International Journal of Speech Technology (2018) 21:167–183
https://doi.org/10.1007/s10772-018-9495-8

Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition

Shashidhar G. Koolagudi[1] · **Y. V. Srinivasa Murthy**[1] · **Siva P. Bhaskar**[1]

Received: 13 August 2017 / Accepted: 22 January 2018 / Published online: 1 February 2018

Abstract
In this paper, the process of selecting a classifier based on the properties of dataset is designed since it is very difficult to experiment the data on *n*—number of classifiers. As a case study speech emotion recognition is considered. Different combinations of spectral and prosodic features relevant to emotions are explored. The best subset of the chosen set of features is recommended for each of the classifiers based on the properties of chosen dataset. Various statistical tests have been used to estimate the properties of dataset. The nature of dataset gives an idea to select the relevant classifier. To make it more precise, three other clustering and classification techniques such as *K*-means clustering, vector quantization and artificial neural networks are used for experimentation and results are compared with the selected classifier. Prosodic features like pitch, intensity, jitter, shimmer, spectral features such as mel frequency cepstral coefficients (MFCCs) and formants are considered in this work. Statistical parameters of prosody such as minimum, maximum, mean (μ) and standard deviation (σ) are extracted from speech and combined with basic spectral (MFCCs) features to get better performance. Five basic emotions namely anger, fear, happiness, neutral and sadness are considered. For analysing the performance of different datasets on different classifiers, content and speaker independent emotional data is used, collected from Telugu movies. Mean opinion score of fifty users is collected to label the emotional data. To make it more accurate, one of the benchmark IIT-Kharagpur emotional database is used to generalize the conclusions.

Keywords Properties of dataset · Normality tests · Selection of classifier · Spectral and prosodic features · Jitter · Shimmer

1 Introduction

The design principle of classifiers is well defined based on certain mathematical criteria. The classifiers are expected to perform better for specific kind of data. For instance, the data that follows Gaussian distribution is clearly and better classified by the classifiers such as Gaussian mixture models (GMMs). However, at present, the selection process of classifiers is happening blindly and experimenting the datasets with all the classifiers. It is a time consumption operation as well as not an appreciable task. The process of identifying a relevant classifier for a particular dataset solves many advantages such as computational complexity, performance improvement and so on. With this motivation, an effort has been made in this work to identify the properties of datasets before choosing the classifiers to work on them. As a case study, the task of emotion recognition has been selected. Emotion recognition from speech has been one of the challenging tasks due to its ambiguity. Many times even humans cannot recognize emotions correctly. Emotion recognition task helps to identify the emotional state of a human being from their voice. Automatic emotion recognition from speech has many applications such as making human–machine communications practical and more interactive, patient monitoring, telephone-based customer service systems, psychological health care initiatives and so on (Reddy et al. 2011). Some of the important basic emotional states are anger, fear, happiness, sadness, neutral, boredom, surprise, disgust, etc. Many factors influence the difficulties in modeling emotions effectively. The factors include lack of a proper emotional database, identification, and extraction of

✉ Y. V. Srinivasa Murthy
urvishnu@gmail.com

Shashidhar G. Koolagudi
koolagudi@yahoo.com

Siva P. Bhaskar
sivabhaskar3@gmail.com

[1] Department of CSE, National Institute of Technology Karnataka, Mangalore 575 025, India

proper features which discriminate emotions clearly and so on. The highly varying modulation in the emotional speech of different persons makes the speaker-independent emotion recognition more difficult.

Speech signal mainly contains information about linguistics, speaker's identity, emotion, and so on (Nwe et al. 2001). Of these, emotion is one important attribute which provides naturalness to speech. The emotion can be estimated from the speech, text, and facial expressions. Anyhow, it is hard to determine the emotions from text due to the ambiguities at syntactic and semantic levels. Moreover, it has been stated that all the emotions cannot be identified from facial expressions. Hence, speech is the one possible reliable source to recognize the emotions effectively when compared to text and image. In general, the process of speech emotion recognition can be done by extracting features and using classification models. Several features and an *n*—number of classification models are already proposed in the literature for the task of emotion recognition. However, all the features may not be useful and it is important to identify the suitable features for the present task. Here, the information related to text is not taken into consideration to make the investigation independent of that factors. The features that are found in the literature of emotion recognition are inherited from the features of speech tasks. The approach for identifying a suitable set of features for any classifier is very essential. Moreover, it is also important to recommend a classifier based on the properties of a feature set is also important since the performance of the emotion recognition systems (ERSs) also highly depends on classification models (Reddy et al. 2011). The performance of the classifier equally depends on quality and properties of the dataset used. Most of the research works focused on selecting the input features based on the task. However, while selecting features for any speech task, classification model should also be taken into consideration. It is practically not possible to get the better performance with all the classification models for the given dataset.

In this work, a metric has been introduced to understand the properties of the given speech emotional dataset. The features that are relevant to speech emotion have been extracted as a prior step of selecting the classifier. The combinations of computed features have been tested with the classifier to select the emotion-specific features. Further, the statistical techniques have been applied to the best possible combinations to determine the nature of the dataset. The relevant classifier has been suggested based on the results of statistical analysis. The results have been compared with two clustering algorithms and one more classification model to defend the proposed classifier. The models that are explored in this work are vector quantization (VQ), *K*-means clustering, GMMs and artificial neural networks (ANNs). Spectral features like MFCCs and formants; prosodic features such as pitch, intensity, jitter, and shimmer have been computed from the speech signals due to their ability in discriminating emotions. It has been observed that the emotional data falls in the normal distribution and hence, GMM has been suggested since it outperforms for the datasets that are normally distributed. In addition to that, IIT-KGP benchmark emotional database is also used to generalize the conclusions.

Rest of the paper is organized as follows. Detailed literature of related work is discussed in Sect. 2. Section 3 gives complete information about proposed approach in detail including database collection, feature extraction, subset construction and classification models. Results have been provided with analysis in Sect. 4. Section 5 concludes the paper with some future research directions.

2 Related work

In this section, the features and classifiers used for emotion classification are discussed in brief. Though different feature sets have specific statistical properties, pieces of evidence are not found to use those properties while deciding the classifiers. The existing literature on the task of emotion recognition is detailed below:

2.1 Feature selection

To automatically distinguish emotions from the speech signal, it is important to identify relevant features. The combination of features also plays an important role in improving the efficiency of recognition system. From the literature, emotion recognition (ER) related features are broadly categorized into (a) prosodic, (b) spectral, (c) combinations of (a) & (b) and (d) multi-modal features. The following subsections will descibe the importance of these features in recognizing emotions.

2.1.1 Prosodic features

Prosodic is an adjectival form of prosody (Nooteboom 1997). It is believed that, emotional state of speaker is primarily indicated by prosody (Ortony 1990; Chen et al. 2006). There are mainly three categories of prosodic features (i) pitch, (ii) intonation and (iii) intensity. It is not useful to derive these features at frame level, hence they are extracted at syllable, word, utterance and sentence levels (Koolagudi and Rao 2012). Pitch (*also referred as* Fundamental frequency) is a rate of vocal folds region vibration. It mainly depends on air pressure at subglottal and tension of vocal folds. Hence, it is one of the important features which carries emotion specific information (Ververidis and Kotropoulos 2006). There are many approaches to estimate the pitch value if signal is quasi-stationary (Hess 2008). Studies in Iida et al. (2003)

stated that high recognition of emotions can be achieved using autocorrelation based pitch value. In some cases intervention of first formant may affect the fundamental frequency. It is easy to recognize voiced segments from speech using energy criteria as it is high at voiced regions compared to unvioced ones. Overall amplitude, energy distribution in spectrum and duration of pauses are affected with the arousal state of a speaker (Scherer 1999, 2003; Cowie and Cornelius 2003). Hence, energy and duration featuers are useful in recognizing emotions. In general, energy level in males is higher compared to females in anger state (Heuft et al. 1996). For the same state speech rate of female speakers is high compared to males (Iida et al. 2000). Pitch, energy and speech rate is less in the state of disgust. High pitch value and high intensity values are reported in the state of fear. Compared to neutral state speech rate is less in sadness (Ververidis and Kotropoulos 2006). In case of sad emotions male speakers speech rate is high compared to the anger (or) disgust states (Cowie and Cornelius 2003; Iida et al. 2000). Statistical values of prosodic features are also useful in characterizing emotions efficiently (Chung-Hsien and Liang 2011; Rao et al. 2013). Statistical values of pitch include range, minimum, maximum, mean, standard deviation, median, slope maximum, slope mininum, relative pitch, skewness, kurtosis, 4th order legendre parameters, first order difference (Δ), jitter and so on. Similarly for energy and duration statistical features like shimmer, speech rate, duration of voiced to unvoiced sounds' ratio along with mean, minimum, maximum, standard deviation are considered. In Rao et al. (2010), dynamics of prosody features such as pitch, energy and duration contours are used as features to recognize seven emotions. Analysis is done on individual features with the classifier support vector machines (SVM). Statistical values of pitch and energy are used as features in Bhatti et al. (2004; Schuller et al. 2003; Luengo et al. 2005) to recognize emotions from speech. Different classifiers like modular neural network (MNN), GMM and Hidden Markov Models (HMMs) are used for the same. In Rao et al. (2013), characterization of eight emotions of IITKGP-SESC corpus through signal analysis at syllable, word and utterance levels of speech segments is done. Local and global features are extracted at these levels. Critical analysis is done individually and as a group of features for both male and female speakers. With the combination of local and global features, improved recognition rate is reported for last syllables in final words. In Jawarkar et al. (2007), statistical features of pitch and energy are extracted to recognize four emotions. Fuzzy min–max neural network (FNN) is used as classifier and it is reported that it requires less time to learn compared to back propagation neural network (BPN).

