IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A CONTENT ANALYSIS OF THE RESEARCH APPROACHES IN SPEECH EMOTION RECOGNITION

Turgut Özseven [*1], Muharrem Düğenci [2], Alptekin Durmuşoğlu [3]

[*1] Department of Computer Engineering, Gaziosmanpaşa University, Tokat, Turkey
[2] Department of Industrial Engineering, Karabük University, Karabük, Turkey
[3] Department of Industrial Engineering, Gaziantep University, Gaziantep, Turkey

DOI: 10.5281/zenodo.1133547

ABSTRACT

Numerous researchers have conducted studies on the recognition of emotion from human speech with different study designs. Speech Emotion Recognition (SER) is a specific class of signal processing where the main goal is to identify the emotional state of people from voice. SER processes are extensively initialized with the extraction of acoustic features from speech signal via signal processing. Subsequent to selection of the most relevant speech features, a model explaining the relations between the emotions and the voice is searched. Effects of acoustic parameters, the validity of the data used, and performance of the classifiers have been the vital issues for emotion recognition research field. In this study, a content analysis of the studies on the SER based on acoustic parameters was performed. 81 articles (published in the indexed journals) have been assessed by the approaches used for emotion labelling, acoustic features and classifiers and the database used. In addition to that analysis, effect of the acoustic parameters on the status of emotion is also extracted as a summary. The main aim of this study is to: describe the features of the databases in use and to create a brief on the efficiency of acoustic parameters and the classifiers employed by the previous studies. Thereby, it is expected to shed light on the study design for the future studies.

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

I. INTRODUCTION

The communication ability, converting the sound to the form of the speech, is the most important aspect that is distinguishing the human from other living human beings. Speech is a complex function which occurs via audio path processing [1]. Speech, in addition to being a communication tool, is also an indicator of a person's identity, mental state and physical health and etc. Therefore, the automatic Speech Emotion Recognition (SER) has a huge potential in applications of fields such as psychology, psychiatry and the affective computing technology [2].

There have been numerous studies focusing on the relation between speech and the personal aspects/emotions. On the other hand, the studies examining the SER studies by the “approaches/classifiers”, “the emotions included”, “acoustic features”, “data sources” has been limited. To some extent, the contextual analysis can be very beneficial to: understand, interpret, analyze and synthesize of a research area.

For this purpose, in this content analysis, 81 papers that were published about SER between 2005 and 2015 (January) were analyzed. Publication search was performed using the search engine of Web of Science (WoS) where indexed publications are listed. Within the scope of this research only on the journal publications were included and conference/symposium papers were excluded. The following keywords were searched and 81 articles were retrieved:

- 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 process of conducting a contextual analysis can help to establish a framework within the research area typical steps of SER analysis covering five phases was presented as a framework shown in Fig. 1. These steps explained in Fig. 1 correspond to the subsections (presented in the parenthesis in the figure) included in this paper. In this regard, the flowchart given in Fig.1. can also be used as a roadmap for the readers.

The *data collection step* contains the acquisition of data regarding to the voice records which will be used in the study. Some researchers prefer the use the data that they collected and some others use existing databases. Various paid and free speech/emotion databases are also available for researchers. It is also seen that the data corpuses, used to reveal the emotional state, are also used intensively in the studies.

Data acquisition does not equally mean that the data is ready to be processed. *Preprocessing step* may be an essential need to make the data ready to be further analyzed. The ones, who are collecting the data for their studies, should convert the speech and audio recordings (that is in analog format) into numeric format for further digital signal processing (DSP). It is also difficult to identify the outputs (corresponding emotions) in SER studies. There are objective and subjective methods that are employed for emotion labelling. Perceptual evaluation is a subjective evaluation method which is simply an interpretation of records by the experts. However, experts may not have the same conclusions about the given records. There are objective evaluation methods that are used to overcome this subjectivity problem [3]. Acoustic analysis has been a widely utilized method to objectively evaluate the speeches which is an inexpensive method providing objective, noninvasive data in a short time. Software packages also exist for acoustic analysis [4] to make analysis easier. In some of the studies; situations that trigger the emotions (emotional stimulus) can be used to investigate the changes in speech.

