFCCΔ[0-3,5,7,9,10,12-14]*
	Percentile range		*F0 Env+Δ,MFCC+Δ[0-14],Log Mel[0,2,3,5-7], LogMelΔ[2]*	Percentile range		*F0EnvΔ,VoicingProb+Δ, MFCC[0,4,8,9,11,12],MFCCΔ[0,3,5-8,10,12-14], LogMel[3-5,7],LogMelΔ[0,2-4], LSP [0-7], LSPΔ[1-6]*
	Up-level time		____	Up-level time	75	*F0SubΔ,F0Env, VoicingProb, Jitter LocalΔ, MFCC[0,3-6,8-11,13],MFCCΔ[0,2-11,14],LogMel[0-7],Log MelΔ[2,3,6, 7], LSP[0,1,2,4,5], LSPΔ[0,1,3,5-7]*
					90	*F0Sub,F0Env,ShimmerLocalΔ, Jitter DDP+Δ, VoicingProb, Jitter LocalΔ, MFCC[1,6,13,14], MFCCΔ[2], Log Mel[0,1,6,7], LSP[1,3-7], LSPΔ[5]*

Tab. 2. Best INTERSPEECH 2010 paralinguistic features selected using UFSOL, FSASL, UFSOL-SuFS and FSASL-SuFS algorithms for the proposed SER system for EMO-DB and IEMOCAP databases.

Database	Method	No. of Features	Linear Kernel			RBF Kernel		
			Training	Testing		Training	Testing	
			Time (sec)	Time (sec)	Acc (%)	Time (sec)	Time (sec)	Acc (%)
EMO-DB	Baseline	1602	6.4	0.17	84.1	1.3	0.22	76.6
	UFSOL	**500**	**0.22**	**0.05**	**85**	**0.41**	**0.06**	**75.7**
	FSASL	**600**	**0.28**	**0.06**	**86.8**	**0.47**	**0.07**	**75.7**
	UFSOL -SuFS	450	0.21	0.043	84.9	0.57	0.08	74.8
	FSASL-SuFS	350	0.165	0.032	86	0.29	0.04	77.6
IEMOCAP	Baseline	1602	39.35	5.4	56.05	46.4	10.9	71
	UFSOL	1250	35	5.4	56.05	**29.4**	5.9	70.9
	FSASL	1250	34.1	5.3	57.3	**34.6**	7.7	70
	UFSOL-SuFS	800	21.2	3.4	60.6	14	2.9	77.5
	FSASL-SuFS	650	24.6	2.9	59.7	21	4.1	70.4

Tab. 3. Performance comparison of baseline and proposed SER systems for EMO-DB and IEMOCAP databases using SVM classifier with hold-out validation.

Database	Method	No. of Features	Linear Kernel		RBF Kernel	
			Time (sec)	Acc(%)	Time (sec)	Acc(%)
EMO-DB	Baseline	1602	3.2	85(±0.8)	12.13	81(±1.5)
	UFSOL	500	2.5	**86(±1.0)**	4.07	78((±1.5)
	FSASL	600	1.86	85(±1.3)	5.1	78((±1.4)
	UFSOL-SuFS	**450**	1.71	**84(±0.8)**	5.4	81((±1.4)
	FSASL-SuFS	**350**	1.4	**85(±0.8)**	2.68	78((±1.3)
IEMOCAP	Baseline	1602	304.9	58(±0.3)	430	69(±0.4)
	UFSOL	1250	289.4	58(±0.5)	310	69(±0.4)
	FSASL	1250	277.5	59(±0.5)	309	69(±0.4)
	UFSOL-SuFS	**800**	216.7	57(±0.5)	125.5	**77(±0.4)**
	FSASL-SuFS	**650**	199	58(±0.4)	199.8	**69(±0.4)**

Tab. 4. Performance comparison of the baseline and proposed SER system for EMO-DB and IEMOCAP databases using SVM classifier with 10-fold cross-validation.

Emotion	Ang	Anx	Bor	Dis	Hap	Neu	Sad
Ang	**94.1%**	0	0	0	5.9%	0	0
Anx	5.5%	**77.7%**	0	0	16.8%	0	0
Bor	0	0	**79%**	0	0	10.5%	10.5%
Dis	0	0	0	**75%**	0	12.5%	12.5%
Hap	6.2%	0	0	6.3%	**87.5%**	0	0
Neu	0	5.9%	5.9%	0	0	**88.2%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Tab. 5. Confusion matrix of baseline SER system for EMO-DB database.