2.1.2 Spectral features

The shapes of the vocal tract system are unique for different emotions, and the shape of the vocal tract can be estimated by using spectral analysis. Hence, spectral features are also useful to categorize the emotions (Rao et al. 2013; Bitouk et al. 2010). In general, the spectral features are computed by dividing the speech signal into small segments (*called* frames) of length 20–50 ms. The speech signal is assumed to be stationary in the specified length. Studies in Banse and Scherer (1996; Kaiser 1962; Nwe et al. 2003) reported that the value of energy is high in the state of happiness where it is low in the case of sadness. The high and low energy values are observed in the high-frequency regions. There are different approaches to extract the spectral features. Some popular techniques include traditional linear predictor coefficients (LPC), one-sided autocorrelation linear predictor coefficients (OSALPC), short-time coherence method (SMC) and least-squares modified Yule-Walker equations (LSMYWE) (Rabiner and Schafer 1978; Hernando et al. 1997; Le Bouquin 1996; Bou-Ghazale and Hansen 2000). In Chauhan et al. (2010), LP residual is used as a feature to design emotion recognition system (ERS). The auto-associative neural network (AANN) and GMMs are found to be good and used as classifiers for a majority of the tasks mentioned above. Further, the sequence of glottal pulses is considered as the excitation source of voiced speech (Ananthapadmanabha and Yegnanarayana 1979). The glottal closure instance (GCI) (also known as Epoch) is highly helpful for estimating the pitch as well as vocal tract frequency response. The epoch information has been extracted using the approaches such as LP residual and zero frequency filtered (ZFF) speech signal to recognize the emotions from IITKGP—simulated emotion speech corpus(Koolagudi et al. 2010). GMMs and SVMs have been used as classifiers and found that the GMM is highly compatible when compared to SVMs. Moreover, human perception of pitch may not always follow a linear scale. Hence, some approaches have been introduced to estimate the non-linear scales using Bark scale, Mel-frequency scale, modified Mel-frequency scale, and ExpoLog scale (Rabiner and Juang 1993; El Ayadi et al. 2011). The log magnitude spectrum has been computed for the same, also known as Cepstral analysis. The cepstral analysis has been done to extract the features like linear predictive cepstral coefficients (LPCC), Mel-frequency cepstral coefficients (MFCC), one-sided autocorrelation linear predictive cepstral coefficients (OSALPCC), and so on. The process of detecting the stress from the speech signal using non-linear scale is found to be better when compared to linear scale analysis (Bou-Ghazale and Hansen 2000). The Mel-energy spectrum dynamic coefficients (MEDCs) are extracted based on spectral energy dynamics to recognize the emotions of both male and female speakers (Lee et al. 2004). In some other works, the Mel frequency based short time speech power

coefficients (MFSPCs) are extracted, and VQ based HMM is used as a classifier to recognize emotions (Nwe et al. 2001).

2.1.3 Combination of prosodic and spectral

It is found that the temporal information is missing with the short-time features such as MFCCs and Perceptual linear predictive (PLP) values. The temporal information is very useful while estimating the emotions (Siqing et al. 2011). Based on this, the modulation spectral features (MSF) are introduced to estimate the temporal information in the speech signal that further helps to determine the emotion (Razak et al. 2005). The combination of prosody and spectral features are also used to improve the efficiency while recognizing emotions from speech (Nicholson et al. 2000). Critical Analysis has been done with open and closed tests on both male and female Japanese speakers database. The results stated that if the number of speakers is increasing classification performance is decreasing with the combination mentioned above (Li and Zhao 1998). The filter bank coefficients from 300 to 3400 Hz are used to extract standard MFCCs. In the case of low-MFCCs, filter banks are applied in the range of 20–300 Hz frequency regions to model fundamental frequency (F0) variations. The features such as MFCCs, low-MFCCs, pitch, and Δ pitch are found to efficient to improve the performance of the ERS (Neiberg et al. 2006). GMM is used as a classifier to recognize emotions. It is reported that low-MFCCs perform well in extracting stable pitch. Further, some analysis has been done with the combination of prosody and short-time with rough sets (Zhou et al. 2006). At the outset, it has been stated that the temporal variations play a major role in discriminating emotions.

2.1.4 Multi-modal features

Human feelings can be expressed by using tone, gestures, facial expressions, key spotting techniques and so on. The initial efforts are done with facial analysis to identify the human emotions (Black and Yacoob 1995; Essa and Pentland 1997; Kenji 1991; Tian et al. 2000; Yacoob and Davis 1994) and also auditory voice (Ververidis and Kotropoulos 2006; El Ayadi et al. 2011; Krothapalli and Koolagudi 2013) individually. Further, the facial expressions and voice have been combined to improve the accuracy (Busso et al. 2004; Schuller et al. 2004). The room is open to work on the multimodal features.

2.2 Classifier selection

There are a *n*—number of classifiers such as HMM, ANN, GMM, SVM, VQ, k-NN and so on that are used for the task of emotion recognition from speech.

Studies in El Ayadi et al. (2011), Womack and Hansen (1999), Lee and Hon (1989) state that the majority of the previous works have been focused with HMM as a classifier to recognize emotions from speech due to its popularity and efficiency in various speech applications. The phonemes are extracted and modeled using HMM in the case of automatic speech recognition (ASR) applications (Deller et al. 2000). The state transition matrix is useful to capture the temporal dynamics in speech signal (Rabiner 1989). Since the phonemes follow left-to-right sequence in speech, HMM usually adopts the left-to-right structure to recognize speech. The same phenomenon has been used to recognize the emotions from speech (Schuller et al. 2003; Kwon et al. 2003; Nogueiras et al. 2001; Polzin and Waibel 1998; Bitouk et al. 2010; Sato and Obuchi 2007). However, it is not possible to observe the sequential flow of emotional cues incorporated in an utterance. For instance, it is difficult to fix a time for the pause which appears in an utterance of sad emotion. It may appear at any place such as in the beginning, middle or end events in an utterance (Yamada et al. 1995). The concept of ergodic model HMM is considered as a classifier to overcome this problem (Nwe et al. 2003). In this model, it is possible to reach from any single state to any other in a single step. However, none of the works explains the reasons for choosing HMM for a given dataset.

For data density estimation, a probabilistic model GMM is designed. GMM is considered as the state-of-art classifier and mostly used in the tasks of speaker identification and verification (Reynolds et al. 2000). It provides flexible-basis representations to model diversified data with large dimensions (Li and Barron 1999; Vlassis and Likas 2002). It can be treated as a special case of continuous single state hidden Markov model (Douglas and Richard 1995). In general, second-order parameters like mean and standard deviation are used in GMM to capture the hyperplane distribution of data points (Koolagudi et al. 2010). In the case of multimodal distributions, GMMs are found to be an adequate and minimal train, and test sets are sufficient compared to normal continuous HMMs (Bishop 1995). Hence, GMMs are more appropriate in the case of global features which are extracted from the speech signal to recognize emotions. One of the limitations with this model is difficulty in modeling the temporal structure of the training data due to the independent structure of feature vectors. It is also a challenging attempt to decide the optimum number of components. Modelling order section principles such as minimum description length (MDL) (Rissanen 1978), classification error with respect to a cross-validation set, goodness of fit (GOF) based on kurtosis (Vlassis and Likas 1999) and Akaike information criterion (AIC) (Akaike 1974) are the common approaches to decide the optimal number of components (El-Yazeed et al. 2004). To estimate both the model order and components together expected maximization (EM) algorithm is designed

which is based on greedy approach (Vlassis and Likas 2002). GMM is widely used in Neiberg et al. (2006), Ververidis and Kotropoulos (2005), Yang and Lugger (2010) to recognize emotions in speech. Tang et al. proposed a boosted GMM to recognize emotions (2009).

