DEVELOPING AN INTEGRATED MODEL USING SIGNAL AND SPEECH PROCESSING, CONTENT AND ACOUSTIC ANALYSIS TO ENHANCE THE EFFICACY OF SER USING PREDICTIVE STATISTICAL METRICS.

Karan Gupta

Deenbandhu Chhotu Ram University of Science and Technology, Murthal, Haryana,131039

ABSTRACT

Different authorities have coordinated examinations on the affirmation of feeling from human talk with different examination designs. Speech Emotion Recognition (SER) is a specific class of signal processing where the principal objective is to recognize the energetic state of people from voice. Effects of acoustic parameters, the authenticity of the data used, and execution of the classifiers have been the pivotal issues for feeling affirmation investigate the field. 81 s (distributed in the listed diaries) have been assessed by the approaches used for feeling stamping, acoustic features and classifiers and the database used. The principle point is to examine: portray the highlights of the databases being used and to make a brief on the proficiency of acoustic parameters and the classifiers utilized by the past examinations.

***KEYWORDS**: Content analysis, emotion recognition, acoustic analysis, signal processing, speech processing.*

INTRODUCTION

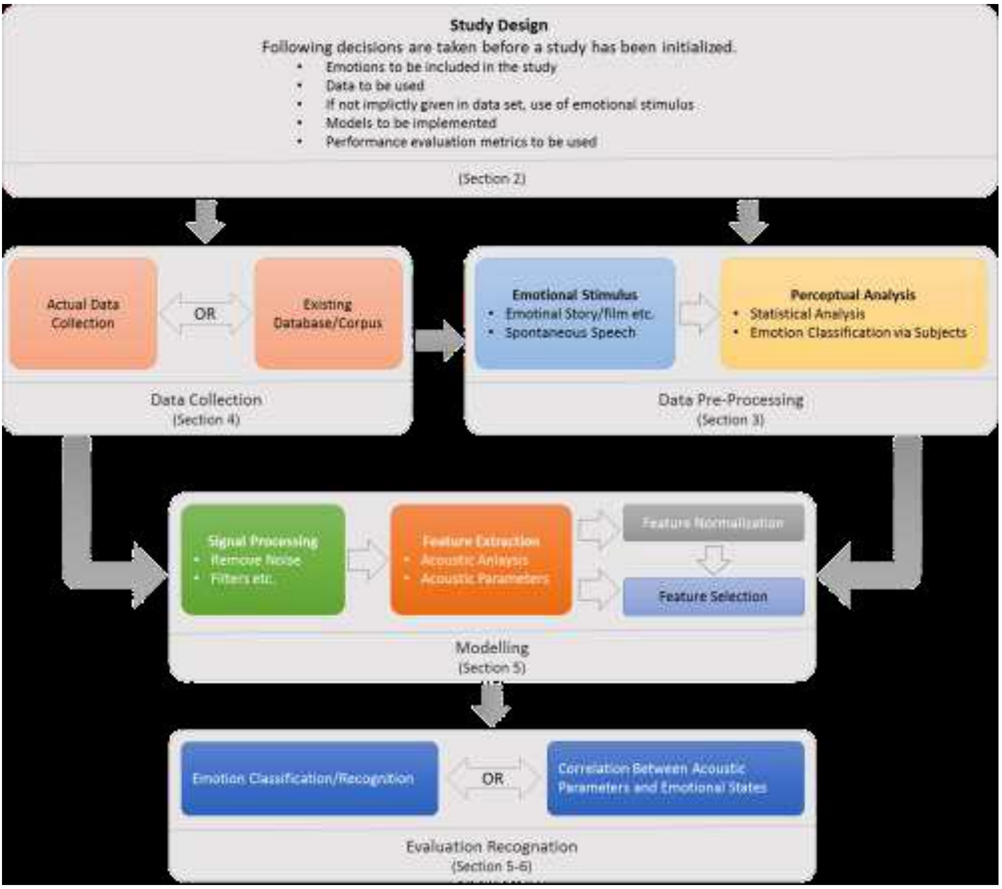
- The correspondence capacity, changing over the sound to the type of the discourse, is the most vital perspective that is recognizing the human from other living individuals. Discourse is an unpredictable capacity which happens by means of sound way preparing [1]. Discourse, notwithstanding being a specialized device, is likewise a pointer of a man's personality, mental state and physical wellbeing and so on. In this manner, the programmed Speech Emotion Recognition (SER) has an immense potential in the utilization of fields, for example, brain research, psychiatry and the full of feeling processing innovation [2]. There have been various examinations concentrating on the connection between discourse and individual perspectives/feelings. Then again, the examinations looking at the SER ponder by the "approaches/classifiers", "the feelings included", "acoustic highlights", "information sources" has been restricted. To some degree, the logical investigation can be extremely advantageous to: comprehend, translate, examine and integrate of an exploration territory. For this reason, in this substance examination, 81 papers that were distributed about SER somewhere in the range of 2005 and 2015 (January) were dissected. Production seek was performed utilizing the web index of Web of Science (WoS) where filed distributions are recorded. Inside the extent of this exploration just on the diary productions were incorporated and gathering/symposium papers were avoided. The accompanying watchwords were sought and 81 articles were recovered:

- speech emotion recognition
- vocal emotion recognition
- acoustic and emotion
- acoustic and emotional dysregulation
- acoustic and emotional disorder
- acoustic and affective disorder
- acoustic and affect dysregulation
- acoustic and mood disorder

Since the way toward directing a relevant examination can set up a system inside the exploration region run of the mill ventures of SER investigation covering five stages was exhibited as a structure appeared in Fig.1. These steps explained in Fig.

In such manner, the flowchart given in Fig.1. can likewise be utilized as a guide for the readers. The information gathering step contains the obtaining of information with respect to the voice records which will be utilized in the examination. A few scientists incline toward the utilization the information that they gathered and some others utilize existing databases. Different paid and free discourse/feeling databases are additionally accessible for scientists. It is likewise observed that the information corpora, used to uncover the passionate state, are additionally utilized seriously in the investigations.