Emotion	Ang	Anx	Bor	Dis	Hap	Neu	Sad
Ang	**92.4%**	0	0	0	7.6%	0	0
Anx	5.5%	**84.5%**	0	0	10%	0	0
Bor	0	0	**79%**	0	0	10.5%	10.5%
Dis	0	0	0	**87.5%**	0	12.5%	0
Hap	12.5%	0	0	0	**87.5%**	0	0
Neu	0	0	5.9%	0	0	**94.1%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Tab. 6. Confusion matrix of proposed FSASL based SER system for EMO-DB database.

Emotion	Ang	Anx	Bor	Dis	Hap	Neu	Sad
Ang	**94.1%**	0	0	0	5.9%	0	0
Anx	5.6%	**83.3%**	0	0	11.1%	0	0
Bor	0	0	**84.2%**	0	0	5.3%	10.5%
Dis	0	0	0	**75%**	0	25%	0
Hap	25%	0	0	0	**75%**	0	0
Neu	0	0	5.9%	0	0	**94.1%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Tab. 7. Confusion matrix of proposed UFSOL based SER system for EMO-DB database.

Emotion	Ang	Anx	Bor	Dis	Hap	Neu	Sad
Ang	**90.3%**	0	0	0	9.7%	0	0
Anx	9%	**82%**	0	9%	0	0	0
Bor	0	0	**84.2%**	0	0	5.3%	10.5%
Dis	12.5%	0	0	**87.5%**	0	0	0
Hap	12.5%	0	0	0	**87.5%**	0	0
Neu	0	0	5.9%	0	0	**94.1%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Tab. 8. Confusion matrix of proposed FSASL-SuFS based SER system for EMO-DB database.

Emotion	Ang	Anx	Bor	Dis	Hap	Neu	Sad
Ang	**96%**	0	0	0	4%	0	0
Anx	5.6%	**83.3%**	0	0	1.1%	0	0
Bor	0	0	**76.4%**	0	0	11.8%	11.8%
Dis	0	12.5%	0	**75%**	0	0	12.5%
Hap	6.2%	0	0	6.3%	**87.5%**	0	0
Neu	0	4.1%	5.9%	0	0	**90%**	0
Sad	0	0	16.7%	0	0	0	**83.3%**

Tab. 9. Confusion matrix of proposed UFSOL-SuFS based SER system for EMO-DB database.

Emotion	Ang	Hap	Neu	Sad
Ang	**84%**	1.3%	13.8%	0.9%
Hap	10.8%	**18%**	50.4%	20.8%
Neu	4.5%	3.5%	**80.2%**	11.8%
Sad	2.2%	1.3%	24.1%	**72.4%**

Tab. 10. Confusion matrix of baseline SER system for IEMOCAP database.

Emotion	Ang	Hap	Neu	Sad
Ang	**79.1%**	1.8%	17.8%	1.3%
Hap	10%	**19%**	45.9%	25.1%
Neu	4.5%	2%	**83%**	10.5%
Sad	3.5%	1.3%	24.2%	**71%**

Tab. 11. Confusion matrix of proposed FSASL based SER system for IEMOCAP database.

Emotion	Ang	Hap	Neu	Sad
Ang	**80.5%**	1.3%	16.9%	1.3%
Hap	11.7%	**18.1%**	46.8%	23.4%
Neu	3.1%	2.5%	**83.3%**	11.1%
Sad	3.5%	1.7%	23.2%	**71.6%**

Tab. 12. Confusion matrix of proposed UFSOL based SER system for IEMOCAP database.

Emotion	Ang	Hap	Neu	Sad
Ang	**85.5%**	5.8%	6.9%	1.8%
Hap	6.3%	**20.7%**	55%	18%
Neu	4.2%	3.5%	**80.9%**	11.4%
Sad	4%	2.5%	24.6%	**68.9%**

Tab. 13. Confusion matrix of proposed FSASL-SuFS based SER system for IEMOCAP database.

Emotion	Ang	Hap	Neu	Sad
Ang	**97.3%**	0.9%	0.9%	0.9%
Hap	8.1%	**22.6%**	50.4%	18.9%
Neu	1.7%	2.4%	**86.5%**	9.4%
Sad	1.3%	2.6%	22.8%	**73.3%**

Tab. 14. Confusion matrix of proposed UFSOL-SuFS based SER system for IEMOCAP database.