The main hypothesis in SER analysis indicates that there is a relation between the voice and the emotions. On the other hand, sound signal and audio path features vary by age, gender, body weight and height and the length of the audio path. Therefore, it may be very difficult to distinguish the emotional state among other factors. 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 feature extraction phase of *modelling step*, various acoustic parameters belonging to speech or sound are tested to see whether they have considerable relation with the emotions or not. If the size of the significant features (considerable acoustic parameters) is too large, the number of the features can be reduced by using dimension reduction methods which can reduce computation time. The most favorable dimension reduction methods 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 some of the studies, acoustic features, linguistic and contextual information have been used in combination for the emotion detection [6].

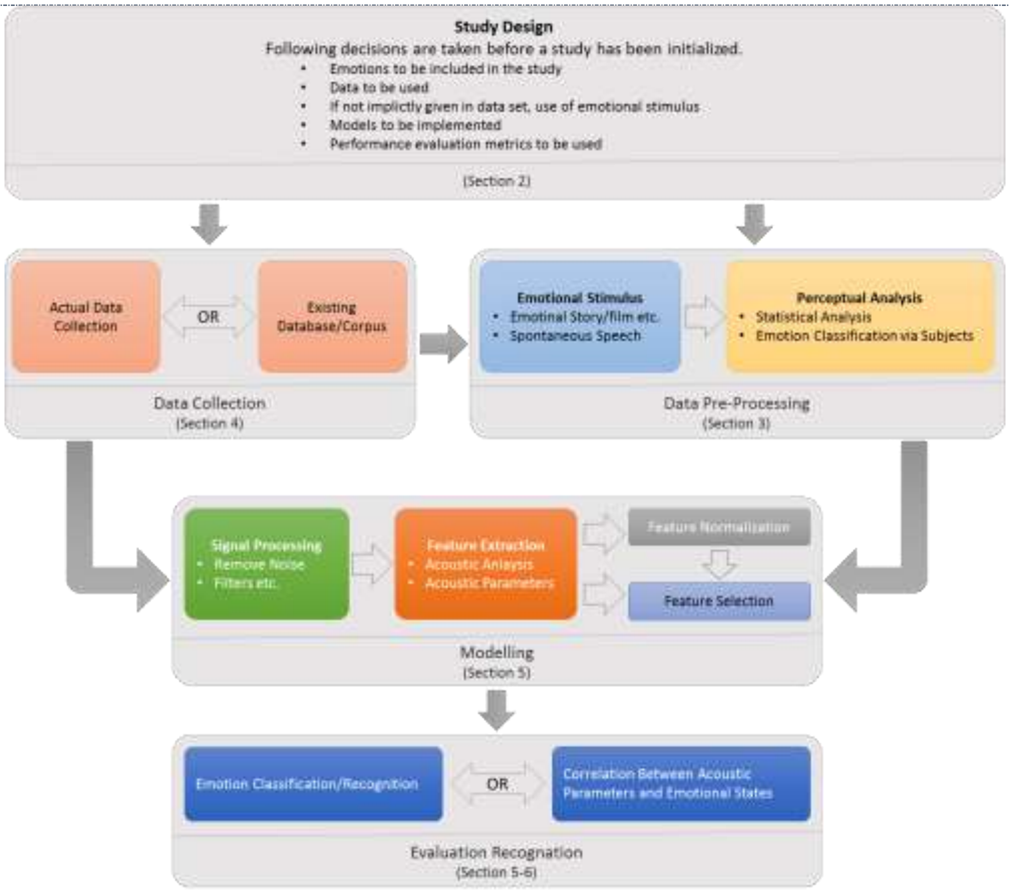

Fig.1. Flowchart of typical speech emotion recognition studies

II. 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 purpose of the study (emotion restricted/demographic restricted)?
- Which data to be used (an existing database, corpus or a new data set 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 participants to the speech recording (which privacy issues should be stated and guaranteed?)
    - 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? (dark room-neural room, quietness of the room)
    - How will those analog speech records be converted into numeric format?
    - Where will these numeric data be stored?
    - How will the speech records be matched with the emotions included? Will there be an expert group to identify the emotion (or an automated methodology be inserted in the analysis)
  - If an existing database or corpus will be used:
    - Is that database reliable?
    - How was the data obtained?