Information securing does not similarly imply that the information is prepared to be handled. Preprocessing step may be an essential need to make the data ready to be further analyzed. data for their studies should convert the speech and audio recordings (that is in analog format) into a numeric format for further digital signal processing (DSP). It is likewise hard to recognize the yields (relating feelings) in SER examines. There are objective and abstract techniques that are utilized for feeling marking. Perceptual assessment is an abstract assessment technique which is just an elucidation of records by the specialists. Nonetheless, specialists might not have similar decisions about the given records. There are target assessment techniques that are utilized to beat this subjectivity issue [3]. The acoustic examination has been a broadly used technique to impartially assess the discourses which is a cheap strategy giving a target, noninvasive information in a brief timeframe. Programming bundles likewise exist for acoustic examination [4] to make investigation simpler. In a portion of the examinations; circumstances that trigger the feelings (enthusiastic improvement) can be utilized to research the adjustments in discourse. Then again, the stable flag and sound way include shift by age, sex, body weight, and tallness and the length of the sound way. In modelling step; signal processing techniques and filtration of the speech signal are used to remove the factors that are out of interest. In the event that the span of the noteworthy highlights (extensive acoustic parameters) is too vast, the quantity of the highlights can be decreased by utilizing measurement decrease techniques which can diminish calculation time. The best measurement decrease strategies can be listed as: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Sequential Forward Selection (SFS). Evaluation recognition step includes emotion recognition/classification or the detection of relationship between emotions and acoustic parameters. With the advances in computer architectures, complex emotion recognition algorithms have been in use [5]. In a portion of the investigations, acoustic highlights, etymological and logical data have been utilized in blend for the feeling recognition[6].



.

Fig.1. Flowchart of typical speech emotion recognition studies

RESEARCH DESIGNS OF SER STUDIES

To ensure integrity and quality of a research study, it should be designed, reviewed and undertaken. The main considerations in a typical SER study should answer the following questions:

What is the motivation behind the examination (feeling limited/statistic confined)?

Which information to be utilized (a current database, corpus or another informational collection is required)?

➔ If data will be collected:

- What is the population under interest?
- Will sampling methodologies used or whole population to be included?
- How to call the potential members to the discourse recording (which protection issues ought to be expressed and ensured?)
- Will there be emotional stimulus (such as: text reading, film watching) or Spontaneous speech?
- What kind of recording technologies will be used?
- How will the “record environment” be prepared?
- How will the discourse records be coordinated with the feelings included? Will there be a specialist gathering to recognize the feeling (or a computerized system be embedded in the investigation)?

➔ If database which is already exists will be used:

- database consistent?
- data obtained techniques?
- condition of data use agreement (payment)?
- Which studies used that database before? What did they do? What were the findings?
- Are the results comparable?
- How to eliminate the effects of other factors from the data (noise; effect of age and etc.)?
- Is there a magnitude incompatibility which requires data transformation/normalization?
- How to combine/eliminate (feature selection/ dimension reduction) some acoustic parameters to increase computing speed and accuracy of findings?
- What models to be used to search the potential relationships?

- How to verify the model and check validity?
- How to compare results? Is there statistical test required?

These research design issues for the reviewed 81 papers will be discussed in the subsequent sections of this paper.

EMOTIONS USED IN THE SER LITERATURE

The general population's passionate state fluctuates relying upon the current mental condition, the earth, or the challenges happened previously. In some of the SER studies, emotions examined in a wide category like: positive and negative [6], [7]. In some other SER studies (Fig. 2.) “anger”, “sadness”, “happiness” and “fear” have been the emotions that are mostly under interest.