Vapnik and Chervonenkis utilized the concepts of statistical learning theory to introduce a new classification and regression technique, and the result of this effort is SVM (Burges 1998; Wang 2005). It mainly uses kernel functions to map the non-linearity in the feature set to the large dimensional feature space through which the linear separation can be obtained. In various pattern recognition applications, SVM is found to give better results when compared to many other classifiers (Shen et al. 2011). Especially 75–80% of classification rate is obtained in the case of speaker independent applications using SVM classifier (Zhou et al. 2006). At the outset, SVM classifier is constructed for two classes, and it is possible to reduce the classification error of test samples through finding the optimal hyperplane. Several methods were introduced to use SVM for multi-class classification. Among those the one-vs.-all Method is developed and used in Takahashi (2004) to recognize emotions. The LIBSVM is considered to classify five emotional states using the Mel energy spectrum dynamics coefficients (MEDC) feature vector (Lin and Wei 2005). Since SVM is good at classifying two classes, to discriminate n classes Zhou et al. combined it with bin-tree (Zhou et al. 2006). SVM classifier is used to recognize emotions from the speech signal in some of the works but found that the accuracy is not markable (Grimm et al. 2007; Seehapoch and Wongthanavasu 2013; Yu et al. 2011; Pan et al. 2012; Chavhan et al. 2010). The performance of an ERS with SVM is around 80.09% for the gender and situation independent database (Giannoulis and Potamianos 2012; Muthusamy et al. 2015).

The other efficient classifier to capture the non-linear relations and used in various pattern recognition applications is ANNs. It has some significant advantages compared to other classifiers. In the case, if training samples are lesser in number, then classification performance is usually better compared to HMMs and GMMs. Multi-layer perceptron (MLP), radial basis function networks (RBF) and recurrent neural networks (RNN) are the main categories of ANNs (Bishop 1995). Among those, MLP is the one which is commonly used in emotion recognition applications, and RBF is the least used (El Ayadi et al. 2011). MLP is easy to implement and it is built with well-defined training algorithm. ANN performance totally depends on its parameters such as the number of hidden layers and number of hidden neurons in each hidden layer. More than one ANN is used in some speech emotion recognition applications to achieve better performance (Nicholson et al. 2000), (Firoz et al. 2009), (Dai et al. 2008). ANN is used in Petrushin (2000) to distinguish agitation and calm emotional states. Generalized discriminant analysis (GerDA) is introduced in Stuhlsatz et al. (2011) for the task of acoustic emotion recognition using deep neural networks (DNN). Better performance is reported with DNNs compared to SVMs for this task (Han et al. 2014). In Khanchandani and Hussain (2009) MLP and generalized feed forward neural networks (GFNNs) are used to recognize emotions in speech signal and their results are compared. Results stated that GFNNs perform better compared to MLPs. Auto associate neural networks (AANNs) are used in Koolagudi and Rao (2012) to recognize basic emotions in semi natural speech. 2D-neural classifier is used in Partila and Voznak (2013) for classifying the emotional state of a man's voice.

The classifiers such as HMMs, SVMs and ANNs are found to give less if the number of samples is small. VQ is introduced to provide better recognition in such case (Huang et al. 2012). In VQ a fixed size-quantized vector V_i is created with dimension n for the vectors V of the same dimension. If all the components of V and the corresponding components of V_i are close enough, then V_i is treated as a quantized vector of V (Konar and Chakraborty 2014). Learning vector quantization (LVQ) technique is used in several facial emotion recognition applications (Konar and Chakraborty 2014). Variance-based Gaussian Kernel Fuzzy Vector Quantization (VGKFVQ) method is proposed in Huang et al. (2012) to recognize emotions in short speech. In Khanna and Kumar (2011) LBG-VQ method is introduced to recognize human emotions.

There are some works available to suggest a classifier by experimenting feature vector on all available candidate algorithms (Demšar 2006). They are also called *meta-learning* (Soares and Brazdil 2000; Muslea et al. 2006). Computational complexity is the important issue while developing meta-learning systems. All the classification algorithms have to be tested with the given dataset, and it may be the reason for complexity issues. In contrast, the dataset is analyzed, and suitable classifiers have been suggested in this work. In literature, there is no specific strategy to decide the suitable classifier for the task of speech emotion recognition (El Ayadi et al. 2011). Their advantages and limitations may confuse researchers to choose the best one.

Moreover, there are several other approaches have been proposed to categorize the emotions. Since the feature vector contributes much while categorizing the emotions, the approaches towards estimating the relevant features are essential. The technique called incomplete sparse least square regression (ISLSR) has been proposed to select the features that can highly contribute to categorize the emotions of six classes (Zheng et al. 2014). A novel set of features based on Fourier parameter model are also proposed with their derivatives to categorize the emotions. However, if the number of emotional classes increases, then the performance is getting degraded (Wang and An 2015). The parameters

Table 1 Summary of source, features and classifiers used in existing work to recognize emotions from speech

Category	Feature set	Vector size	Database	Classifier	Recognition rate (%)	Remarks	Ref.
Prosodic features	stat{pitch + amplitude}	17	English, Chinese Urdu and Indonesian	MNN	80.69	Language independent and implemented for four languages.	Bhatti et al. (2004)
	stat{pitch + energy}	20	German	HMM	64.7	Language specific ERS.	Schuller et al. (2003)
	stat{pitch + energy}	72	Basque	GMM	84.7	Feature vector size and number of mixtures used for GMM may increase computational complexity	Luengo et al. (2005)
Spectral features	MEDCs	39	Danish (DES)	SVM	88	MEDCs are introduced and better performance is reported.	Lin and Wei (2005)
	MFSPCs	12	Burmese	VQ based HMM	60.1	Database is created only for two speakers.	Nwe et al. (2001)
	LP residual	40 samples	IITKGP-SESC	AANN GMM	56	It is stated that LP residual contains emotion specific information.	Chauhan et al. (2010)
	Epoch parameters	4 values	IITKGP-SESC	GMM SVM	60.6	Better performance is achieved even signal has a noisy behavior	Koolagudi et al. (2010)
Combo. of prosody & spectral	stat{log energy + cepstral coefficients+ pitch}	3809	Berlin (EMO-DB) English (SUSAS)	SMO with RBF	78	Review is done on different classifiers for ERS using WEKA on selected features.	Vogt et. al. (2008)
	MSFs + prosodic	41	Berlin (EMO-DB) VAM DB	SVM	91.6	3.1% recognition rate is improved with MSFs than MFCCs and PLPs.	Wu et. al. (2011)
	LPCCs + pitch + energy + duration+jitter	18	Malay English	Fuzzy sets	60	Gender dependent ERS.	Razak et. al. (2005)
	MFCCs + Δ MFCCs stat{F1+F2}	39	Semi-professional female actress	HMM	88	Gender depended ERS.	Lee et al. (2004)
	speech power + pitch + LPC + ΔLPC	26	Japanese	ANN	50	Speaker and contet independent ERS.	Nicholson et. al. (2000)
	stat{acoustic features}	100	English speakers English movies	SVM	71.62	Speaker and content independent ERS.	Schuller et. al. (2005)
	Pitch + energy + spectral features	11	Spanish	HMM	70	Speaker dependent ERS.	Nogueiras et al. (2001)
	stat{pitch+formants + log energy}	20	English	GMM	62	Content independent ERS.	Li and Zhao (1998)

Table 1 (continued)

Category	Feature set	Vector size	Database	Classifier	Recognition rate (%)	Remarks	Ref.
	MFCCs + low-MFCCs + ΔMFCCs + pitch + Δpitch	42	Swedesh VP and English ISL meeting corpus	GMM	79	Content independent ERS.	Neiberg et al. (2006)
	Pitch + energy + LPCCs+MFCCs	27	Real life data	GMM	75	Review is done with different classifiers on same data.	Ingale et. al. (2012)
	stat{pitch + energy + formants}	37	Real life data	SVM	74.75	Statistical values of prosodic features are useful in recognizing emotions.	Zhou et. al. (2006)
	GFCCs + jitter + shimmer	14	human speech and animal voice	HMM with 4-mixture GMMs	69.1 and 82.7	82.7% recognition rate in case of caller dependent. Jitter and shimmer are useful to identify arousal levels in human & animal voices	Li et. al. (2007)
Other methods	ISLSR based features	648 and 487	eNTERFACE DB FAU-AIBO	ISLSR	69.33 and 60.50	A method to select the relevant features for the task of SER has been proposed.	Zheng et. al. (2014)
	UNIVERSUM	384	ABC GeWEC EMODB SUSAS	Autoencoder based Adaptation model	63.30 (UAR)	An auto encoder based unsupervised adaptation model called Universum has been proposed to recognize emotions.	Deng et. al. (2017)
	DNN	988	IEMOCAP	ELM	0.543 (WAR)	The technique of deep neural networks (DNNs) has been used to extract the features from raw signal and fed to extreme learning machine (ELM).	Han et al. (2014)

called time-lapse and linguistic information have been considered as knowledge information and used to estimate the emotions from the spontaneous speech recorded from call centers (Chakraborty et al. 2016). Few models based on auto encoder based unsupervised domain adaptation technique are also proposed for the task of speech emotion recognition (Deng et al. 2014, 2017). They constructed the model without using any label information. However, the dimension of the feature vector is large and severe complexity issues may raise. A modified brain emotional learning model is also proposed to categorize three emotional classes (Motamed et al. 2017). A majority of the works done for speech emotion recognition have focused on estimating the differences between datasets instead of differentiating different corpora (Song et al. 2014). Hence the concept of non-negative matrix factorization (NMF) and transfer NMF have been considered for emotion recognition (Song et al. 2016). Further, the task has been extended to extract the features by directly feeding the raw speech signal to deep neural networks (DNNs). A few works have been done by extracting the features from DNNs and passing them to various classifiers. One such classifier is extreme learning machine (ELM) (Han et al. 2014; Trigeorgis et al. 2016). The summary of literature including features, classifiers, database, and remarks has been given in Table 1. It is found that the majority of above works have experimented on high dimensional feature vectors that generally lead to the computational issues. Hence, an approach with relevant optimal feature vector and the suitable classifier is always essential. In this paper, the properties of the dataset have been estimated to suggest the classifier.

