On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion recognition

Abdenour Hacine-Gharbi, Philippe Ravier

On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion recognition

Abdenour Hacine-Gharbi[a], Philippe Ravier[b],
[a]LMSE laboratory, University of Bordj Bou Arréridj, Elanasser, 34030 Bordj Bou Arréridj, Algeria
[b]PRISME Laboratory, University of Orléans, 12 rue de Blois, 45067 Orléans, France

On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion recognition

Abstract:

Mutual information (MI) has been previously used to select the relevant features for the task of speech emotion recognition (SER). However, the procedure does not deliver the optimal number of relevant features. We propose MI based criterion for estimating this number defined as the minimum number of features that explains the variable of the class indices. In order to minimize the MI estimation errors, we also search the best histogram binning choice considering three formulas: Sturges, Scott and LMSE. Four selection strategies MMI, CMI, JMI and TMI have been implemented and applied on 39-features vectors and on large dimension vectors. The feature selection results have been validated on independent text SER system, based on GMM classifier and evaluated on EMO-db database. Results demonstrate that LMSE bin choice gives the best MI estimation and ensures a minimal number of features with slight performance drop. Particularly, using the proposed stopping criterion, the CMI strategy achieves reduction of 48.72% in the case of the 39-features vectors size and 67.86% in the case of large dimension vectors. Moreover, using the recognition rate criterion, the JMI strategy gives a comparable feature reduction with slight improvement of performance but requiring very high computation capabilities.

Keywords: speech emotion recognition, mutual information, binning of joint histogram, features selection, MFCC coefficients, GMM models.

1 INTRODUCTION

Speech Emotion Recognition (SER) has received much attention over the last decade due to its wide application in security fields, human-computer interaction, interactive teaching, computer games or marketing (Huang, Gong, Fu, & F, 2014). It aims at automatically identifying the emotion state of a speaker from speech signal using tools of signal processing and pattern recognition fields. Specifically, an SER system involves a speech analysis tool to efficiently extract features from speech signal and a pattern classifier to identify the emotion class of input speech signals. In the literature, several SER systems based on different feature extraction methods and classification approaches have been presented (G Shashidhar: K & Sreenivasa, 2012).

The features which are commonly used and discussed in several papers are the short-term spectral features including Linear Prediction Cepstrum Coefficients (LPCC), Perceptual Linear Prediction (PLP) and Mel-Frequency Cepstrum Coefficients (MFCC) and the prosodic features including the energy and pitch feature (Basu, Chakraborty, Bag, & Aftabuddin, 2017) (Wu, Falk, & Chan, 2011). According to several authors, results have demonstrated the efficiency of MFCC descriptor for the task of SER (Wu, Falk, & Chan, 2011) (Pan, Shen, & Shen, 2012) (Zaidan & Salam, 2016) (Trabelsi & Bouhlel, 2016). Furthermore, researchers have proposed several classification approaches, such as Hidden Markov Model (HMM) (Schuller, Rigoll, & Lang, April 2003), Gaussian Mixture Model (GMM) (Neiberg, Elenius, & Laskowski, 2006) (Vijesh Joe & Shinly Swarna Sugi, 2016), K-nearest Neighbors (KNN) (Lanjewar, Swarup, & Patel, 2015), Support Vector Machines (SVM) and Artificial Neural Networks (ANN) (Pao, Chen, Yeh, & Li, September 2006) (Mannepalli, Sastry, & Suman, 2018).

An important step that may be needed in the SER system is the selection of relevant features after feature extraction step. It has the principal aim to select the important features that contain relevant information about the emotion classes without redundancy. This decreases the computing time and memory capacity and furthermore may improve the accuracy while avoiding the well known curse of dimensionality phenomenon (Jain, Duin, & Mao, 2000).

The feature selection methods have been grouped into two principal categories (Kohavi & John, 1997). The first one is the wrappers category which methods are based on the accuracy of the classification system as relevance measure of the features subsets. Accordingly, methods from wrappers category depend on the classifier to be built (Giannoulis & Potamianos, 2012), which needs huge computational cost for reducing the size of high dimensional feature space. The second one is the Filters category which methods are based on the relevance of features useful for describing the classes. The information quantity shared between the features and the classes is used as relevance measure of the subset features. Accordingly methods from filters category do not depend on the classifier, which considerably reduces the computational cost, compared to the wrappers methods.

In (Manolov, Boumbarov, Manolova, Poulkov, & Tonchev, 2017), the authors have applied a filter approach algorithm based on the mutual information estimation as relevance measure of the features for the task of speech emotion recognition. They have used several strategies based on the mutual information maximization criterion using the Brown's Toolbox (Brown, Pocock, Zhao, & Lujan, 2012). However, two major concerns have not been introduced using this Toolbox. The authors (Brown, Pocock, Zhao, & Lujan, 2012) have mentionned that the computation of entropies for continuous and ordinal variables is highly non-trivial and requires an estimation of distributions which are not known practically. In fact, they have proposed the histogram approach using fixed-width bin to estimate the entropy and the mutual information. Indeed the toolbox operates on discrete data, which needs the discretization of continuous variables before applying mutual information maximization criteria. But the histogram proposed approach is a critical point that must be carefully driven before applying Brown's Toolbox. The discretization issue is thus one concern. Moreover, the toolbox does not allow getting out the estimated values of the mutual information. These values can be used in order to determine the optimal number of relevant features. The optimal number issue is the main concern of this work.

We firstly propose to give more details about the entropy and the MI computation using histogram approach and to study the influence of the histogram bin number choice on the estimation quality. Based on this study, we secondly discuss the optimal number of features for the task of SER using MI values based criterion. In order to validate this proposal, an SER system has been carried out first using MFCC features with energy and their first and second derivatives and second using a combination of several types of spectral features (MFCC, LPCC, PLP) and prosodic features (pitch and energy). The latter validation investigates large features dimensions.

Also, the system performance was evaluated using the Berlin Database of Emotional Speech (Emo-DB) that considers different emotion classes such as anger, boredom, disgust, fear, happiness, sadness and neutral (Burkhardt, Paeschke, Rolfes, & Sendlm, 2005).