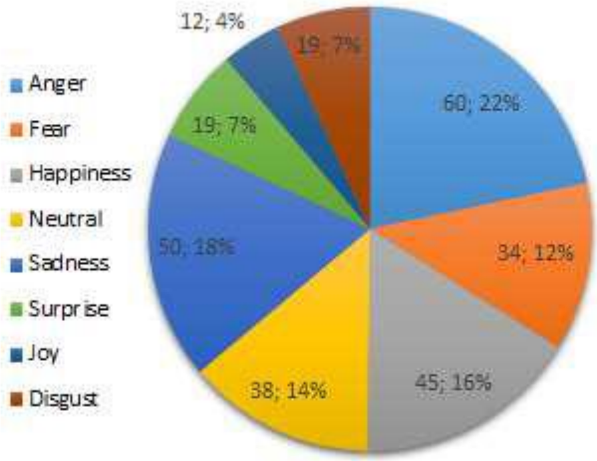


Fig.2. The emotions distribution under consideration

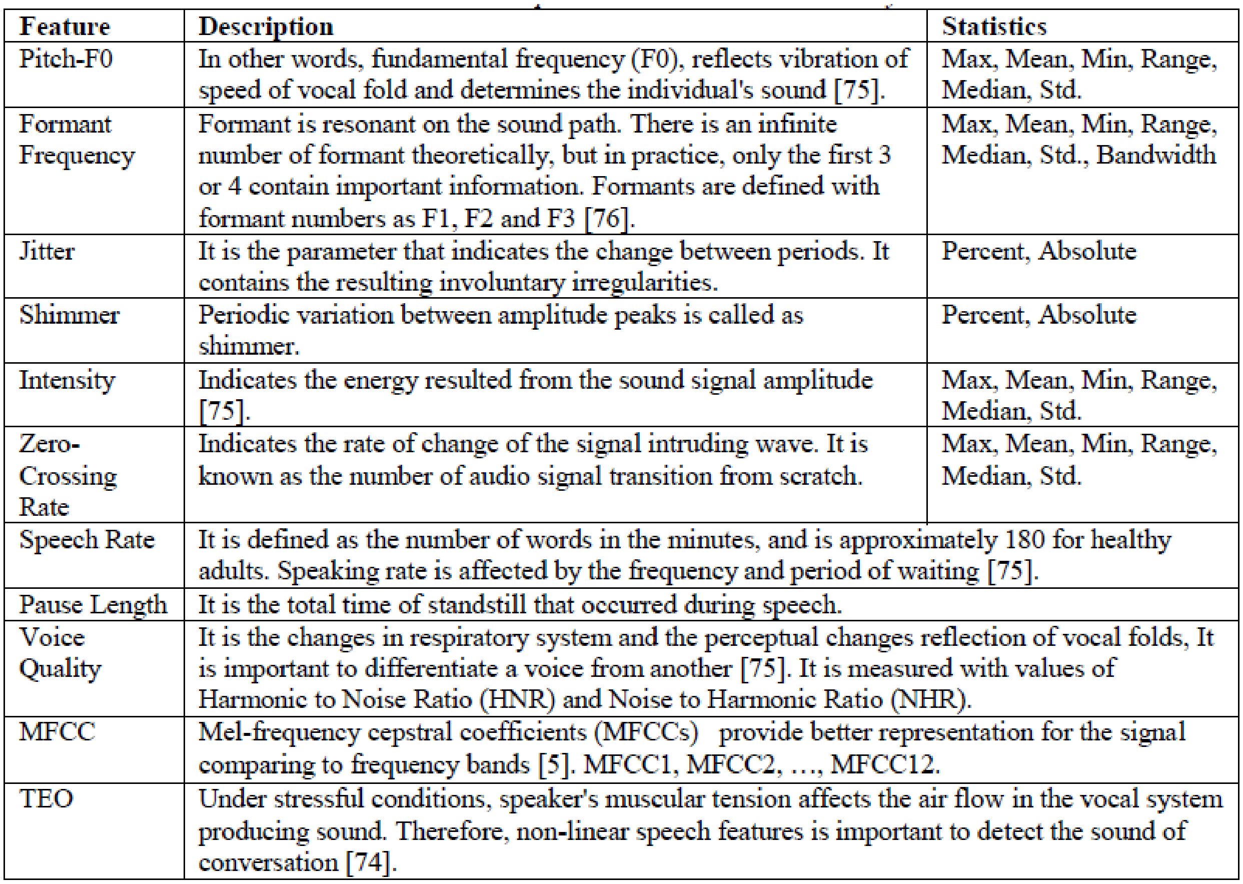
There are likewise some different investigations that consider some different emotions like :Milton and Tamil Selvi, 2014, Ramakrishnan and El Emary, 2013; Siegert et al., 2014; Truong et al., 2012; Zao et al., 2014), tired [15], [16], rest [17], [18], emphatic [15], [17], [18], appreciation [19], awe [19], calm [20], fidgetiness [16], achievement [21], gloating [19], gratitude [19], interest [22], [23], polite [24], reproach [19], serenity [25], taunting [26] and tickling [26]. Despite the fact that the specialists may recalibrate the feeling list by including or evacuating a specific feeling that they are keen on their investigation, the central thought remains that feeling is discrete and can be measured utilizing discourse [27]. Even the data on the available databases were labeled according to the emotional states; the sound records conducted through the corpus requires to be labeled with the emotions. In the reviewed studies; reading emotional stories to participants or having participants listen to audio records has been used to create an emotional stimulus. It is apparent that the majority of participants in those tests are the ordinary people who have no expertise in the area.

DATA COLLECTION METHODOLOGIES USED IN SER STUDIES

One of the major challenges on the emotion recognition studies is to obtain a data set where the performance was tested and which contains natural emotional states. The values in the grids indicate the corresponding publications. It is understood that most of the SER studies conducted in English and German languages and anger, joy and fear emotions have been mostly focused on emotions. s Interactive Multimedia Project (CHIMP)" involving the interaction of the machines with the kids in the games [61]. acoustic demonstrating for programmed discourse acknowledgment (ASR), dialect and exchange displaying, and multimodal-sight and sound UI outline. Acoustic displaying adjustment and vocal tract [36] standardization calculations that yielded cutting edge ASR execution on youngsters' discourse are portrayed [31] to collect the data indicating excessive emotional symptoms in the life-threatening situations. With this purpose, Clavel et al. have developed SAFE corpus (situation analysis in a fictional and emotional corpus) including fear and neutral feelings based on fiction films in 2008. [33] created a database named "Multilingual Emotional Speech Database of North East India" (MESDNEI). Members were given (in five distinctive local dialects) to peruse short sentences covering the feelings: anger, disgust, fear, happiness, sadness, surprise and neutral. [62] introduced two databases: BHUDES- Beihang University Database of Emotional Speech and BHUDEP-Beihang University Database of Emotional Points. [21] have built up a corpus containing feelings, for example, accomplishment/triumph, delight, erotic joy, and help, got through two ladies and two men members. List of databases available to be used in SER studies. Although there are several databases that contain the emotional speech, their accuracy of emotion labeling may be a misleading factor. The accuracy of labeling may be affected from the simulation of the emotions or the factors like recording environment, sound recorder, sound quality, and actors used for the record may distort the labeling performance [74].