Various pieces of evidence from statistical analysis have been taken from an emotional database which is collected from Telugu movies and people who speak *Telugu* language to determine the suitable classifier for emotion recognition. For the selected case-study, two clustering and two classification techniques such as VQ, k-means clustering, GMM, and ANN have chosen based on their relevance. The SVM is found to give less performance in the case of speech emotion recognition. Hence, SVM and other random forest classifiers have been ignored in this work due to their less performance and complexity issues. Critical analysis is done with different combinations of features and by modifying the parameters of classifiers. Normalization tests namely Kolmorgov–Smirnov, Shapiro–Wilk and Mardia's tests are performed to observe the distribution of data.

3 Proposed methodology

The proposed flow diagram is shown in Fig. 1. The semi-natural emotional clips collected from Telugu movies are used as database. In addition to that, standard emotional speech

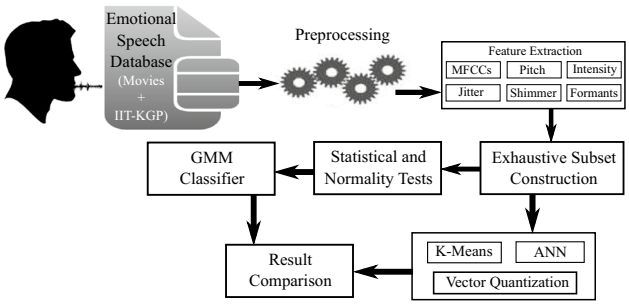

Fig. 1 The proposed framework to determine the classifier based on the properties of dataset

database of IIT-KGP has been considered. After preprocessing, spectral and prosodic features are extracted from the speech clips as they are effective in modelling emotional speech. Exhaustive subset is constructed to identify the task specific features. Different statistical tests are conducted to study the properties of dataset and a suitable classifier is suggested. The results of three classifiers are compared with the suggested classifier to evaluate the proposed classifier. The following subsections elaborate the each block in detail.

3.1 Database collection

The proper and complete emotion database is essential for efficient modeling of emotions. There are several ways of collecting emotional database and the ideal way is to collect the speech data from natural conversations since these include real emotions. However, recording sufficient amount of such natural conversations with a good quality is extremely difficult task. So movies are preferred for collecting sound clips that contain emotions assuming that the actors express emotions in a precised natural way. It is also called as semi-natural database.

Five basic emotions such as anger, fear, happiness, neutral, and sadness are considered for this work. All the emotional clips are collected from Telugu movies. Care has been taken to include different genders, context independent speech and ignored the overlapping of voices. Speech samples are collected at high sampling frequency of 44.1 KHz and later decimated to 16 KHz. It is done because the emotion specific information will be retained with in 8 KHz, to avoid the complexity issues and to consider Nyquist theorem (Rao and Koolagudi 2012). For each clip, the emotion is labelled based on mean opinion score (MOS) collected from fifty users. The users of Telugu linguistic base are selected to collect MOS. In addition to that, contextual information is also considered. Total 1000 clips of length 2–5 s are considered. Of these, fifty sophisticated clips have been used to understand the properties of data. Later, the same technique is applied to all 1000 clips. To generalize the conclusions,

comparison has been done on one of the standard speech emotion database of IIT-KGP (Koolagudi et al. 2009).

3.2 Feature extraction

From the literature, it is observed that spectral and prosodic features are best suited for emotion recognition (Zhou et al. 2009). It is also true that the variations in prosody helps to model emotional speech. Hence, an analysis has been done in this aspect and it is found that the statistical variations of pitch such as minimum, maximum, mean (μ) and standard deviation (σ) are also prominent to determine emotion from human speech. Therefore the same features are used in this work along with jitter and shimmer. Usually, speech signal is assumed to be stationary for the purpose of analysis over a short duration of time. Hence the spectral features are extracted from a frame of around 20 ms, as variation in speech signal within 20 ms. is ignorable. An overlap of 50% is considered to retain the continuation. 13 MFCC features are extracted at frame level. Average of all frame wise MFCC feature vectors of an utterance is calculated to represent utterance level spectral feature vector. The detailed feature extraction process is explained below.

3.2.1 Mel freqency cepstral coefficients (MFCCs)

MFCCs are most widely used features for almost every speech task. This is due to their ability to imitate human auditory perception mechanism. They are derived based on the characteristics of the human hearing system. These features are derived from a mel-frequency cepstrum where the frequency bands are non linearly spaced based on the mel-scale. General block processing approach was used for extraction of MFCCs (Wenjing et al. 2009).

3.2.2 Pitch

Pitch is the fundamental frequency (F0) of vocal folds' vibration. Speech signal is observed to be periodic for this reason. There are many methods to estimate pitch of the speech. In this work, auto-correlation method is used for pitch estimation. Initially, pitch analysis have been done for both the genders to recognize the gender of a speaker. Further, the analysis for emotion recognition is done based on the gender. The average pitch observed from the database for females is about 210 Hz whereas the same for males is about 120 Hz (Traunmüller and Eriksson 1995).

Pitch is one of the important attribute which adds naturalness to speech. Pitch contains information about emotion, gender, accent, speaking manner and so on (Wenjing et al. 2009). Pitch of a speech expressing one emotion is usually different from the other one (Kostoulas and Fakotakis 2006). Thus, it can be used as discriminating feature in emotion recognition. Statistical parameters of pitch contour like minimum, maximum, mean, standard deviation are used as features at utterance level. These statistical measures are derived from pitch values of all frames of an utterance.

3.2.3 Intensity

Intensity refers to the volume or energy of speech signal. Energy depends on the loudness of the voice of the speaker. The speaker who speaks louder enforces higher energy in the signal than that of the one who speaks mutedly (Fu et al. 2008). As the average energy of complete signal do not give any information w.r.t. emotions, short time energy is computed for a frame of length 20 ms. shown in Eq. (1) (Anagnostopoulos et al. 2012).

$$E_n = \sum_{m=n-N+1}^{n} [x(m)w(n-m)]^2 \tag{1}$$

where E_n is the energy value, N represents the length of the frame, $w()$ represents analysis window which can be rectangular or hamming and n is the sample where the analysis window is focused.

It has been reported that anger, fear and happiness have high intensity values than that of sad (Scherer 2003). So intensity can also be considered as one of the emotion discriminating features. Statistical measures of intensity such as minimum, maximum, mean (μ), standard deviation (σ) are computed as features at utterance level.

3.2.4 Jitter

Pitch period slightly changes over consecutive pitch cycles. This variation of pitch period over time, depends on many factors like text, intonation, emotional state of the speaker etc. Hence, cycle-to-cycle pitch variation is computed, also known as *Jitter* (Farrus and Hernando 2009). It is given by

$$J = \frac{1}{N-1} \sum_{i=1}^{N-1} |T_i - T_{i+1}| \tag{2}$$

where T_i is the extracted pitch period and N is the number of cycles considered. The same is shown in Fig. 2.

3.2.5 Shimmer

It is possible to observe the energy variation when there is high emotion (Farrus and Hernando 2009). The variations in energy have been clearly observed using the shimmer feature.

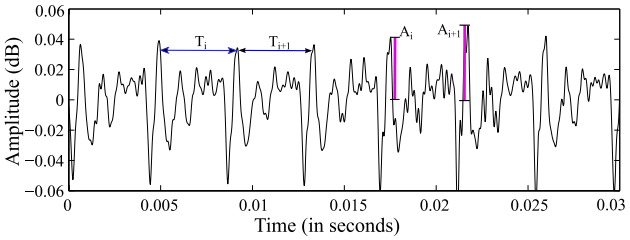

Fig. 2 The proces of extracting Jitter and Shimmer

Shimmer is the parameter which represents the variation in the amplitude of samples between the adjacent pitch periods. It is given by

$$Shimmer = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(\frac{A_{i+1}}{A_i}\right)\right| \tag{3}$$

where A_i is the extracted peak-to-peak amplitude data and N is the number of extracted pitch periods. The process of extracting shimmer is given in Fig. 2.