2 EMOTIONAL SPEECH RECOGNITION SYSTEM

An automatic emotion recognition system is a pattern recognition sytem generally composed of two important phases, the traning (learning) phase and the testing (classification) phase. Both of these phases require a features extraction step to transform each temporal signal into a sequence of short-term feature vectors. The training phase aims to learn the classes of emotion patterns from the occurences of training database using a classes modeling approach. In the testing phase, the system uses a classifier to identify the class the unknown input signal belongs to. Generally, a testing database of signals is used to classify each signal and finally evaluate the system performance using accuracy criterion. Based on this general description, we hereafter present the architecure of the SER system we developped with a selection step useful for estimating the optimal number of selected features. Feature selection will be introduced in section 3.

2.1 Proposed architecture

Figure (1) presents the diagram of our SER system. It shows the training phase for the learning of the emotion occurences and the testing phase for the performance study of the automatic identification system. Each phase takes into count a short-term feature extraction. The diagram also includes the selection procedure provided in this study for dimensionality reduction.

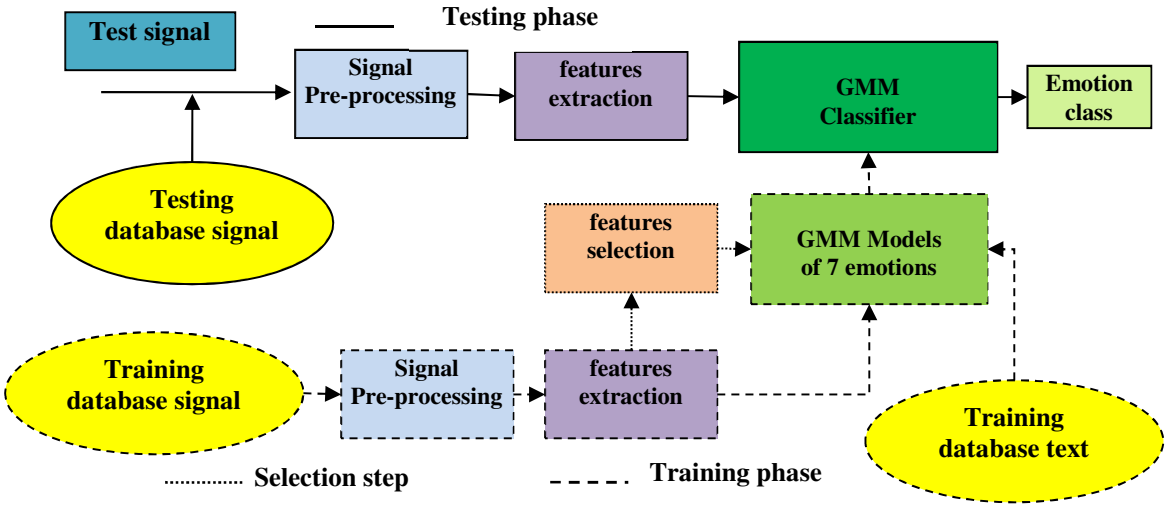

Figure 1 Diagram of the SER system: the training phase (dashed lines) learns the GMM emotion models using the occurrences of the training database with their corresponding text; the selection step (dotted lines) extracts the most relevant features which reduces the dimensionality of the problem; the testing phase (full lines) decides in which emotional class the test signal belongs to.

2.2 Emotion modelling

The emotions recognition system used in this work is based on the GMM approach that models each emotion by GMM with n Gaussians (Vijesh Joe & Shinly Swarna Sugi, 2016). Implementation of the system is carried out using HTK tools (Hidden Markov Model Toolkit), in which we consider GMM as HMM model with one state modeled by GMM of n Gaussians components with diagonal covariance matrix (Young, Kershaw, Odell, & Ollason, 1999).

The HMM model is fed by speech descriptors and according to several authors (Wu, Falk, & Chan, 2011), results have demonstrated the efficiency of MFCC descriptor for the task of SER (Pan, Shen, & Shen, 2012) (Zaidan & Salam, 2016). In (Trabelsi & Bouhlel, 2016), the authors have compared performance results of different descriptors such as MFCC, PLP, LPCC, and Rasta PLP. The study has shown that the best performance results are obtained with 12-coefficient MFCC. In order to take into account dynamic evolutions of the data, we used MFCC with energy and their 1st and 2d derivatives which forms 39-components feature vectors (Wu, Falk, & Chan, 2011).
Each utterance of the database is preprocessed by suppression of silence of signals boundaries and is filtered by a high-pass filter with a pre-emphasis coefficient of 0.97 (Wu, Falk, & Chan, 2011). Then, each obtained utterance signal corresponding to an emotion class is converted into a sequence of short-term vectors of 39 features computed each 10 ms on 30 ms hamming-windowed speech frame, using the 'Hcopy' command of HTK library. The vectors sequences of the training database are used to model each emotion class by a GMM model using the command 'HEREST'. Next, each sequence of vectors of the testing database is classified using the command 'HVITE'. Finally, performance evaluation is done by using the command 'HResult'.
The quality of the classification system is evaluated by a recognition rate RR defined as:

$$\mathrm{RR} = \frac{O - M}{O}$$

where O is the total number of occurrences given at the input of the classifier and M is the number of misclassified occurrences.

2.3 Speech database

We have used the Berlin Database of Emotional Speech (EMO-DB) to evaluate the system performances (Burkhardt, Paeschke, Rolfes, & Sendlm, 2005). The dataset is composed of 10 German sentences of different texts (5 short sentences constitute a set A and 5 longer sentences constitute a set B) pronounced by 10 actors (5 male, 5 female) who simulated seven primary emotion states (anger, boredom, disgust, fear, happiness, sadness)

including neutral. The sentences come from everyday communication and can be interpretable in all applied emotions. The total collection consists of 800 utterances including some second versions but the final collection considered only 535 utterances because a human validation was performed for each utterance (20 listeners had to decide in which emotional state the speaker had been and the decision for dataset inclusion was made when the recognition rate was higher than 80% and also considering for more than 60% of the listeners sentences are natural). Table1 details the distribution of the sentence recordings among the 7 emotional states as well the numbers of occurences used for testing and training phases, respectively. The sentences have a length of 1-2 seconds. Recordings were taken with a sampling frequency of 48 kHz and later downsampled to 16 kHz.

Table 1 Distribution of the sentences of EMO-DB among the 7 emotional states and per state for testing / training.