ACOUSTIC PARAMETERS AND METHODS USED IN THE SER STUDIES

The most important matters for the SER studies are; the determination of the speech features that expresses different emotions and their corresponding change with respect to emotions. The difficulty on the determination of acoustic features for a certain emotion is due to: the individual variation in sound, age and gender differences [75]. The parameters that are widely used on acoustic features are given in Table 1



Feature	Description	Statistics
Pitch-F0	In other words, fundamental frequency (F0), reflects vibration of speed of vocal fold and determines the individual's sound [75].	Max, Mean, Min, Range, Median, Std.
Formant Frequency	Formant is resonant on the sound path. There is an infinite number of formant theoretically, but in practice, only the first 3 or 4 contain important information. Formants are defined with formant numbers as F1, F2 and F3 [76].	Max, Mean, Min, Range, Median, Std., Bandwidth
Jitter	It is the parameter that indicates the change between periods. It contains the resulting involuntary irregularities.	Percent, Absolute
Shimmer	Periodic variation between amplitude peaks is called as shimmer.	Percent, Absolute
Intensity	Indicates the energy resulted from the sound signal amplitude [75].	Max, Mean, Min, Range, Median, Std.
Zero-Crossing Rate	Indicates the rate of change of the signal intruding wave. It is known as the number of audio signal transition from scratch.	Max, Mean, Min, Range, Median, Std.
Speech Rate	It is defined as the number of words in the minutes, and is approximately 180 for healthy adults. Speaking rate is affected by the frequency and period of waiting [75].	
Pause Length	It is the total time of standstill that occurred during speech.	
Voice Quality	It is the changes in respiratory system and the perceptual changes reflection of vocal folds, It is important to differentiate a voice from another [75]. It is measured with values of Harmonic to Noise Ratio (HNR) and Noise to Harmonic Ratio (NHR).	
MFCC	Mel-frequency cepstral coefficients (MFCCs) provide better representation for the signal comparing to frequency bands [5]. MFCC1, MFCC2, …, MFCC12.	
TEO	Under stressful conditions, speaker's muscular tension affects the air flow in the vocal system producing sound. Therefore, non-linear speech features is important to detect the sound of conversation [74].	

Table 1

The usage frequency of the acoustic parameters (employed in ten or more publications) is as given. Twenty-seven different types of acoustic parameters were used in the reviewed studies. Twelve of them have been very common in use as depicted.

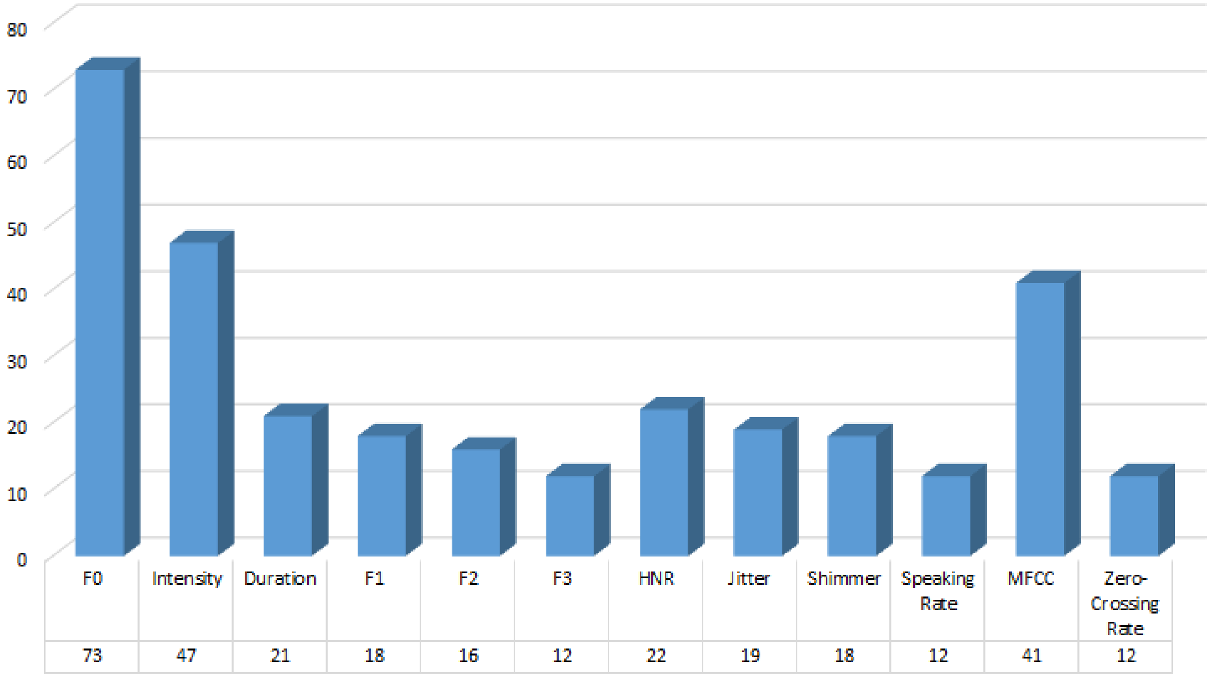


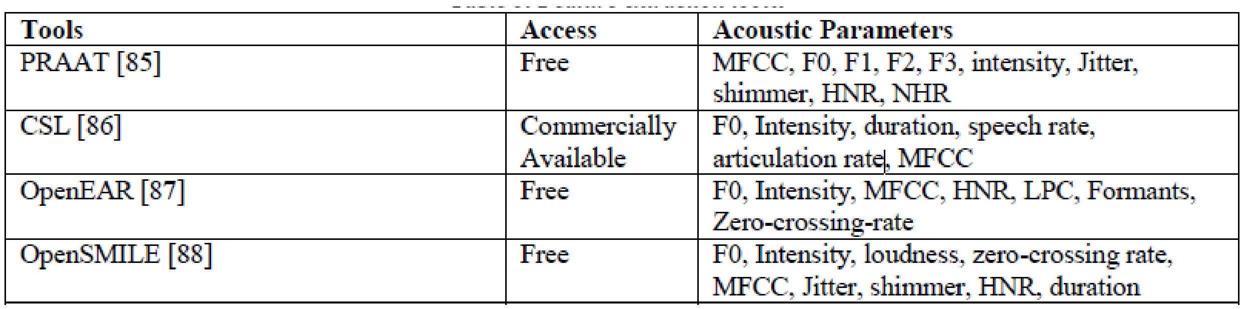
Fig.3. The frequency of the acoustic parameters used