3.2.6 Formants

Formants correspond to the resonances of human vocal tract system. At each formant frequency, there exists a high degree of energy. They are mostly observed in the signal corresponding to vowels (Anagnostopoulos et al. 2012). Formants are the features which depend on vocal tract characteristics. These vocal tract characteristics change as the emotions change. Hence formant features are explored for discriminating emotions (Koolagudi et al. 2009). In this work, first three formants (F1, F2, F3) are considered.

3.3 Subset construction

All the above specified features are concatenated at utterance level to form a complete feature vector of length 26 in the order MFCCs (13), prosodic: pitch (4), intensity (4) and others: jitter (1), shimmer (1), formants (3). It is also true that all features may not be useful for the specified task. It is very important to identify the relevant and suitable features, known as feature selection. There are several feature selection algorithms already available to reduce the dimensionality of feature vector (Liu and Lei 2005; Tang et al. 2014). In contrast to them, in this work, an exhaustive search has been done by testing the accuracy for all feature combinations with four classifiers such as GMM, VQ, *K*-means and ANN. As MFCCs are considered as baseline features for this work it is considered as single feature. By considering MFCCs as single feature vector 2^{14} feature subsets are generated and tested against each classifier. Out of them, the best five combinations for each classifier has been shown in Table 2.

3.4 Analysis to understand the properties of featureset

The featureset is analyzed critically by applying various statistical and normality tests to figure out the properties of the dataset. The best subset which is giving better results for all the four classifiers is identified and considered for analysis. As MFCCs are considered as baseline features for the selected case study, they are excluded from this step. From Table 2, it is observed that GMM is giving best performance with the combination of MFCCs, pitch (min, max), intensity (max), jitter, shimmer, F1 and F3 features. Based on this, the same features excluding MFCCs are considered for normality tests. Three different standard tests such as, K–S test, S–W test and Mardia's test are used in this paper. The detailed explanation for each test is given in the subsequent subsections.

3.4.1 Kolmogorov–Smirnov test

Kolmogorov–Smirnov test is used to assess the similarity between empirical cumulative distribution function (ECDF) of the sample space and cumulative distribution function (c.d.f.) of the reference distribution. It is a non-parametric test, where one dimensional probability distributions can be used to compare a sample with reference probability distribution (Lilliefors 1967).

The ECDF $\hat{F}_k$ for k data samples (Y_j) is defined in Eq. (4)

$$\hat{F}_k(y) = \frac{1}{k}\sum_{j=1}^{k} I_{Y_j \le y} \tag{4}$$

where $F(y)$ represents the c.d.f, $I_{Y_j \le y}$ is the indicator function which equal to 1 if $Y_j \le y$ and 0 otherwise.

3.4.2 Shapiro–Wilk test

To test whether data samples $(y_1, y_2, \ldots, y_n)$ are in normal distribution or not, S–W test calculates S_w and is given by Eq. (5).

$$S_w = \frac{\left(\sum_{i=1}^{n}\alpha_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{5}$$

where y_i is the ith data sample, α_i is a constant value based on y and $\bar{y}$ is the mean of n samples (Shapiro and Wilk 1965).

3.4.3 Mardia's test

In order to check the similarity of the multivariate normal distribution generally multivariate tests are conducted (Mardia 1970). In this paper, we considered multivariate skewness measure to support null-hypotheses (H_0) and is given in Eq. (6).

Table 2 Best five combinations of features for four classifiers (Acc. → accuracy)

S. No.	MFCC	Pitch				Intensity				Jitter	Shimmer	Formants			Acc.(%) (N = 2)	Acc.(%) (N = 4)	Acc.(%) (N = 8)
		Min.	Max.	Mean	SD	Min.	Max.	Mean	SD			F1	F2	F3			
Vector quantization																	
1	✓				✓		✓	✓	✓	✓	✓	✓	✓	✓	49.16	61.32	**76.98**
2	✓				✓	✓	✓				✓	✓	✓		53.24	61.98	**76.14**
3	✓			✓				✓	✓	✓	✓	✓		✓	48.65	59.66	**75.32**
4	✓				✓	✓	✓	✓	✓			✓	✓		54.16	61.82	**75.92**
5	✓						✓	✓	✓		✓	✓			52.31	65.65	74.56
K-means clustering																	
6	✓	✓				✓	✓		✓		✓	✓		✓	62.00	**75.59**	61.00
7	✓					✓	✓	✓	✓		✓	✓		✓	58.86	**79.75**	62.38
8	✓	✓				✓	✓	✓	✓		✓	✓			69.61	76.24	69.78
9	✓		✓			✓		✓			✓	✓	✓	✓	63.67	76.42	68.28
10	✓					✓	✓			✓	✓	✓		✓	53.98	71.39	68.42
Gaussian mixture models																	
11	✓	✓	✓				✓			✓	✓	✓		✓	76.00	**84.78**	71.52
12	✓		✓		✓						✓	✓		✓	63.58	80.09	65.36
13	✓	✓	✓		✓		✓	✓						✓	71.34	81.66	55.24
14	✓	✓	✓		✓						✓	✓		✓	61.46	79.72	51.34
15	✓	✓	✓	✓			✓	✓				✓		✓	71.76	79.85	47.26
Artificial neural networks																	
16	✓			✓		✓	✓				✓		✓	✓	**72.47**		
17	✓			✓			✓	✓		✓			✓	✓	**71.49**		
18	✓	✓	✓				✓			✓	✓		✓	✓	**73.58**		
19	✓	✓			✓		✓	✓		✓		✓	✓		69.40		
20	✓		✓	✓	✓		✓	✓		✓	✓	✓	✓	✓	65.38		

Bold face indicates the best performance of a classifier

$$M_{1,m} = \frac{1}{N^2} \sum_{a=1}^{N} \sum_{b=1}^{N} \left[(y_a - \bar{y})' S^{-1} (y_b - \bar{y}) \right]^3 \quad (6)$$

Where $\bar{y}$ is the sample mean and S is the sample co-variance matrix (Rencher and Christensen 2012).

Table 3 shows the results of normality tests. From table, it is observed that the p values of Mardia's test are less than significant value 0.05. It indicates that the data of emotional speech follows normal distribution.

3.5 Classification models

The detailed explanation for four classification methods such as VQ, K-Means, GMM and ANN is given below.

3.5.1 VQ

In VQ approach, a set of feature vectors is mapped to a finite number of vectors called as *code vectors*. The collection of these code vectors form a *code book*. Further, the similar feature vector generated from the test clip will be compared with code vectors for computing deviation, also known as distortion. Lesser the deviation is the more match. The performance of VQ depends on the creation of an effective code book. In this work, the code book is computed by using LBG (Linde–Buzo–Gray) algorithm (Linde et al. 1980). Given a test speech feature vector, the distance for each code book have been calculated to find the one with minimum distance. The emotion class represented by that code book is the emotion of test clip. (Soong et al. 1987). Similarly, five codebooks are obtained for five emotions each of size N. So, all training vectors of an emotion are mapped to the set of code vectors of that specific codebook. The experiments have been conducted by varying the value of N ($N = 2, 4, 8$).

3.5.2 *K*-means clustering

Though K-means clustering is a well-known clustering algorithm, it may also be used as classification tool. The training and testing procedures are same as that of VQ approach. Instead of code books it forms K centroids. Initially, K centroids for each emotion have to be selected randomly. Due to this random selection of initial centroids, the final positions of centroids change every time we run K-means clustering algorithm. Thus, the centroid positions will be converged either by running the algorithm n times or by stopping whenever the centroids show the least average movement.

In this work, a mathematical approach has been used to estimate the distribution of data. The dispersion of a data distribution is measured by coefficient of variation (CV) which is given by $s/\bar{x}$. Where s is the standard deviation and $\bar{x}$ is the mean of the cluster. Generally, higher the CV value indicates that there is a greater the variability in data. K-means clustering tends to form clusters such that CV value is in between 0.3 and 1.0. If the CV value is out of the given range, K-means clustering forms final clusters that are different from true clusters so that CV value attains the prescribed range. This affects the classification accuracy of the patterns. Table 4 shows average CV values of the selected emotions for best five feature combinations.