Emotions	**anger**	**boredom**	**disgust**	**fear**	**happiness**	**sadness**	**neutral**
Number	127	81	46	69	71	62	79
Test / Train	62/65	40/41	21/25	34/35	33/38	30/32	38/41

In the present work, the set A of the short sentences is taken as the training database that has 277 utterances, whereas the set B of the longer sentences is taken as the testing database, that has 258 utterances. Since the sentences of the testing database have not the same text than the text of the training database, hence we obtain an SER system in independent text mode.

3 HISTOGRAM APPROACH BASED MUTUAL INFORMATION FOR FEATURE SELECTION

3.1 The feature selection problem

In large dimension problems, dimensionality reduction of the number of features is often a necessity. The reduction can be achieved either by transforming the features of by selecting them. The first approach consists in transforming the features of an initial set F of n features $\{Y_1, Y_2 \cdots, Y_n\}$ in a low dimension subset of k features. This solution however requires the computation of all the features as well as the choice of an appropriate criterion for the definition of the transformation, which is not easy. The second approach consists in selecting the k most relevant features $\{Y_{P_1}, Y_{P_2} \cdots, Y_{P_k}\}$ from the set F which forms the subset S_{opt}. In opposite to the former, this second solution needs only the k selected features to be computed for the classification task in the testing phase. This approach will be preferred.

The feature selection procedure uses an information measure of a subset of features useful for a classification task. S_{opt} is an optimal subset of features if its information is maximum for the classification task. Mutual information (MI) is often used as a quantity of information measure because of its ability of assessing the nonlinear statistical dependency between variables. So the subset S_{opt} is chosen in such a way that the MI between S_{opt} and the class label C is maximized:

$$S_{opt} = \arg\max_{S \subset F} I(C; S). \tag{1}$$

However, the number of combinations of features for exhaustively constructing the sets S to be tested rapidly becomes prohibitive when the size of S grows. To circumvent this problem, "greedy forward" search strategies can be employed. The search is a one-by-one selection procedure that gives at each step j the best feature Y_{P_j} from the unselected features set. This new selected feature Y_{P_j} grows the already selected subset S_{j-1} by appending it as $S_j = Y_{P_j} \cup S_{j-1}$:

$$Y_{P_j} = \arg\max_{Y_i \in F - S_{j-1}} \left[I\left(C; Y_i, S_{j-1}\right) \right] \tag{2}$$

Since $I\left(C; Y_i, S_{j-1}\right) = I\left(C; S_{j-1}\right) + I\left(C; Y_i \backslash S_{j-1}\right)$ (Cover & Thomas, 1991), (2) can be reduced to:

$$Y_{P_j} = \arg\max_{Y_i \in F - S_{j-1}} \left[I\left(C; Y_i \backslash S_{j-1}\right) \right] \tag{3}$$

Equation (3) can also be expanded in a multivariate MI of order 3 between C, Y_i and S_{j-1} as:

$$Y_{P_j} = \arg\max_{Y_i \in F - S_{j-1}} \left[I(C; Y_i) - I_3\left(C; Y_i; S_{j-1}\right) \right] \tag{4}$$

The I_3 term may be positive which corresponds to withdrawing redundancy introduced by the new feature. If this term is negative, this means that Y_i and S_{j-1} are synergic (Cover & Thomas, 1991).

The evaluation of $I_3(C; Y_i; S_{j-1})$ becomes very difficult when j grows because this evaluation requires the estimation of high-dimensional probability density functions that cannot be precise enough for fixed database sizes (Drügman, Gurban, & Thirian, 2007). Most of the algorithms propose a simplification of (4) following different strategies like MIM, MIFS, MRMR, CMI, DISR, CIFE, TMI, ICAP (Brown, Pocock, Zhao, & Lujan, 2012) (Hacine-Gharbi, Deriche, Ravier, Harba, & Mohamadi, 2013). In (Brown, Pocock, Zhao, & Lujan, 2012), the authors conclude that the JMI strategy provides a good compromise between precision, flexibility and stability when the database is small size. They also point out the MRMR and CMI strategies that perform better than other ones in terms of balance between high relevance and small redundancy. We give below the derivation of (4) for four selected strategies.

- MMI (Maximum MI)

$$Y_{P_j} = \arg \max_{Y_i \in F - S_{j-1}} [I(C; Y_i)] \tag{5}$$

- TMI (Truncated MI)

$$Y_{P_j} = \arg \max_{Y_i \in F - S_{j-1}} \left[I(C; Y_i) - \sum_{k=1}^{j-1} I_3(C; Y_i; Y_{P_k}) \right] \tag{6}$$

- JMI (Joint Mutual Information)

$$Y_{P_j} = \arg \max_{Y_i \in F - S_{j-1}} \left[I(C; Y_i) - \frac{1}{j-1} \sum_{k=1}^{j-1} I_3(C; Y_i; Y_{P_k}) \right] \tag{7}$$

- CMI (Conditional Mutual Information)

$$Y_{P_j} = \arg \max_{Y_i \in F - S_{j-1}} \left[I(C; Y_i) - \max_{Y_{P_k} \in S_{j-1}} I_3(C; Y_i; Y_{P_k}) \right] \tag{8}$$

For JMI and CMI strategies, the term $I_3(C; Y_i; Y_{P_k})$ is actually computed as $I(Y_i; Y_{P_k}) - I(Y_i; Y_{P_k} \backslash C)$. The MI $I(X; Y)$ between variables X and Y is expressed as $I(X; Y) = \iint_{-\infty}^{+\infty} p(x, y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$ where $p(x, y)$ is the joint distribution of (X, Y) and $p(x)$ and $p(y)$ are the marginal distributions. This continuous definition can be estimated by considering the discrete version of the $I(X; Y)$ formula and by applying a histogram partitioning in the estimation of the distributions. Partitioning affects performance of the MI discrete estimator. This constitutes a binning problem that is introduced in the next section.

3.2 The binning problem

All the strategies are faced to MI estimation errors with the increasing number of selected features. Indeed, the maximization procedure of MI is based on the sum of individual MI estimations which number grows with the number of selected features. The result is an accumulation of errors, which can produce very different evolutions of the MI values between the current selected subset and the class label as a function of the selected features, for the same data. Moreover, as the number of samples decreases for a correct estimation with the MI dimension, the estimation error worsens. So care must be taken in the MI computations in order to limit the error accumulations that are harmful for the feature selection criterion and for finding an optimal number of features for a speech emotion recognition task.