The preprocessing is an important step to improve the performance of classification in SER studies. Noise removal, pre-emphasis, and windowing techniques have been the widely used preprocessing techniques. Noise is defined as undesired signals that cause deterioration in the signal during communication, measurement, and signal processing applications [77]. The rate of noise to a signal is defined as Signal to Noise Ratio (SNR). If this value is high, it indicates that noise is low. Therefore, many methods provide a noise remove in the range of 10 to 20 dB [78]. Components of audio signals are frequency and harmonics. Since fundamental frequency is stronger than the noise, it is not affected too much by noise. However, the harmonics having low amplitude value are affected severely. This causes a value of the SNR to decrease. To overcome this problem, SNR value is increased by strengthening the harmonics having high frequency but low amplitude. This process is called as pre-emphasis. Pre-emphasis is a calibrated filter that is used to synchronize the effect of speech transmission through the air [74]. Windowing is used to reduce the effect leakage and the noise level in the speeches. The most widely used method of windowing is Hamming. Impact of noise removal on speech recognition has been studied via several databases (Danish Emotional Speech Corpus -DES, Berlin Emotional Speech Database-EMO-DB, Speech Under Simulated and Actual Stress- SUSAS and it is concluded that: as noise level increases, an accuracy of the emotion recognition decreases [79].

A white noise signal low pass filter was used to control the random change in the acoustic parameters [48]. Relative SpectrAl (RASTA) filter and cepstral mean subtraction (CMS) methods are used to eliminate the spectral changes invisible during speech and eliminate the environmental noise [14].

Feature Extraction Tools

There are configuration apparatuses for the figuring of acoustic parameters the pre-preparing stages to be executed on the speech signals appeared in Table 2. These devices and acoustic highlights that can be identified by them has been utilized in by far most of the investigations (Agrawal et al., 2010; Bänziger et al., 2014; Bejani et al., 2014; Diamond et al., 2010; Hoque et al., 2006; Jia et al., 2011; Laukka et al., 2011, 2005; Leitman et al., 2010; Lima et al., 2013; Livingstone et al., 2014; López-Cózar et al., 2011; Origlia et al., 2014; Patel et al., 2011; Paulmann et al., 2008; Pell et al., 2009; Pérez-Espinosa et al., 2012; Rochman et al., 2008;Scherer, 2013; Scherer et al., 2015; Szameitat et al., 2009; Truong et al., 2012).



Tools	Access	Acoustic Parameters
PRAAT [85]	Free	MFCC, F0, F1, F2, F3, intensity, Jitter, shimmer, HNR, NHR
CSL [86]	Commercially Available	F0, Intensity, duration, speech rate, articulation rate, MFCC
OpenEAR [87]	Free	F0, Intensity, MFCC, HNR, LPC, Formants, Zero-crossing-rate
OpenSMILE [88]	Free	F0, Intensity, loudness, zero-crossing rate, MFCC, Jitter, shimmer, HNR, duration

Table 2

Feature Normalization and Selection Techniques

The performance of a classifier is directly affected by classifier training, the size of the data and data unit differences to be used during the test phase. To overcome this problem feature normalization technique are used.

Mostly z-score normalization technique is used for relevant studies. z-score technique for feature x is given in Equation 1 [74];

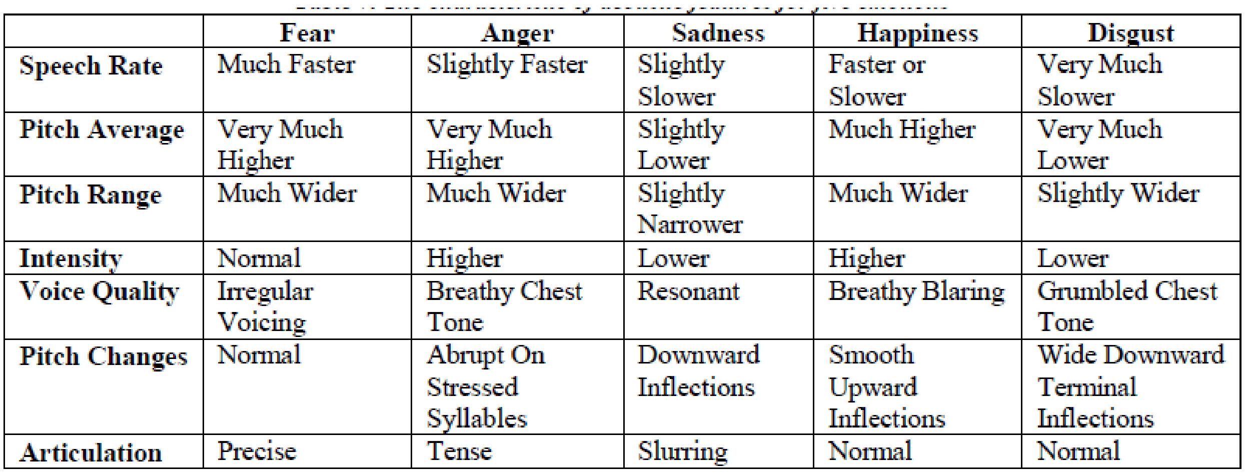
$$x_n = \frac{x - \mu}{\sigma}$$

where μ is the cruel of the x and σ is the standard deviation of it.

Feature selection techniques, are used for the purpose of determining the best classification features from the feature set. Reducing the size of the data set with feature selection, classification performance and accuracy are increased.