Methods		EMO-DB	IEMOCAP
Chen et al. 2016 [10]		77.4%	-
Zhang et al. 2013 [11]		80.85%	-
Zhang and Zhao 2013 [12]		78.5%	-
Yan et al. 2013 [9]		79.23%	-
Gudmalwar et al. 2019 [13]		75.32%	-
Ozseven 2019 [5]		84.07%	-
Sun et al. 2019 [6]		86.86%	-
Huang et al. 2015 [14]		71.16%	-
Sahu et al. 2018 [15]		-	58.38%
Latif et al. 2017 [16]		-	56.42%
Jiang et al. 2019 [17]		-	64%
Proposed SER System	**FSASL**	**86(±1.0)%**	**69(±0.4)%**
	UFSOL	**85(±1.3)%**	**69(±0.4)%**
	FSASL-SuFS	**85(±1.5)%**	**77(±0.4)%**
	UFSOL-SuFS	**85(±0.8)%**	**69(±0.4)%**

Tab. 15. Performance comparison of SER system with the existing literature works.

From the results, it is clearly understood that by using the unsupervised feature selection and inducing SuFS algorithm upon UFSOL and FSASL techniques, the proposed SER system provides improved accuracy with less computational complexity. Further, the performance of the proposed SER system is compared with the different works in the Tab. 15 for EMO-DB and IEMOCAP databases in terms of the Classification Accuracy performance metric. It is clearly evident that the proposed SER system upon using the feature selection process provided improved performance compared to the rest of the SER systems in the literature.

4. Conclusion

In this proposed SER system, the unsupervised feature selection techniques UFSOL and FSASL are employed to optimize the combination of INTERSPEECH 2010 Paralinguistic and GTCC features. Also, a novel SuFS algorithm is proposed upon the UFSOL and FSASL techniques to reduce further the feature dimension acquiring the comparable performance in the proposed SER system. The performance of the proposed SER system is analyzed with EMO-DB and IEMOCAP databases using SVM classifier with Linear and RBF kernels. 10-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 the over-fitting problem and Hold-Out validation scheme to test the performance of the proposed SER system with new data. The proposed SER system for EMO-DB data achieves highest classification accuracy using SVM with Linear kernel with 86% using FSASL and 85% using UFSOL, FSASL-SuFS and UFSOL-SuFS methods. Similarly, the highest classification accuracy for IEMOCAP database is obtained using SVM classifier with RBF kernel with 77% using FSASL-SuFS and 69% using the rest of the methods respectively. It is clearly evident from the results that the proposed SER system outperforms the baseline, i.e., the SER system without feature selection and also with 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.

Acknowledgments

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 Visvesvaraya Ph.D. Scheme for Electronics and IT.

References

[1] EL AYADI, M., KAMEL, M. S., KARRAY, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition,* 2011, vol. 44, no. 3, p. 572–587. DOI: 10.1016/j.patcog.2010.09.020

[2] VERVERIDIS, D., KOTROPOULOS, C. Emotional speech recognition: Resources, features, and methods. *Speech Communication,* 2006, vol. 48, no. 9, p. 1162–1181. DOI: 10.1016/j.specom.2006.04.003

[3] ANG, J. C., MIRZAL, A., HARON, H., et al. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 2015, vol. 13, no. 5, p. 971–989. DOI: 10.1109/TCBB.2015.2478454

[4] ARRUTI, A., CEARRETA, I., ÁLVAREZ, A., et al. Feature selection for speech emotion recognition in Spanish and Basque: On the use of machine learning to improve human-computer interaction. *PloS ONE,* 2014, vol. 9, no. 10, p. 1–23. DOI: 10.1371/journal.pone.0108975

[5] ÖZSEVEN, T. A novel feature selection method for speech emotion recognition. *Applied Acoustics,* 2019, vol. 146, p. 320–326. DOI: 10.1016/j.apacoust.2018.11.028

[6] 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, no. 2, p. 1–14. DOI: 10.1186/s13636-018-0145-5

[7] KUCHIBHOTLA, S., VANKAYALAPATI, H. D., ANNE, K. R. An optimal two stage feature selection for speech emotion recognition using acoustic features. *International Journal of Speech Technology,* 2016, vol. 19, no. 4, p. 657–667. DOI: 10.1007/s10772-016-9358-0

[8] JIN, Y., SONG, P., ZHENG, W., et al. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence (Italy), 2014, p. 4808–4812. ISBN: 978-1-4799-2893-4. DOI: 10.1109/ICASSP.2014.6854515

[9] YAN, J., WANG, X., GU, W., et al. Speech emotion recognition based on sparse representation. *Archives of Acoustics,* 2013, vol. 38, no. 4, p. 465–470. DOI: 10.2478/aoa-2013-0055

[10] CHEN, S. H., WANG, J. C., HSIEH, W. C., et al. Speech emotion classification using multiple kernel Gaussian process. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. Jeju (South Korea), 2016, p. 1–4. ISBN: 978-1-5090-2401-8. DOI: 10.1109/APSIPA.2016.7820708