The uniform histogram partitioning is often used because of some existing formula for an immediate estimation of the number of bins k, or equivalently of the bin width Δ. The formula use the data samples number N and may also require some classical statistical parameters of the data. Three formula were investigated. Sturges proposed $k = 1 + \log_2(N)$ (Sturges, 1926). Scott proposed $\Delta = 3.5\sigma/\sqrt[3]{N}$ where σ stands for the data standard deviation (Scott, 1992). A more recent estimator proposed in (Hacine-Gharbi, Deriche, Ravier, Harba, & Mohamadi, 2013) minimizes the mean square error estimation of MI. This LMSE estimator writes

$$k = \text{round}\left\{\frac{1}{2} + \frac{1}{2}\sqrt{1 + 4\sqrt{\frac{6N\hat{\rho}}{1-\hat{\rho}^2}}}\right\} \tag{9}$$

where the unknown correlation coefficient ρ of the data has been replaced by its estimated value $\hat{\rho}$. Additionally, the MMI criterion makes only use of $I(C; Y_i)$ which is developed as $I(C; Y_i) = H(Y_i) - H(Y_i \backslash C)$.

This computation requires the following LMSE formula for the entropy estimation (Hacine-Gharbi, Deriche, Ravier, Harba, & Mohamadi, 2013):

$$k = \text{round}\left\{\frac{\delta}{6} + \frac{2}{3\delta} + \frac{1}{3}\right\} \tag{10}$$

with $\delta = \sqrt[3]{8 + 324N + 12\sqrt{36N + 729N^2}}$ assuming data follow a Gaussian distribution with range equal to six times the standard deviation.

The three strategies have been used and compared in the frame of the feature selection procedure.

4 EXPERIMENTS AND RESULTS

Several experiments are conducted in order (1) to give the optimal configuration parameters of the SER system; (2) to study the influence of the bin number on the MI estimation for feature selection in the task of SER; (3) to estimate the optimal number of features from the MI curve, and validate the result using accuracy criterion.

4.1 Configuration study of the GMM-based SER system

The design of an SER system based on GMM classifier firstly requires searching the optimal number of Gaussian components of GMM models of the emotions classes, which gives the best accuracy rate. In order to practically demonstrate the importance of adding energy and dynamic features to the static MFCC features, a comparative study is performed. We called this configuration of descriptor as MFCC_EDA, in which E represents the energy, D the derivative Δ (speed) and A the double derivative ΔΔ (acceleration). Hence, this experience aims to find the best combination for the Gaussian components number and the descriptor type.

Table 2 gives the RR of SER system based on GMM models as a function of different number of Gaussian components and different descriptors types.

Table 2 Recognition rate as a function of the Gaussian number n of GMM models and descriptor types.

n	MFCC_EDA	MFCC	MFCC_E	MFCC_ED	MFCC_D
1	61.63	49.22	51.94	62.02	57.36
2	52.71	54.26	53.10	60.08	62.40
4	68.60	60.85	60.85	64.34	66.28
8	64.34	63.95	61.24	64.73	63.18
16	72.09	62.79	64.34	73.26	65.89
32	75.97	67.83	67.44	71.71	71.71
64	80.62	74.81	73.64	77.52	74.03
128	84.50	76.36	80.23	84.88	79.84
256	82.95	78.68	81.01	84.11	82.17

From this table, we can give these points:
- the optimal combination is obtained taking the Gaussian components number equal to 128 and taking the descriptor MDCC_ED or MFCC_EDA;
- the energy and the dynamic features Δ improve the RR;
- the ΔΔ added alone does not improve the RR.

In the following sections, we will consider the MFCC_EDA descriptor.

Table 3 gives the confusion matrix obtained in the case of MFCC_EDA descriptor. From this matrix, fear and happiness classes have the worst performance values.

Table 3 Confusion matrix for MFCC_EDA descriptor of the SER system.

	Anger	Bore.	Disg.	Fear	Happ.	Neut.	Sad.
Anger	96.78	0	1.61	0	1.61	0	0
Bore.	0	85.0	0	0	0	10	5
Disg.	4.76	4.76	90.48	0	0	0	0
Fear	11.77	0	0	52.94	17.65	8.82	8.82
Happ.	33.33	0	0	0	66.67	0	0
Neut.	0	2.63	0	0	0	97.37	0
Sad.	0	3.33	0	0	0	3.33	93.34

4.2 Histogram binning study for feature selection

The aim of this experiment is to study the effect of bin number choice on the MI estimation for the feature selection using MMI, CMI, JMI and TMI strategies. The different binning formulas used are those of Sturges, Scott, LMSE given in subsection (3.2).

Figures 2, 3, 4 and 5 show the results of MI estimation $I(C; Y_j, S_{j-1})$ using respectively MMI, CMI, JMI and TMI selection strategies (the MI estimations are the expressions in the brackets of eq. 2 to 5). For each figure, we consider the previews binning formulas with corresponding ST, SC, LMSE legend.

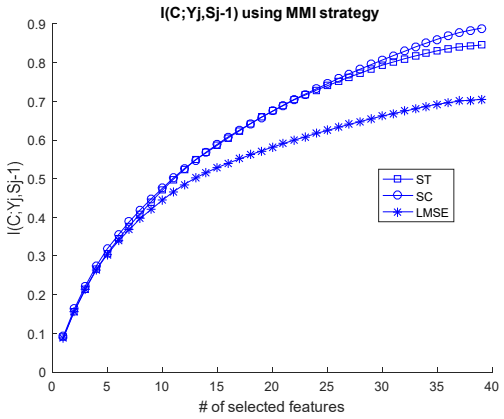

Figure 2 Estimation of $I(C; Y_j, S_{j-1})$ using **MMI** strategy with Sturges, Scott and LMSE bin choices.

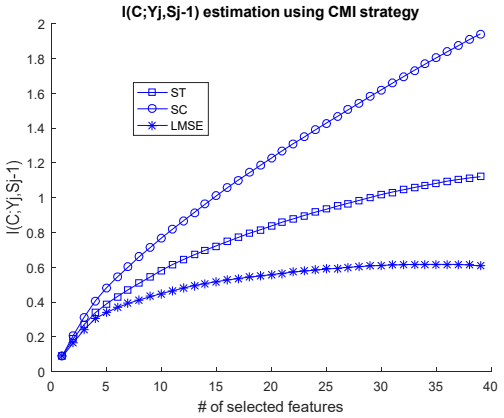

Figure 3 Estimation of $I(C; Y_j, S_{j-1})$ using **CMI** strategy with Sturges, Scott and LMSE bin choices.