Acoustic Cues of Emotion

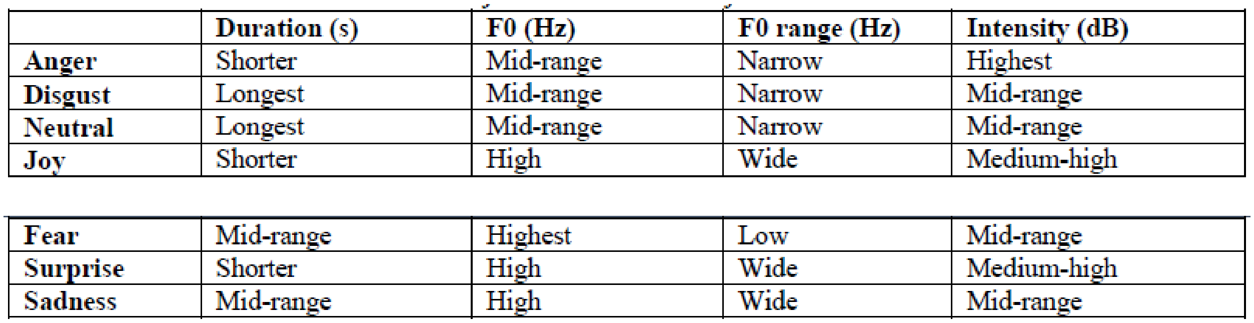
Acoustic features, obtained by processing of sound which is a signal, are used for determination of emotional state. Besides, since acoustic analysis is an objective assessment method, by which, evaluator independent results can have obtained. Murray and Arnott (1993) have created a table that indicates the relationship between basic five emotional states and acoustic features. The relationship between acoustic features and emotional status is given in Table 3 [109].



	Fear	Anger	Sadness	Happiness	Disgust
Speech Rate	Much Faster	Slightly Faster	Slightly Slower	Faster or Slower	Very Much Slower
Pitch Average	Very Much Higher	Very Much Higher	Slightly Lower	Much Higher	Very Much Lower
Pitch Range	Much Wider	Much Wider	Slightly Narrower	Much Wider	Slightly Wider
Intensity	Normal	Higher	Lower	Higher	Lower
Voice Quality	Irregular Voicing	Breathy Chest Tone	Resonant	Breathy Blaring	Grumbled Chest Tone
Pitch Changes	Normal	Abrupt On Stressed Syllables	Downward Inflections	Smooth Upward Inflections	Wide Downward Terminal Inflections
Articulation	Precise	Tense	Slurring	Normal	Normal

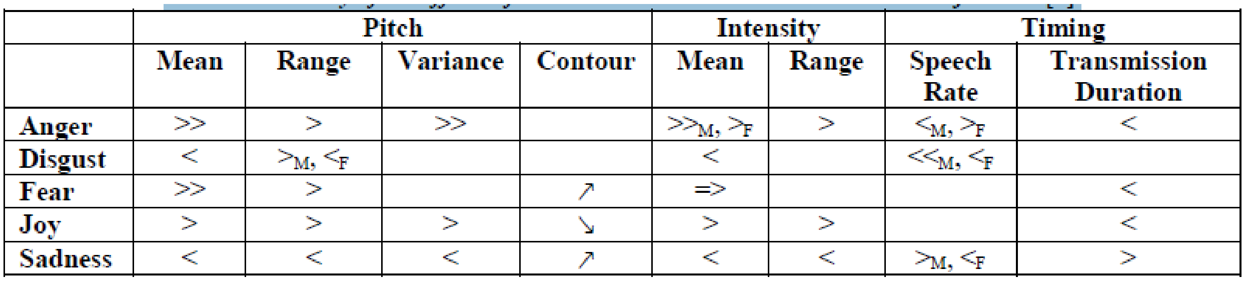
Table 3

As seen in Table 3, the differences in the features of the audio for all of the different emotions can be observed. Emotional state determination can be accomplished by using these differences. Drioli et al. (2003) examined the relationship between the emotion state and acoustic parameters over 5 parameters and six emotions. The results obtained are given in Table 4 [110].



	Duration (s)	F0 (Hz)	F0 range (Hz)	Intensity (dB)
Anger	Shorter	Mid-range	Narrow	Highest
Disgust	Longest	Mid-range	Narrow	Mid-range
Neutral	Longest	Mid-range	Narrow	Mid-range
Joy	Shorter	High	Wide	Medium-high
Fear	Mid-range	Highest	Low	Mid-range
Surprise	Shorter	High	Wide	Medium-high
Sadness	Mid-range	High	Wide	Mid-range

The data, given in Table 4 were obtained for six different emotion states including vowel, consonant, and vowel combinations. By using the combination of acoustic features in the table, emotion recognition is performed. Ververidis, and Kotropoulos (2006) by their literature view, summarized the relationship between the emotion
state and acoustic as given in Table 5 [5]



	Pitch				Intensity		Timing	
	Mean	Range	Variance	Contour	Mean	Range	Speech Rate	Transmission Duration
Anger	>>	>	>>		$>>_{M}, >_{F}$	>	$<_{M}, >_{F}$	<
Disgust	<	$>_{M}, <_{F}$			<		$<<_{M}, <_{F}$	
Fear	>>	>		↗	=>			<
Joy	>	>	>	↘	>	>		<
Sadness	<	<	<	↗	<	<	$>_{M}, <_{F}$	>

Table 5.

Explanation of symbols: >: increases, <: decreases, =: no change from neutral, ↗: inclines, ↘: declines. In this section, it is indicated how acoustic features were affected by emotional state. In the result of literature survey, how the acoustic features such as F0, Intensity, Duration, F1, F2, F3, Jitter, shimmer, HNR, Speaking Rate, MFCC and Zero-Crossing Rate, changes with the emotional states such as anger, fear, happiness, sadness, neutral, surprise, joy and disgust, in terms of emotions given below. By taking into account the results given in the publications and analysis tables, for ratings between the parameter and the emotional state, the ranges of 1-5 (1: very low, 2: low, 3: medium, 4: high, 5: very high) were used.