[11] ZHANG, S., ZHAO, X. Dimensionality reduction-based spoken emotion recognition. *Multimedia Tools and Applications,* 2013, vol. 63, no. 3, p. 615–646. DOI: 10.1007/S11042-011-0887-X

[12] ZHANG, S., ZHAO, X., LEI, B. Speech emotion recognition using an enhanced kernel isomap for human-robot interaction. *International Journal of Advanced Robotic Systems,* 2013, vol. 10, no. 2, p. 1–7. DOI: 10.5772/55403

[13] GUDMALWAR, A. P., RAMA RAO, C. V., DUTTA, A. Improving the performance of the speaker emotion recognition based on low dimension prosody features vector. *International Journal of Speech Technology,* 2019, vol. 22, no. 3, p. 521–531. DOI: 10.1007/S10772-018-09576-4

[14] HUANG, Z. W., XUE, W. T., MAO, Q. R. Speech emotion recognition with unsupervised feature learning. *Frontiers of Information Technology & Electronic Engineering,* 2015, vol. 16, no. 5, p. 358–366. DOI: 10.1631/FITEE.1400323

[15] SAHU, S., GUPTA, R., SIVARAMAN, G., et al. Adversarial auto-encoders for speech based emotion recognition. In *INTERSPEECH*. Stockholm (Sweden), 2017, p. 1243–1247. DOI: 10.21437/Interspeech.2017-1421

[16] LATIF, S., RANA, R., QADIR, J., EPPS, J. Variational autoencoders for learning latent representations of speech emotion: A preliminary study. In *INTERSPEECH.* Hyderabad (India), 2018, p. 3107–3111. DOI: 10.21437/Interspeech.2018-1568

[17] JIANG, W., WANG, Z., JIN, J. S., et al. Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors,* 2019, vol. 19, no. 12, p. 1–15. DOI: 10.3390/s19122730

[18] BURKHARDT, F., PAESCHKE, A., ROLFES, M., et al. A database of German emotional speech. In *INTERSPEECH 2005*. Lisbon (Portugal), 2005, p. 1517–1520.

[19] BUSSO, C., BULUT, M., LEE, C. C., et al. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation,* 2008, vol. 42, no. 4, p. 335–359. DOI: 10.1007/s10579-008-9076-6

[20] KOOLAGUDI, S. G., RAO, K. S. Emotion recognition from speech using source, system, and prosodic features. *International Journal of Speech Technology*, 2012, vol. 15, no. 2, p. 265–289. DOI: 10.1007/s10772-012-9139-3

[21] SCHULLER, B., STEIDL, S., BATLINER, A., et al. The INTERSPEECH 2010 paralinguistic challenge. In *INTERSPEECH 2010.* Makuhari, Chiba (Japan), 2010, p. 2794–2797.

[22] EYBEN, F., WÖLLMER, M., SCHULLER, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia.* Firenze (Italy), 2010, p. 1459–1462. ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1874246

[23] VALERO, X., ALIAS, F. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE Transactions on Multimedia,* 2012, vol. 14, no. 6, p. 1684–1689. DOI: 10.1109/TMM.2012.2199972

[24] GUO, J., QUO, Y., KONG, X., et al. Unsupervised feature selection with ordinal locality. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*. Hong Kong (China), 2017, p. 1213–1218. ISBN: 978-1-5090-6068-9. DOI: 10.1109/ICME.2017.8019357

[25] DU, L., SHEN, YD. Unsupervised feature selection with adaptive structure learning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Sydney (Australia), 2015, p. 209–218. ISBN: 978-1-4503-3664-2. DOI: 10.1145/2783258.2783345

About the Authors …

Surekha Reddy BANDELA was born in Telangana State, India in the year 1990. She received her M.Tech. in VLSI & Embedded Systems from ER&DCI IT, CDAC Thiruvananthapuram, Kerala, India in 2015. At present she is the Ph.D. Scholar at NIT Warangal, India. Her research interests include speech emotion recognition, machine learning, multimodal emotion recognition, and speech synthesis.

T. Kishore KUMAR received his Ph.D. degree in the area of Signal Processing. His research interests include speech signal processing, adaptive signal processing, etc. He was the Former Head of the Dept. of ECE, N.I.T. Warangal, India and deputed to AIT Bangkok as Visiting Professor sponsored by MHRD, Govt. of India. He is currently working as Professor at Dept. of E.C.E. and Head of Computer Center, N.I.T. Warangal, India. He has R&D projects worth of about 1 Crore sponsored by SERB, DRDO and MHRD, Govt. of India. He was awarded as the "Best Engineering Researcher Award" and "Research Excellence Award" in the year 2017. He has published about 25 papers in reputed international journals.