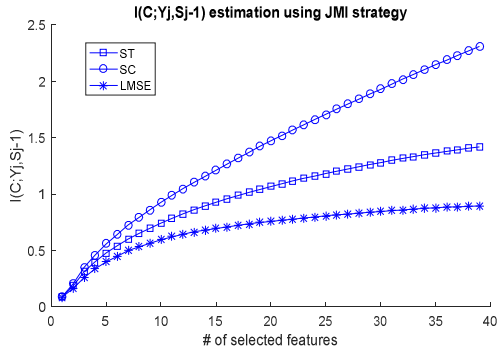

Figure 4 Estimation of $I(C; Y_j, S_{j-1})$ using **JMI** strategy with Sturges, Scott and LMSE bin choices.

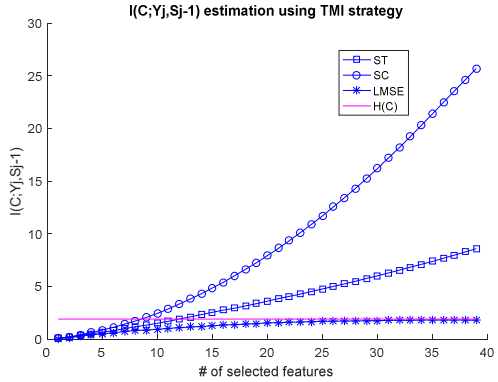

Figure 5 Estimation of $I(C; Y_j, S_{j-1})$ using **TMI** strategy with Sturges, Scott and LMSE bin choices. The entropy $H(C)$ is added showing that this theoretical value is reached only using the LMSE binning choice with the TMI strategy.

From the previews results, adding the 13 features of ΔΔ descriptor does not improve the RR, hence these features will not add any information that explains the emotion classes. Theorically, the $I(C; Y_j, S_{j-1})$ values must not surpass the entropy $H(C)$ of the class index variable C and should reach a plateau for an optimal number of relevant features (Hacine-Gharbi, Deriche, Ravier, Harba, & Mohamadi, 2013). However, practically, the curves of the MI may not reach a plateau after the selection of the relevant parameters because of the MI approximation using heuristic methods and also because of the MI estimations errors caused principally by the limited of samples number. Figures 2, 3, 4 and 5 show that the mutual information increases rapidely with the number of selected features for Scott and Sturges bin choices whatever the selection strategy and reaches great values compared to the LMSE bin choice. Especially in the case of TMI strategy with using Scott bin choice, the MI curves reach far higher values than those obtained with other bin choices and than the entropy $H(C)$. Only the LMSE bin choice allows the MI estimation approximately reach a plateau.
To sum up, figure 6 shows together the curves of the MI for the four strategies using the LMSE bin choice.

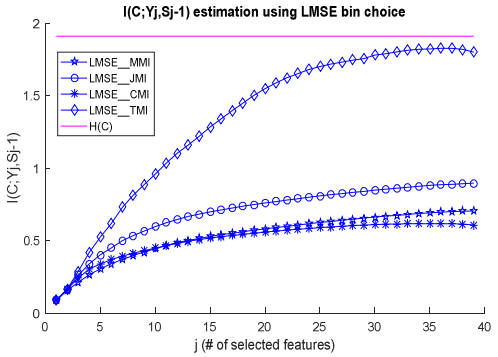

Figure 6 Estimation of $I(C; Y_j, S_{j-1})$ using MMI, JMI, CMI and TMI strategies in the case of LMSE bin choice.

The curve obtained by TMI strategy gives higher values than those of the others strategies, which is probably explained by a higher error acumulation of I_3 estimation in the sum. Furthermore the CMI, JMI and MMI curves present a plateau at approximately 20 features with minimum fluctuations caused by MI estimation errors. However the TMI reaches approximaly the plateau at 30 features.
Hence, in the following section, we consider only the LMSE bin choice.

4.3 Performance study for feature selection using MMI, CMI, JMI and TMI strategies

This experiment aims to study performance of the SER system in terms of RR as a function of the selected features order. We consider the four selection strategies using LMSE bin choice. Figure 7 shows the emotion RR results with respect to the selected features number using MMI, CMI, JMI and TMI strategies.

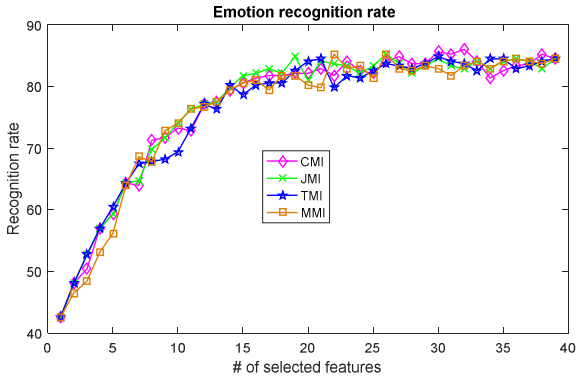

Figure 7 Emotion RR as a function of the selected features order with the four strategies MMI, CMI, JMI and TMI.

It can be observed that about 20 features are sufficient for explaining the classes because a similar RR value is obtained with the total number of 39 features. However discrepancies exist between the strategies in the first selected features. Table 4 and 5 give the numbers of the selected features with the corresponding RR values with the features subset growing, respectively.

Table 4 Number of the first 10 selected features for each strategy.

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
MMI	2	1	4	5	3	9	10	6	8	13
CMI	2	13	5	1	4	9	10	3	8	28
JMI	2	13	1	5	4	9	3	8	10	6
TMI	2	13	1	5	8	9	3	28	4	27

It is clear from these tables, that the number of static features is dominant which confirms the results obtained in (Trabelsi & Bouhlel, 2016). Further, the MMI strategy gives the worst performance results for the 6 first selected features. This can be explained by the MMI selection procedure that does not take into account any redundancy with the already selected features.

Table 5 Recognition rate obtained for the first 10 selected features.