Table 3 Normality test results using three statistical methods

Emotion	K–S test p value	S–W test p value	Skewness p value
Anger	0.037	0.020	0.029
Fear	0.019	0.065	0.011
Happy	0.058	0.086	0.008
Neutral	0.070	0.040	0.013
Sad	0.045	0.072	0.006

3.5.3 Gaussian mixture models

GMMs perform better when data is in normal distribution. In this work, for each emotion, a Gaussian mixture model is developed with $'N'$ Gaussian components. If the data for each emotion follow a multivariate normal distribution. Multivariate normal distribution is a distribution that contains a collection of two or more normal distributions. The features that follow multivariate normal distribution can be effectively modeled by GMMs. Classification accuracy of GMMs also depends on the factors like number of Gaussians in each class, size of the dataset, distribution of data and so on. From the statistical tests done in Sect. 3.4, it is observed that the feature vector with some selected features follows normal distribution. As GMM can effectively model the data if it is in normal distribution, better accuracy is achieved when compared to other techniques.

3.5.4 Artificial neural networks

ANNs capture the complex non-linear relations present in the data as similar to the human brain. They contain many simple processing elements called as neurons that are interconnected together to understand the hidden patterns. In general, ANNs contain three types of layers namely input, hidden and output layers. Each of these layers contains several neurons. ANNs are designed with an input and output layers. Number of neurons in the input layer is equal to the length of a feature vector. Number of neurons in the output layer is equal to the total number of classes. The structure of neural network which is specific to this work is as shown in Fig. 3. A simple feed-forward back propagation neural network (BPNN) algorithm has been used for this task (Han and Kamber 2006; Rojas 2013).

Table 4 Average coefficient of variation (CV) values of five emotions for best five feature combinations

S.No.	MFCC	Pitch		Intensity				Jitter	Shimmer	Formants			Average CV
		Min.	Max.	Min.	Max.	Mean	SD			F1	F2	F3	
1	✓	✓		✓	✓		✓		✓	✓		✓	0.4394
2	✓			✓	✓	✓	✓		✓	✓		✓	0.4296
3	✓	✓		✓	✓	✓	✓		✓	✓			0.4317
4	✓		✓	✓		✓			✓	✓	✓	✓	0.4734
5	✓			✓	✓			✓	✓	✓		✓	0.4664

Table 2 (row numbers 16–20) shows the best five feature combinations which give better accuracy with ANNs. The ANN used in this work contains one hidden layer. Experimentation has been done by varying the number of neurons in the hidden layer from $(n + 1/2)$ to $(2n/3)$, where n is number of input layer neurons. Better accuracy is observed at 1.7 times to the input neurons.

4 Results and analysis

In this section, some important results showing the performance of classifiers and identifying the feature sets for particular classifier are discussed. The justification that why a classifier works better for a specific set of data is also given. Initially, the possible subsets are formed for the feature vector and experimentation has been done to identify the best features that are suitable for emotion recognition. Four different classifiers are used to do this task. Table 2 shows five feature combinations giving the best results for a particular classifier. Columns represent all features and rows represent the chosen feature vector. In each row a cell containing '✓' indicates inclusion of that feature in the feature vector.

Later, the results are validated using cross validation method. As residual evaluation is not able to give information about the capability of the classifier for an unseen test set, *k-fold* cross validation is considered in this work. The entire subsets are divided into k-subsets. For every subset the remaining $n - k$ values are considered for training where n is the total number of feature values (Kohavi 1995). The average accuracy of all subsets are considered as system accuracy. The same experimentation has been done with different k values and the best results are obtained with the k value 10 (shown in Table 5).

The performance measurement considered in this work is *accuracy* and computed using the formula given in Eq. (7).

$$Performanceaccuracy = round\left(\frac{I_e}{T_e} \times 100\right) \tag{7}$$

Where I_e is the total number of emotional clips correctly identified and T_e represents the total number of emotional clips.

From the experimentation, there are few observations related to classification models have been noted here. Classification accuracy of VQ method depends on various factors like number of clusters in each class, size of the dataset, type of data in the dataset and so on. As the number of code vectors per codebook (N) increases, the accuracy increases till certain value of N (in this case $N = 8$). The same thing can be observed from row 1 to 5 of Table 2. Similar to VQ, classification accuracy of K-means clustering also depends on various factors. The algorithm generally tends to form clusters with relatively uniform distribution of cluster sizes (Xiong et al. 2009). Hence, CV value is computed to form the clusters. However, K-means is unable to map the emotional clips due to their ambiguity and non-linearity. GMM is the one which is giving better performance with 84% accuracy for detecting emotions in movie database and 81% accuracy for IIT-KGP emotional database. In addition to that, ANN is giving equivalent performance if the training set increases and the same is found in literature as well (Yegnanarayana 1994). With the movie database collected, it is giving 72% and for IIT-KGP database there is an increase

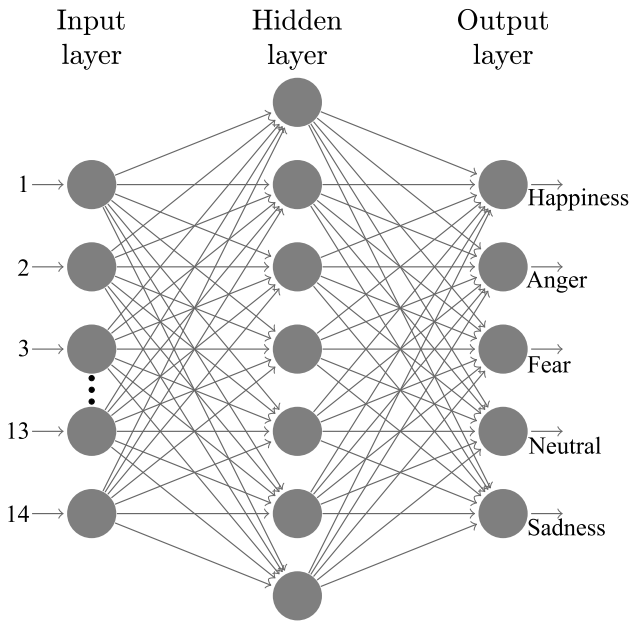

Fig. 3 Structure of artificial neural network for emotion classification

Table 5 Accuracy of classification models on different datasets

Classification model	MFCC	Pitch				Intensity				Jitter	Shimmer	Formants			Accuracy in %	
		Min	Max	Mean	SD	Min	Max	Mean	SD			F1	F2	F3	MDB	IIT-KGP
VQ	✓	✓	✓	✓	✓		✓			✓	✓	✓			57	62
K-means	✓	✓	✓	✓	✓		✓			✓	✓	✓			74	71
GMM	✓	✓	✓	✓	✓		✓			✓	✓	✓			**84**	**81**
ANN	✓	✓	✓	✓	✓		✓			✓	✓	✓			72	79

Bold face indicates the best performance of a classifier

MDB movie database

of 7%. From literature, it is true that ANN can improve the learnability if there is large training set.

5 Conclusion and future work

In this work, the classification model which suits better for a given feature set is suggested. In contrast to existing approaches or meta-learning, some statistical operations have been done on featureset and the classifier is recommended based on the results achieved. This work concludes that the classifier performance always depends on the dataset chosen. In this regard, GMM is giving better accuracy if the data falls in the normal distribution region. From the statistical analysis, it is observed that the emotional data majorly falls in the same region. As emotional data is non-linear and ambiguous, VQ and *K*-means are inappropriate to map them efficiently. ANN always gives better performance if the training set increases. However, it is unable to beat the performance given by classifier suggested. Finally, from several observations and analysis, the work concludes that the relevant emotional featureset falls in normal distribution and GMM is capable to classify it effectively when compared to other classifiers.

A state-of-art classifier has to be recommended for every subtask of speech processing to achieve good performance. This work may be extended to improve the performance by suggesting suitable classifier for other speech processing tasks such as speaker recognition, gender recognition, language identification and so on. However, the present work can be extended by extracting some more relevant features and experimenting with suitable classifiers. Moreover, the future work compares the feature depended systems with deep networks to determine the efficient algorithms for future research.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Anagnostopoulos, C.-N., Iliou, T., & Giannoukos, I. (2012). Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review*, *43*, 1–23.

Ananthapadmanabha, T., & Yegnanarayana, B. (1979). Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *27*(4), 309–319.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614.

Bhatti, M. W., Wang, Y., & Guan, L. (2004). A neural network approach for human emotion recognition in speech. In *ISCAS'04. Proceedings of the 2004 international symposium on Circuits and systems, 2004*, (Vol. 2, pp II–181). IEEE

Bishop, C. M., et al. (1995). Neural networks for pattern recognition. New York: Oxford University Press

Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-level spectral features for emotion recognition. *Speech Communication*, *52*(7), 613–625.

Black, M. J., & Yacoob, Y. (1995). Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In *Proceedings, fifth international conference on Computer vision, 1995*, (pp. 374–381). IEEE.

Bou-Ghazale, S. E., & Hansen, J. H. L. (2000). A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Transactions on Speech and Audio Processing*, *8*(4), 429–442.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., & Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 205–211). ACM

Chakraborty, R., Pandharipande, M., & Kopparapu, S. K. (2016). Knowledge-based framework for intelligent emotion recognition in spontaneous speech. *Procedia Computer Science*, *96*, 587–596.