In the classification performed by using wavelet transform and MFCC for the feature set, using the wavelet-based WPCC2 and tfWPCC2 feature sets increased the success of feature set that [33]. Besides, Teager-Energy Operator (TEO) use, increased the average success [13], [33]. While Zero-crossing rate and MFCC parameters improving the recognition rate of anger and happiness reduces the rate of recognition for sadness and neutral [13], [28]. While contextual and acoustic features are offering the best result for the detection of frustration emotion, lexical information yielded better results than acoustic and contextual information in respect of politeness determination [24]. Also, classification performance also varies with age and sex [24]. Especially for the detection of politeness, higher accuracy classification was provided in the range of 10-11 years old ladies compared to the ladies with other age ranges and men [24]. Pinnacle execution has been gotten over MFCC include set [24]. At the point when [7] the trouble of distinguishing State of feelings is analyzed, it is discovered that the location of annoyance and misery feelings is the least demanding and that of dread is the most troublesome [62]. In the examination where ANOVA investigation of acoustic [7 parameters and energetic states are fused, it is seen that the extension in F0 is huge only for fulfillment, and the development in F1 is noteworthy for F1, and differentiations for F2 isn't found vital for F2. The changes in the duration are meaningless [55].

Anger: n the investigations performed under indignation feeling state, it is discovered that F0 is high, mean F0 is high, max F0 is high, min F0 is high, standard deviation of F0 is medium, and F0 territory is high [9], [21], [32], voice quality is high [11], [21], [46], [48], [60]; and high in some studies [21], [47]; mean intensity and standard deviation intensity are high [9], [21], [36], standard deviation is F1, F2 and standard deviation of F2 is very high [55].

Fear: In the investigations performed under dread feeling state, it is discovered that maximum F0 is low, min F0 is medium, mean F0 is high, standard deviation of F0 is medium, F0 territory is medium [9], [21], [32], [41], [46], [48], [49], [51], [52], [58]; medium [46] and high [52]; duration parameter is low [21], [48]; mean intensity and standard deviation intensity are high [9], [21], [46], [48], [51]; pause and speaking rate are medium [32], [51].

Happiness: In the examinations performed under Happiness feeling state, it is discovered that mean F0 is so high, standard deviation of F0 is high, go F0 is medium, max F0 is so high, min F0 is high [32], [34], [36], [41];duration is medium [36], [47], [48]; mean intensity is high and standard deviation intensity is medium [36], [46]–[48]; speaking rate is high [32].

Sadness: In the investigations performed under Happiness feeling state, it is discovered that mean F0 is low and high, max F0 is high, min F0 is medium, standard deviation of F0 is medium, scope of F0 is high [9], [21], [32], [34], [41], [45]–[48], [51], [52], [59]; voice quality is low [21], [58], high [45], [46], [52]; in some studies, duration is high [45], [48] in some studies it is low [21], [47]; mean intensity is high, and standard deviation intensity is medium and low [9], [21], [46]–[48], [51]; pause is high, and speaking rate is medium and low [32], [51].

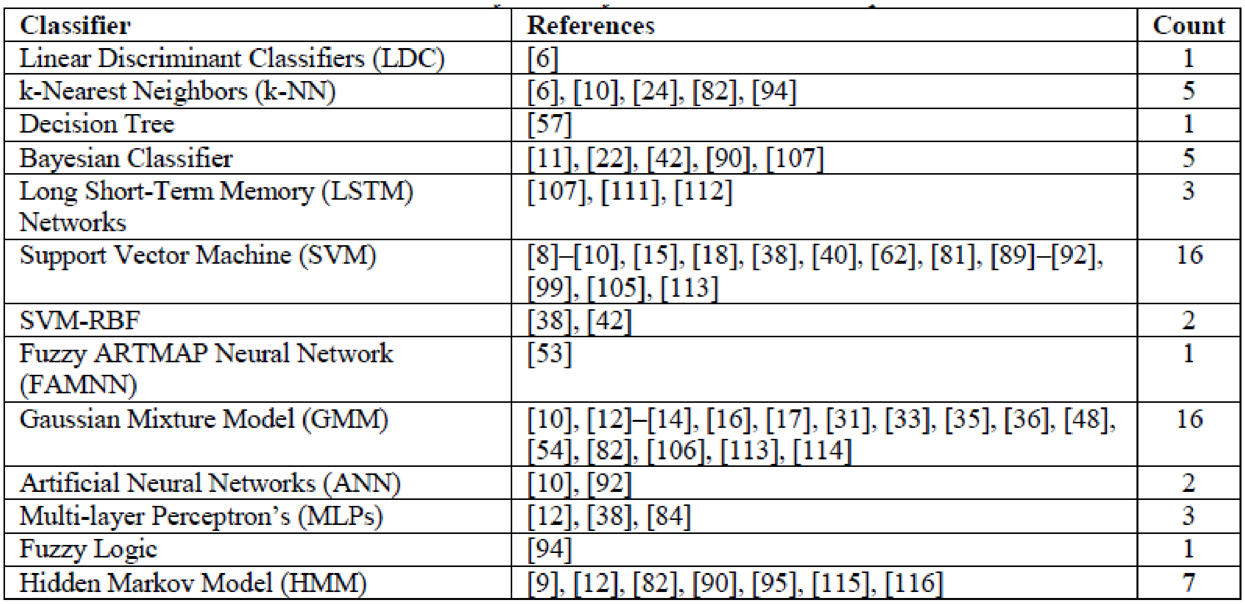
Neutral: In the studies performed under Neutral emotion state, it is found that mean F0 is very low, standard deviation of F0 is medium and low, range F0 is very low [32], [49], [55]; speaking rate is high [32]; F1, standard deviation of F1, F2 and standard deviation of F2 are medium [55].

Joy: In the studies performed under Joy emotion state, it is found that mean F0 is high, max F0 is high, standard deviation of F0 is high [9], [26], [45], [51], [52], [55]; voice quality is low and high [26], [45], [52]; duration is medium and high [26], [45]; mean intensity is high [9], [51]; pause is medium [51]; jitter and shimmer are low [52]; F1 high and low, standard deviation of F1 is so high, F2 is high and low, standard deviation of F2 is high [26], [55]; speaking rate is high [9].

Disgust: In the investigations performed under Joy feeling state, it is discovered that mean F0 is low, max F0 is low, extend F0 is low, standard deviation of F0 is high [9], [21], [32], [45], [46]; voice quality is medium and low [21], [45], [46]; duration is medium [21], [45]; mean intensity and standard deviation intensity are medium and high [21], [46]; speaking rate is low [32].