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
MMI	42.64	46.51	48.45	53.10	56.20	63.95	68.60	67.83	72.87	74.03
CMI	42.64	48.06	50.39	56.98	59.30	64.34	63.95	71.32	71.71	73.26
JMI	42.64	48.06	52.71	56.98	59.30	64.34	64.73	69.77	71.71	74.03
TMI	42.64	48.06	52.71	56.98	60.47	64.34	67.44	67.83	68.22	69.38

4.4 Study of optimal number of features

From the previews experiments, we have noticed from the RR and MI curves, that a number of features greater than 20 can approximately reach a plateau. This experiment now aims at searching for the minimum number of features that gives the best performance results, taking independently two criteria respectively based on the MI estimation values (classifier-independent) and the RR values (classifier-dependent).

Firstly, we follow the same criterion as the one described in (Hacine-Gharbi & Ravier, 2018). The criterion considers the maximum value $MI_{max} = \max_j MI(C; Y_j, S_{j-1})$ of the MI estimation and a parameter α which corresponds to a small reduction of the MI_{max} value. This MI reduction that equals to $(1-\alpha)\%$ can be explained by MI estimation errors and by heuristic MI approximations. Hence, the experiment searches for the minimum number of selected features that reaches αMI_{max}. Table 6 shows the evolution of the minimum number of selected features (#SF) as a function of α values given in % taken as 100, 98, 96, 94, 92 and 90, for the four tested strategies. The RR values are also given with $\overline{RR}$ values that are normalized with respect to the RR obtained with the total number of features (84.50 % for 39 features).
By considering a 10% reduction of MI_{max}, the CMI strategy gives the best reduction of the number of features (20) but the $\overline{RR}$ drops of 2.76%. Further the JMI, TMI and CMI give a good compromise between the reductions of feature numbers and a small $\overline{RR}$ drop of about 1%. Even if this criterion is independent of the RR, it succeeds

in providing a reduced features number that gives an acceptable performance results, without the necessity of highly costly classifier-dependent feature number searching procedure setup.

Table 6 Results of optimal number of features in the classifier-independent case. The #SF value is the minimum number that reaches α % of the MI_{max} value. The RR and $\overline{RR}$ values are the corresponding recognition rates with their normalized versions (w.r.t. RR(39)).

	α	100	98	96	94	92	90
MMI	**#SF**	39	35	33	31	29	27
	RR	84.50	84.11	84.11	81.78	83.33	82.95
	$\overline{RR}$	100	99.54	99.54	96.78	98.62	98.17
CMI	**#SF**	35	29	26	24	22	20
	RR	82.56	83.72	84.11	82.56	81.78	82.17
	$\overline{RR}$	97.70	99.08	99.54	97.70	96.78	97.24
JMI	**#SF**	39	35	32	30	27	25
	RR	84.50	84.11	82.95	84.50	83.72	83.33
	$\overline{RR}$	100	99.54	98.17	100	99.08	98.62
TMI	**#SF**	36	32	29	26	24	23
	RR	82.95	83.72	83.72	83.72	81.40	81.78
	$\overline{RR}$	98.17	99.08	99.08	99.08	96.33	96.78

Secondly, we consider a classifier-dependent criterion which is based on the RR performance values. The criterion searches for the minimal number of features (#FT) that gives a greater or equal RR value than the value obtained for the total feature number (RR(39)). Table 7 shows #FT values, the corresponding normalized $\overline{RR}_T$. In order to study the curse of dimensionalty phenomenon, the maximum RR value is also pointed out for each strategy. So the feature number that gives the maximum RR value (#FM) is reported, as well as the corresponding normalized $\overline{RR}_M$ values. For both cases, α values to be applied on MI are also reported in order to make the link with MI criterion used in the classifier-independent case.

Table 7 Results of optimal number of features in the classifier-dependent case. The #FT value is the minimal number of features that gives a greater or equal RR value than RR obtained for the 39-total feature number. The $\overline{RR}_T$ values are the corresponding normalized rates with their α values. The #FM value is the number of features that gives the maximum RR value (with their $\overline{RR}_M$ and α corresponding values).

	#FT	$\overline{RR}_T$	α(#FT)	#FM	$\overline{RR}_M$	α(#FM)%
MMI	22	100.91	85.03	22	100.91	85.03
CMI	27	100.45	97.45	32	101.84	99.77
JMI	19	100.45	83.58	26	100.91	91.22
TMI	21	100.00	87.00	30	100.45	97.46

This criterion ensures a minimum number of features with performance improvement compared to those obtained in the case of 39 features. The JMI strategy gives the lowest number of features (19 parameters). Using the first MI criterion, this last result is obtained by taking α equal to 83%, which represents a drop of MI of 17% with respect to the MI_{max} value. Therefore, in order to ensure a minimum number of features that can achieve good performance results, it is necessary to take α values between 80 and 100. As previously mentioned, this variation can be explained by the estimation error of MI and the approximations of the MI proposed by the different strategies. The results show a maximum RR of 86.05% with 32 features (not reported in Table 7) using CMI strategy. This peaking value can be explained by the curse of dimenstionnality phenomena.
We conclude from this study that using the MI curve can inform about the minimum number of features without considering classifier performance. With the second RR criterion that requires more computing time, results demonstrate the best compromise of features reduction and RR improvement, particularly with JMI strategy.

4.5 Combination of different feature types

This section has principaly the aim to validate the proposed algorithm of optimal feature number estimation for the case of different feature types in large dimension. Furthermore, it has the aim to compare the relevance of different feature types for the identification of dominant feature types useful for explaining the emotion classes. In this work, we investigate common combinations of features (Pan, Shen, & Shen, 2012) which take into account spectral features with prosodic features extracted from a short-term analysis. In particular, each MFCC spectral feature vector is enriched with 12 LPCC and 12 PLP features with their 1st and 2d derivatives using the same signal preprocessing (suppression of silence of signals boundaries, filtering, windowing). This new vector forms

a vector of 108 spectral components (36 features for each type). The short-term prosodic features include the energy with their 1st and 2d derivatives (3 features) and the pitch (1 feature). The pitch is estimated at each 10 ms using Praat Software (Boersma & Weenink, 2018). Hence each signal is converted into a sequence of vectors, each of 112 components. In the training phase and testing phase, the same configuration of the SER system described in section 2 is taken with using 112 features (except in the following section for searching the number of Gaussians for each features combination) .