Chauhan, A., Koolagudi, S. G., Kafley, S., & Rao, K. S. (2010). Emotion recognition using lp residual. In *Students' technology symposium (TechSym), 2010 IEEE* (pp. 255–261). IEEE.

Chavhan, Y., Dhore, M. L., & Yesaware, P. (2010). Speech emotion recognition using support vector machine. *International Journal of Computer Applications*, *1*(20), 6–9.

Chen, C., You, M., Song, M., Bu, J., Liu, J. (2006). An enhanced speech emotion recognition system based on discourse information. In *Computational Science–ICCS 2006* (pp. 449–456). New York: Springer (2006).

Chung-Hsien, W., & Liang, W.-B. (2011). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, *2*(1), 10–21.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1), 5–32.

Dai, K., Fell, H. J., & MacAuslan, J. (2008). Recognizing emotion in speech using neural networks. *Telehealth and Assistive Technologies*, *31*, 38–43.

Deller, J. R. P., John G., & Hansen, J. H.L. (2000). *Discrete-time processing of speech signals*. New York: IEEE.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Deng, J., Xinzhou, X., Zhang, Z., Frühholz, S., & Schuller, B. (2017). Universum autoencoder-based domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, *24*(4), 500–504.

Deng, J., Zhang, Z., Eyben, F., & Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, *21*(9), 1068–1072.

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

El-Yazeed, M. F., El Gamal, M. A., & El Ayadi, M. M. H. (2004). On the determination of optimal model order for gmm-based text-independent speaker identification. *EURASIP Journal on Applied Signal Processing*, *1078–1087*, 2004.

Essa, I. A., & Pentland, A. P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on Pattern analysis and machine intelligence*, *19*(7):757–763.

Farrus, M., & Hernando, J. (2009). Using jitter and shimmer in speaker verification. *IET Signal Processing*, *3*(4), 247–257.

Firoz, S.A., Raji, S.A., & Babu, A.P. (2009). Automatic emotion recognition from speech using artificial neural networks with gender-dependent databases. In *ACT'09. International conference on Advances in computing, control, & telecommunication technologies, 2009*, (pp. 162–164). IEEE

Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech Communication*, *41*(4), 603–623.

Fu, L., Mao, X., & Chen, L. (2008). Relative speech emotion recognition based artificial neural network. In *Computational intelligence and industrial application, 2008. PACIIA'08. Pacific-Asia workshop on* (Vol. 2, pp. 140–144). IEEE

Giannoulis, Panagiotis, & Potamianos, Gerasimos (2012). A hierarchical approach with feature selection for emotion recognition from speech. In *LREC* (pp. 1203–1206)

Grimm, M., Kroschel, K., & Narayanan, S. (2007). Support vector regression for automatic recognition of spontaneous emotions in speech. In *IEEE international conference on acoustics, speech and signal processing, 2007. ICASSP 2007*, (vol. 4, pp. IV–1085). IEEE

Han, J., & Kamber, M. (2006). *Data Mining*. Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann.

Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*.

Hernando, J., Nadeu, C., & Mariño, J. B. (1997). Speech recognition in a noisy car environment based on lp of the one-sided autocorrelation sequence and robust similarity measuring techniques. *Speech Communication*, *21*(1), 17–31.

Hess, W. J. (2008). Pitch and voicing determination of speech with an extension toward music signals. In *Springer Handbook of Speech Processing*, (pp. 181–212). Berlin: Springer.

Heuft, B., Portele, T., & Rauth, M. (1996). Emotions in time domain synthesis. In *Proceedings, fourth international conference on Spoken Language, 1996. ICSLP 96*, (Vol. 3, pp. 1974–1977). IEEE

Huang, J., Yang, W., & Zhou, D. (2012). Variance-based gaussian kernel fuzzy vector quantization for emotion recognition with short speech. In *2012 IEEE 12th international conference on Computer and information technology (CIT)*, (pp. 557–560). IEEE.

Iida, A., Campbell, N., Higuchi, F., & Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, *40*(1), 161–187.

Iida, A., Campbell, N., Iga, S., Higuchi, F., Yasumura, M. (2000). A speech synthesis system with emotion for assisting communication. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE), 2*(1), 235–238.

Jawarkar, N. P., et al. (2007). Emotion recognition using prosody features and a fuzzy min-max neural classifier. *The Institution of Electronics and Telecommunication Engineers*, *24*(5), 369–373.

Kaiser, L. (1962). Communication of affects by single vowels. *Synthese*, *14*(4), 300–319.

Kenji, M. A. S. E. (1991). Recognition of facial expression from optical flow. *IEICE TRANSACTIONS on Information and Systems*, *74*(10), 3474–3483.

Khanchandani, K. B., & Hussain, M. A. (2009). Emotion recognition using multilayer perceptron and generalized feed forward neural network. *Journal of Scientific and Industrial Research*, *68*(5), 367.

Khanna, P., & Kumar, M. S. (2011). Application of vector quantization in emotion recognition from human speech. In *Information intelligence, systems, technology and management* (pp. 118–125). New York: Springer.

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, *14*, 1137–1145.

Konar, A., & Chakraborty, A. (2014). *Emotion recognition: A pattern analysis approach*. Wiley: Hobroken, NJ.

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

Koolagudi, S. G., Maity, S., Kumar, V. A., Chakrabarti, S., & Rao, K. S. (2009). Iitkgp-sesc: Speech database for emotion analysis. In *International conference on contemporary computing* (pp. 485–492). New York: Springer

Koolagudi, S. G., Nandy, S., & Rao, K. S. (2009). Spectral features for emotion classification. In *Advance computing conference, 2009. IACC 2009. IEEE International* (pp. 1292–1296). IEEE

Koolagudi, S. G., Reddy, R., & Rao, K. S. (2010). Emotion recognition from speech signal using epoch parameters. In *2010 international conference on Signal processing and communications (SPCOM)*, (pp. 1–5). IEEE.

Kostoulas, T.P., & Fakotakis, N. (2006). A speaker dependent emotion recognition framework. In *Proceedings 5th international symposium, communication systems, networks and digital signal processing (CSNDSP), University of Patras* (pp. 305–309)

Krothapalli, S. R., & Koolagudi, S. G. (2013). Speech emotion recognition: A review. In *Emotion recognition using speech features*, pp. 15–34. New York: Springer.

Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion recognition by speech signals. In *INTERSPEECH*.

Le Bouquin, R. (1996). Enhancement of noisy speech signals: Application to mobile radio communications. *Speech Communication*, *18*(1), 3–19.

Lee, K.-F., & Hon, H.-W. (1989). Speaker-independent phone recognition using hidden markov models. *IEEE transactions on acoustics, speech and signal processing* , *37*(11), 1641–1648.

Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., & Narayanan, S. (2004). Emotion recognition based on phoneme classes. In *INTERSPEECH* (pp. 205–211).

Li, Y., & Zhao, Y. (1998). Recognizing emotions in speech using short-term and long-term features. In *ICSLP*.

Li, J. Q. & Barron, A. R. (1999). Mixture density estimation. In *Advances in neural information processing systems 12*. Citeseer.

Li, X., Tao, J., Johnson, M. T., Soltis, J., Savage, A., Leong, K. M., & Newman, J. D. (2007). Stress and emotion classification using jitter and shimmer features. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on* (Vol. 4, pp. IV–1081). IEEE.

Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399–402.

Lin, Y.-L., & Wei, G. (2005). Speech emotion recognition based on hmm and svm. In *Proceedings of 2005 international conference on Machine learning and cybernetics, 2005*, (Vol. 8, pp. 4898–4901). IEEE.

Linde, Y., Buzo, A., & Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE transactions on Communications*, *28*(1):84–95

Liu, H., & Lei, Y. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, *17*(4), 491–502.

Luengo, I., Navas, E., Hernáez, I., Sánchez, J. (2005). Automatic emotion recognition using prosodic parameters. In *INTERSPEECH* (pp. 493–496).

Mardia., K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519–530.

Motamed, S., Setayeshi, S., & Rabiee, A. (2017). Speech emotion recognition based on a modified brain emotional learning model. *Biologically Inspired Cognitive Architectures*, *19*, 32–38.

Muslea, I., Minton, S., & Knoblock, C. A. (2006). Active learning with multiple views. *Journal of Artificial Intelligence Research*, *27*, 203–233.

Muthusamy, H., Polat, K., & Yaacob, S. (2015). Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals. *Mathematical Problems in Engineering*, 2015.

Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion recognition in spontaneous speech using gmms. In *INTERSPEECH*.

Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural Computing & Applications*, *9*(4), 290–296.

Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech emotion recognition using hidden markov models. In *INTERSPEECH* (pp. 2679–2682).

Nooteboom, S. (1997). The prosody of speech: Melody and rhythm. *The Handbook of Phonetic Sciences*, *5*, 640–673.

Nwe, T. L., Wei, F. S., & De Silva, L. C. (2001). Speech based emotion classification. In *TENCON 2001, Proceedings of IEEE region 10 international conference on electrical and electronic technology, IEEE*, (Vol. 1, pp. 297–301).