THE CLASSIFICATION TECHNIQUES USED IN THE SER STUDIES

In the vast majority of studies involving emotional state emotion recognition, classification of emotional states is realized, and the purpose of the classification is emotion recognition process. Traditional classification techniques were implemented in almost all of the presented emotion recognition systems. Current studies focus on hybrid classifiers and their effects on the acoustic parameters. Classification techniques used in the publications are given in Table 6.



Classifier	References	Count
Linear Discriminant Classifiers (LDC)	[6]	1
k-Nearest Neighbors (k-NN)	[6], [10], [24], [82], [94]	5
Decision Tree	[57]	1
Bayesian Classifier	[11], [22], [42], [90], [107]	5
Long Short-Term Memory (LSTM) Networks	[107], [111], [112]	3
Support Vector Machine (SVM)	[8]–[10], [15], [18], [38], [40], [62], [81], [89]–[92], [99], [105], [113]	16
SVM-RBF	[38], [42]	2
Fuzzy ARTMAP Neural Network (FAMNN)	[53]	1
Gaussian Mixture Model (GMM)	[10], [12]–[14], [16], [17], [31], [33], [35], [36], [48], [54], [82], [106], [113], [114]	16
Artificial Neural Networks (ANN)	[10], [92]	2
Multi-layer Perceptron's (MLPs)	[12], [38], [84]	3
Fuzzy Logic	[94]	1
Hidden Markov Model (HMM)	[9], [12], [82], [90], [95], [115], [116]	7

Table 6 that; SVM, GMM, HMM, K-NN and Bayesian Classifier are the most common classifiers.

CONCLUSIONS

In this study, Speech Emotion Recognition (SER) studies considering acoustic features have been analyzed via data acquisition methods; data preprocessing, feature extraction, classifiers, acoustic features, and emotional states.

The general tendency in these studies can be summarized as bellows:

It has been understood that "anger", "sadness", "happiness" and "fear" emotions have been the most widely included emotions in the studies. In addition to those studies on basic emotions, there are also the studies which analyzed the emotions dimensionally in respect of their direction and violence [9], [16].

Most of the SER studies identified (labeled) the emotions by perceptual analysis and listening tests before the acoustic analysis.

- In SER studies, existing databases were used and in some of the other studies; data were collected through the college students, phone calls [6], spontaneous dialogue, and speech records obtained through a scenario. In addition to those studies, there are also studies created their own emotional database [21], [31], [33], [61], [62]. Berlin Emotional Database has been the most frequently used database with the highest success [67].

- The most widely used acoustic parameters have been: F0, intensity, MFCC, HNR, duration, Jitter, shimmer, F1, F2, F3, speaking rate and the zero-crossing rate.

- The most commonly used tool of detecting acoustic parameters is PRAAT software [85].

- The most used feature normalization methods are z-score and feature selection method is PCA. In some studies, feature normalization and/or feature selection methods were not used. There are also studies which use all available features and resulted with a higher success rate of classification [94]. The use of PCA and LDA together yields better results when compared to their separate use [7]. Among SFS, LSBOUND, and R2W2 feature selection methods, the success of LSBOUND and R2W2 is more than that of SFS and LSBOUND [99]. Fisher selection method provides higher success when compared to PCA [92].

- Apart from these results obtained, wavelet transform provides greater success than that of the other acoustic parameter and feature selection methods, and the use of TEO even more increases the success rate [33], [106].

- When the relationship of sound parameters with emotion state is examined, it is found that fundamental frequency itself, particularly its average value is active on all the emotions. Anger emotion is with high intensity and F0 value. Disgust emotion is with low mean F0 value and medium-high intensity value. Fear emotion is associated with a high F0 level and level of intensity has been increasing. Speaking rate of fear emotion is higher than that of disgust. Joy emotion is with high mean F0 and intensity and its speaking rate has increased. Sadness emotion is with high medium intensity and very low and high mean F0 level.

- In the studies examined, the most precise classification was obtained through Berlin Emotional Database and GMM classifiers [113].

The fundamental problem encountered in the improvement of the success rate in SER has been about the processing of data. In this regard, researchers tend to prefer the databases available where the validity is confirmed by the previous studies. There is certainly a challenging problem for the studies using their own data while it could be difficult to assign the emotions to speech records by using self-expression. Therefore, it has been apparent that, as the experience level of the participant's increases, the number of the participant's decreases.

For creating a future perspective; it is important to state that, most of the SER studies based on acoustic parameters tend to use conventional classifiers. This may be a great opportunity for the researchers to direct their research while the artificial intelligence methods have not been widely considered before.

REFERENCES

1. M. Gerçeker, İ. Yorulmaz, and A. Ural, "Ses ve Konuşma," KBB Ve Baş Boyun Cerrahisi Derg., vol. 8, no. 1, pp. 71–78, 2000.
2. R. T. Sataloff, Treatment of Voice Disorders. San Diego: Plural Publishing, 2005.
3. M. Belyk and S. Brown, "The Acoustic Correlates of Valence Depend on Emotion Family," J. Voice, vol. 28, no. 4, p. 523.e9-523.e18, Jul. 2014.
4. A. B. Kandali, A. Routray, and T. K. Basu, "Vocal emotion recognition in five native languages of Assam using new wavelet features," Int. J. Speech Technol., vol. 12, no. 1, pp. 1–13, Mar. 2009.
5. J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai, "Emotional Audio-Visual Speech Synthesis Based on PAD," IEEE Trans. Audio Speech Lang. Process., vol. 19, no. 3, pp. 570–582, Mar. 2011
6. T. Bänziger and K. R. Scherer, "Introducing the geneva multimodal emotion portrayal (gemep) corpus," Bluepr. Affect. Comput. Sourceb., pp. 271–294, 2010.
7. S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," Speech Audio Process. IEEE Trans. On, vol. 10, no. 2, pp. 65–78, 2002.