4.5.1 Performance study with features combination

In this section, we provide a comparative performance study using the different spectral and prosodic types of features described in the last section and taking different combinations of them. Table 8 shows recognition rates using different combinations of these types and also choosing for each case the number of Gaussians (between 1 and 256 with power of two progression) giving the best recognition rate.

Table 8 Recognition rates for different combinations of MFCC, LPCC, PLP feature types plus energy and pitch F0. As studied in section 4.1, the dynamic features DA are always included (either applied on prosodic or spectral features or both).

Feature types	*Prosodic*		*Spectral*		*Prosodic + spectral*							
Feature combinations	**E_DA**	**E_DA _F0**	**MFCC _LPCC _PLP**	**MFCC _LPCC _PLP _DA**	**MFCC _E_DA**	**LPCC _E_DA**	**PLP _E_DA**	**MFCC _LPCC _PLP _E_DA**	**MFCC _E_DA _F0**	**LPCC _E_DA _F0**	**PLP _E_DA _F0**	**MFCC _LPCC _PLP _E_DA _F0**
Recognition Rate	54.65	53.88	82.17	81.78	**84.50**	77.52	81.78	81.01	83.33	79.46	83.72	84.11
Number of Features	3	4	36	108	39	39	39	111	40	40	40	112
Number of Gaussians	64	32	128	128	128	128	128	64	128	128	128	128

From Table 8, we give the following points:

- spectral features alone give better performance results than prosodic features alone; this result confirms (Pan, Shen, & Shen, 2012) in which the bad score of prosodic features may be caused by the weak number of features (energy and pitch);
- in some cases, the combination prosodic + spectral features can improve performance by comparison with using only prosodic or only spectral features;
- the best combination between spectral and prosodic features is MFCC with energy and their dynamic features; it increases the recognition rate of 30.62% by comparison with the prosodic features only (E_DA_F0); adding LPCC and PLP features to the latter slightly worsens performance results which is probably caused by the redundancy (Pan, Shen, & Shen, 2012) and the large dimension of feature vectors giving rise to curse of dimensionality problems;
- the pitch F0 slightly improves performance results by combination with LPCC or PLP but not with MFCC or energy; this last result can be probably justifed by the feature redundancy.

In order to reduce dimensionality and probably improve performance, we give in the next section dimensionality reduction results from a set of 112 previoulsy described features using MI based feature selection startegies.

4.5.2 Performance study for 112-feature selection using MMI, CMI, JMI and TMI strategies

The purpose of this study is to try to select the most relevant features among a high dimensional space of 112 features composed of a ranked vector of MFCC_EDA (1-39), LPPC (40-75), PLP (76-111) and pitch F0 (112). The energy and their dynamic features are respectively numbered as 13, 26 and 39.

Table 9 Number of the first 10 selected features among 112 for each strategy.

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
MMI	**2**	77	**1**	76	**4**	**5**	79	87	**3**	78
CMI	**2**	**13**	**5**	76	40	**4**	**9**	86	42	80
JMI	**2**	**13**	76	**5**	77	40	**1**	**4**	**9**	79
TMI	**2**	**13**	76	40	**5**	**1**	**8**	41	**9**	83

Table 10 Recognition rate obtained for the first 10 selected features among 112.

	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}
MMI	42.64	46.90	46.90	49.22	51.94	58.53	58.14	61.24	69.38	66.67
CMI	42.64	48.06	50.39	58.91	61.63	64.34	65.89	69.77	70.93	65.12
JMI	42.64	48.06	52.71	58.91	58.14	60.08	62.02	63.57	65.89	68.22
TMI	42.64	48.06	52.71	56.98	61.63	57.75	65.12	66.28	68.99	68.99

From Table 9, it can be noticed that the static MFCC coefficients are mostly selected with at least 4 MFCC features among the first 10 selected features. The other selected features mostly come from the PLP type. This selection confirms the result given in (Trabelsi & Bouhlel, 2016) which shows the importance of MFCC type in the SER task. Note that features in PLP domain (77 and 76) that are similar in rank to features in MFCC domain (2 and 1) are selected together using MMI strategy but not using CMI or JMI strategies. This can be explained by the fact that the MMI strategy does not take into account the redundancy between the features.
Table 9 also shows that static features are always prefered than dynamic ones. The static relevance domination confirms the previous results obtained in Table 4 and by (Trabelsi & Bouhlel, 2016). Furthermore, the energy is the second feature selected by CMI, JMI and TMI strategies, which confirms the relevance of the prosodic feature type.
By comparing Table 10 with Table 5, RR performance results worsen using 112-feature selection procedure. This may be due to the limitations of the selection algorithms faced to many redundant features. Many reasons can probably explain such limitations: the selection strategies remain heuristic and take redundancies up to order 3 in MI computation; histogram binning procedure causes error accumulation of MI in high dimension.

4.5.3 Optimal number of features

Since the number of features is higher than in the preceeding study, which causes more error accumulation, largeur steps beween α values are considered. Results are given in Table 11 for only the CMI startegy.

Table 11 Results using CMI strategy of optimal number of features in the classifier-independent case. The #SF value is the minimum number that reaches α % of the MI_{max} value. The RR and $\overline{RR}$ values are the corresponding recognition rates with their normalized versions (w.r.t. RR(112)).

	α	100	95	90	85	80	75
CMI	#SF	81	53	43	36	30	25
	RR	82.56	85.27	82.56	83.33	80.62	80.23
	$\overline{RR}$	98.16	101.38	98.16	99.10	95.85	95.39

Whatever the α values, the RR values are above 80% like in Table 6. The optimal number of features decreases at the same time the α value decreases and it can reach dimensionnality reduction of 67.86% taking α equal to 85%. However this number is always higher by considering selection among 112 features than among only 39 MFCC_E_DA features. Note that the MFCC coefficients are known to be good decorrelators between variables and this property is not provided by LPCC and PLP. This explains some higher performance results when using MFCC features.

5 CONCLUSIONS

The aim of this study was to estimate the optimal number of selected features for the task of speech emotion recognition. We have investigated four selection strategies that select features sets according to their relevance, based on MI computation. The optimal number was then estimated as the minimum number of features using a criterion based on the maximum value of MI over the feature-selected sets. A comparison study was carried out using the recognition rate criterion for the optimal number estimation of features.