Ortony, A. (1990). *The cognitive structure of emotions*. Cambridge: Cambridge University Press.

Pan, Y., Shen, P., & Shen, L. (2012). Speech emotion recognition using support vector machine. *International Journal of Smart Home*, *6*(2), 101–107.

Partila, P., & Voznak, M. (2013). Speech emotions recognition using 2-d neural classifier. In *Nostradamus 2013: Prediction, modeling and analysis of complex systems* (pp. 221–231). New York: Springer.

Petrushin, V. A. (2000). Emotion recognition in speech signal: experimental study, development, and application. *Studies*, *3*, 4.

Polzin, T. S. & Waibel, A. (1998). Detecting emotions in speech. In *Proceedings of the CMC* (Vol. 16). Citeseer

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals* (Vol. 100). Englewood Cliffs: Prentice-hall.

Rabiner, L. R., & Juang, B.-H. (1993). In *Fundamentals of speech recognition* (Vol. 14). Englewood Cliffs: PTR Prentice Hall .

Rao, S. K., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology*, *16*(2), 143–160.

Rao, K. S., & Koolagudi, S. G. (2012). *Emotion recognition using speech features*. New York: Springer Science & Business Media.

Rao, K. S., Reddy, R., Maity, S., & Koolagudi, S. G. (2010). Characterization of emotions using the dynamics of prosodic. In *Proceedings of speech prosody* (Vol. 4).

Razak, A. A., Komiya, R., Izani, M., & Abidin, Z. (2005). Comparison between fuzzy and nn method for speech emotion recognition. In *ICITA 2005. Third international conference on Information technology and applications, 2005*, (Vol. 1, pp. 297–302). IEEE

Reddy, S. Arundathy, Singh, Amarjot, Kumar, N. Sumanth, & Sruthi, K.S. (2011). The decisive emotion identifier. In *2011 3rd international conference on electronics computer technology (ICECT)*, (Vol. 2, pp. 28–32). IEEE.

Rencher, A. C., & Christensen, W. F. (2012). *Methods of multivariate analysis* (Vol. 709). New York: Wiley.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, *10*(1), 19–41.

Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on speech and audio processing*, *3*(1), 72–83.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Rojas, R. (2013). *Neural networks: A systematic introduction*. Berlin: Springer Science & Business Media.

Sato, N., & Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, *2*(3), 835–848.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227–256.

Scherer, K. R. (1989). Vocal correlates of emotional arousal and affective disturbance. In *Handbook of social psychophysiology* (pp. 165–197).

Schuller, B., Müller, R., Lang, M., & Rigoll, G. (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Ninth European Conference on Speech Communication and Technology*.

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden markov model-based speech emotion recognition. In *Proceedings. (ICASSP'03). 2003 IEEE international conference on acoustics, speech, and signal processing, 2003*, (Vol. 2, pp. II–1). IEEE.

Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Proceedings (ICASSP'04). IEEE international conference on acoustics, speech, and signal processing, 2004*, (Vol. 1, pp. I–577). IEEE.

Seehapoch, T., & Wongthanavasu, S. (2013). Speech emotion recognition using support vector machines. In *2013 5th international conference on Knowledge and smart technology (KST)* (pp. 86–91). IEEE.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611.

Shen, P., Changjun, Z., & Chen, X. (2011). Automatic speech emotion recognition using support vector machine. In *2011 International*

conference on electronic and mechanical engineering and information technology (EMEIT), (Vol. 2, pp. 621–625). IEEE.

Siqing, W., Falk, T. H., & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, *53*(5), 768–785.

Soares, C., & Brazdil, P. B. (2000). Zoomed ranking: Selection of classification algorithms based on relevant performance information. In *European conference on principles of data mining and knowledge discovery*, (pp. 126–135). New York: Springer

Song, P., Jin, Y., Zhao, L., & Xin, M. (2014). Speech emotion recognition using transfer learning. *IEICE TRANSACTIONS on Information and Systems*, *97*(9), 2530–2532.

Song, P., Zheng, W., Shifeng, O., Zhang, X., Jin, Y., Liu, J., et al. (2016). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Communication*, *83*, 34–41.

Soong, F. K., Rosenberg, A. E., Juang, B.-H., & Rabiner, L. R. (1987). Report: A vector quantization approach to speaker recognition. *AT&T Technical Journal*, *66*(2):14–26

Stuhlsatz, A., Meyer, C., Eyben, F., ZieIke, T., Meier, G., & Schuller, B. (2011). Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 5688–5691). IEEE

Takahashi, K. (2004). Remarks on svm-based emotion recognition from multi-modal bio-potential signals. In *ROMAN 2004. 13th IEEE international workshop on Robot and human interactive communication, 2004*, (pp. 95–100). IEEE.

Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, p. 37

Tang, H., Chu, S. M., Hasegawa-Johnson, M., & Huang, T. S. (2009). Emotion recognition from speech via boosted gaussian mixture models. In *IEEE international conference on Multimedia and expo, 2009. ICME 2009*, (pp. 294–297). IEEE.

Tian, Y., Kanade, T., & Cohn, J. F. (2000). Recognizing lower face action units for facial expression analysis. In *Proceedings, fourth IEEE international conference on Automatic face and gesture recognition, 2000*, (pp. 484–490). IEEE.

Traunmüller, H., & Eriksson, A. (1995). The frequency range of the voice fundamental in the speech of male and female adults. *Unpublished Manuscript*

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 5200–5204). IEEE

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

Ververidis, D., & Kotropoulos, C. (2005). Emotional speech classification using gaussian mixture models. In *IEEE international symposium on circuits and systems, 2005. ISCAS 2005*, (pp. 2871–2874). IEEE.

Vlassis, N., & Likas, A. (1999). A kurtosis-based dynamic approach to gaussian mixture modeling. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *29*(4), 393–399.

Vlassis, N., & Likas, A. (2002). A greedy em algorithm for gaussian mixture learning. *Neural Processing Letters*, *15*(1), 77–87.

Vogt, T., André, E., & Bee, N. (2008). EmoVoice—A framework for online recognition of emotions from voice. In *Perception in multimodal dialogue systems* (pp. 188–199). Springer.

Wang, K., An, N., Li, B. N., Zhang, Y., & Li, L. (2015). Speech emotion recognition using fourier parameters. *IEEE Transactions on Affective Computing*, *6*(1), 69–75.

Wang, L. (2005). *Support vector machines: Theory and applications*, (Vol. 177). Springer Science & Business Media.

Wenjing, H., Haifeng, L., & Chunyu, G. (2009). A hybrid speech emotion perception method of vq-based feature processing and ann recognition. In *WRI global congress on Intelligent systems, 2009. GCIS'09*, (Vol. 2, pp. 145–149). IEEE.

Womack, B. D., & Hansen, J. H. L. (1999). N-channel hidden markov models for combined stressed speech classification and recognition. *IEEE Transactions on Speech and Audio Processing*, *7*(6), 668–677.

Wu, S., Falk, T. H., & Chan, W. Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech communication, 53*(5), 768–785.

Xiong, H., Junjie, W., & Chen, J. (2009). K-means clustering versus validation measures: A data-distribution perspective. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *39*(2), 318–331.

Yacoob, Y., & Davis, L. (1994). Computing spatio-temporal representations of human faces. In *1994 IEEE computer society conference on Computer vision and pattern recognition, 1994. Proceedings CVPR'94*, (pp. 70–75). IEEE

Yamada, T., Hashimoto, H., & Tosa, N. (1995). Pattern recognition of emotion with neural network. In *Proceedings of the 1995 IEEE IECON 21st international conference on Industrial electronics, control, and instrumentation, 1995*, (Vol. 1, pp. 183–187). IEEE

Yang, B., & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal Processing*, *90*(5), 1415–1423.

Yegnanarayana, B. (1994). Artificial neural networks for pattern recognition. *Sadhana*, *19*(2), 189–238.

Yu, C., Tian, Q., Cheng, F., & Zhang, S. (2011). Speech emotion recognition using support vector machines. In *Advanced research on computer science and information engineering* (pp. 215–220). New York: Springer

Zheng, W., Xin, M., Wang, X., & Wang, B. (2014). A novel speech emotion recognition method via incomplete sparse least square regression. *IEEE Signal Processing Letters*, *21*(5), 569–572.

Zhou, Y., Sun, Y., Zhang, J., & Yan, Y. (2009). Speech emotion recognition using both spectral and prosodic features. In *ICIECS 2009. International conference on Information engineering and computer science, 2009*, (pp. 1–4). IEEE.

Zhou, J., Wang, G., Yang, Y., & Chen, P. (2006). Speech emotion recognition based on rough set and svm. In *5th IEEE international conference on Cognitive informatics, 2006. ICCI 2006*, (Vol. 1, pp. 53–61). IEEE.