In this work, we have used the histogram approach to estimate MI values for its simplicity. However, this approach is faced to the binning problem that was discussed by taking several bin number choices such as Sturges, Scott and LMSE formulas. Results were obtained by carrying out an SER system based on a GMM classifier combined with features extraction step that takes 39-features vectors and secondly 112-features vectors. The features vectors were composed of the static of MFCC coefficients, the energy and their dynamic features Δ and ΔΔ. The SER system performance was evaluated using the EMO-db database. Also other results were obtained by taking large dimension vectors including spectral and prosodic features.
The results demonstrate that LMSE choice gives the best estimation of MI, which succeeds in approximately reaching the expected plateau in the MI curve. The study has shown that the MI based criterion gives acceptable performance results compared to the criterion based on the curves of recognition rates. Practically, the CMI

strategy combined with MI based criterion gives the high features reduction of 48.72% (from 39 to 20 features) and of 67.86% in a large dimension case (from 112 to 36) with a slight drop of performance results. On the other hand, the JMI combined with the RR criterion gives the best feature reduction of 51.28% (from 39 to 19 features) with performance improvements. However, this last result is classifier-dependent and requires very high computation capabilities.

We conclude that taking MI estimation with a good choice of bin number can help estimating the minimal number of relevant features for the task of SER, without taking into account classifier performance. This result is particularly interesting in high-dimensional systems.

A principal perspective is to extend this study for multimodal emotion recognition with speech and face modalities.

REFERENCES

Basu, S., Chakraborty, J., Bag, A., & Aftabuddin, M. (2017). A review on emotion recognition using speech. *International Conference on Inventive Communication and Computational Technologies (ICICCT)*, (pp. 109-114).

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer. [Computer program]. Version 6.0.37, www.praat.org.*

Brown, G., Pocock, A., Zhao, M.-J., & Lujan, M. (2012). Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *Journal of Machine Learning Research*(13), 27-66.

Burkhardt, F., Paeschke, A., Rolfes, M., & Sendlm, W. (2005). A database of german emotional speech. *INTERSPEECH- ICSLP*, (pp. 1517-1520). Lisbon.

Cover, T., & Thomas, J. (1991). *Elements of information theory.* New York: Wiley Series in telecommunications.

Drügman, T., Gurban, M., & Thirian, J. (2007). Relevant Feature Selection for Audio-visual Speech recognition. *International Workshop on Multimedia Signal Processing (MMSP).* Chania, Crete, Greece.

G Shashidhar: K, K., & Sreenivasa, R. (2012, June). Emotion recognition from speech: a review. *International Journal of Speech Technology, 15*(2), 99–117 .

Giannoulis, P., & Potamianos, G. (2012). A Hierarchical Approach with Feature Selection for Emotion Recognition from Speech. *the 8th International Conference on Language Resources and Evaluation (LREC'12)*, (pp. 1203–1206). Istanbul, Turkey.

Hacine-Gharbi, A., & Ravier, P. (2018). A binning formula of bi-histogram for joint entropy estimation using mean square error minimization. *Pattern Recognition Letters, 101*, 21-28 .

Hacine-Gharbi, A., Deriche, M., Ravier, P., Harba, R., & Mohamadi, T. (2013). A new histogram-based estimation technique of entropy and mutual information using mean squared error minimization. *Computers and Electrical Engineering, 39*(3), 918-933.

Huang, C., Gong, W., Fu, W., & F, D. (2014). A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM. *Mathematical Problems in Engineering*, 1-7.

Jain, A., Duin, R., & Mao, J. (2000, Jan). Statistical pattern recognition: a review. (IEEE, Ed.) *Trans. Pattern Analysis and Machine Intelligence, 22,*(1), 4-37.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1-2), 273-324.

Lanjewar, R. B., Swarup, M., & Patel, N. (2015). Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model (GMM) and K- Nearest Neighbor (K-NN) Techniques. *Procedia Computer Science, 49*, 50-57.

Mannepalli, K., Sastry, P., & Suman, M. (2018). Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University - Computer and Information Sciences*, in Press.

Manolov, A., Boumbarov, O., Manolova, A., Poulkov, V., & Tonchev, K. (2017). Feature Selection in Affective Speech Classification. *TSP*, (pp. 354-358).

Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. *INTERSPEECH- ICSLP*, (pp. 809–812). Pittsburgh PA, USA.

Pan, Y., Shen, P., & Shen, L. (2012). Speech Emotion Recognition Using Support Vector Machine. *International Journal of Smart Home, 6*(2), 101-108.

Pao, T., Chen, Y., Yeh, J., & Li, P. (September 2006). Mandarin emotional speech recognition based on SVM and NN. *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, *1*, pp. 1096-1100. Hong Kong, China .

Schuller, B., Rigoll, G., & Lang, M. (April 2003). Hidden Markov model-based speech emotion recognition. *Proceedings of the IEEE ICASSP Conference on Acoustics,Speech and Signal Processing*, *2*, pp. 1-4. Hong Kong, China .

Scott, D. (1992). *Multivariate density estimation: theory, practice and visualization.* New York: Wiley.

Sturges, H. (1926). The choice of a class interval. *J. Amer. Statis. Assoc.*, 65-66.

Sugi, C. V. (2016). Optimal Feature for Emotion Recognition from Speech. *African Journal of Basic & Applied Sciences, 8*(3), 136-144.

Trabelsi, I., & Bouhlel, M. S. (2016). Comparison of Several Acoustic Modeling Techniques for Speech Emotion Recognition. *International Journal of Synthetic Emotions (IJSE), 7*(1), 58-68.

Vijesh Joe, C., & Shinly Swarna Sugi, S. (2016). Optimal Feature for Emotion Recognition from Speech. *African Journal of Basic & Applied Sciences*, 136-144.

Wu, S., Falk, T. H., & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication, 53*, 768–785.

Yixiong Pan, P. S. (2012). Speech Emotion Recognition Using Support Vector Machine. *International Journal of Smart Home, 6*(2), 101-107.

Young, S., Kershaw, D., Odell, J., & Ollason, D. (1999). *The HTK Book.* Cambridge: Entropic Ltd.

Zaidan, N., & Salam, M. (2016). MFCC Global Features Selection in Improving Speech Emotion Recognition Rate. *Advances in Machine Learning and Signal Processing. Lecture Notes in Electrical Engineering, vol 387.* Springer.