  - Is there a requirement 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 determine and implement if there is a necessity of data elimination (biased data-wrong entry, misdirecting speech)?
- How to eliminate the effects of other factors from the data (noise; effect of age and etc.)?
- Which acoustic parameters to be calculated?
- 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.

III. EMOTIONS USED IN THE SER LITERATURE

The people's emotional state varies depending on the current psychological condition, the environment, or the difficulties occurred in the past. This case is naturally reflected through people's voice, speech and facial expression. 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.

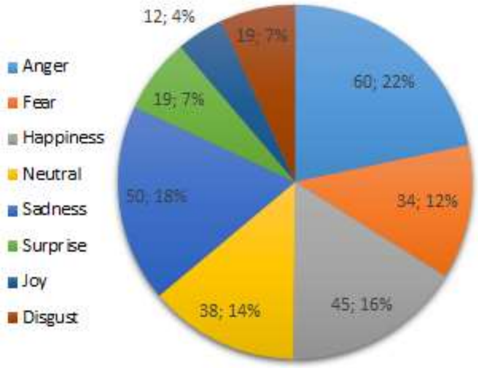

Fig.2. The distribution of the emotions that is under consideration

There are also some other studies that consider some other emotions like: boredom (Albornoz et al., 2011; Milton and Tamil Selvi, 2014, Ramakrishnan and El Emary, 2013; Siegert et al., 2014; Truong et al., 2012; Yang and Lugger, 2010; 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]. Although the researchers may recalibrate the emotion list by adding or removing any particular emotion that they are interested in their study, the fundamental notion remains that emotion is discrete and is able to be quantified using speech [27].

Studies Considering Emotions from Dimensional Perspective

According to the *dimensional perspective of the emotions*; each emotion has gradual membership to the different dimensions such as arousal dimension, potency dimension, intensity dimension and activation dimension. Arousal dimension indicates alertness, excitement or the engagement level of the emotion [28]. An emotion's membership to Valence dimension takes a value in between displeasure and pleasure. Similarly, activation dimension changes between sleep and frenetic excitement. Potency includes cognitive assessment and particularly relates to negative emotions. Intensity indicates the degree of importance for the behavioral and psychological respond of an emotion [29]. Fig.3 [9], [16] demonstrates arousal and valence dimensions of the some emotions.

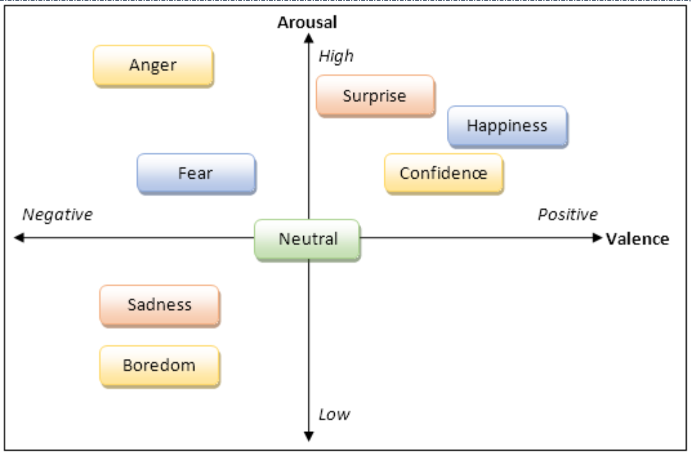

Fig.3. The arousal and valence dimensions of emotions.

Laukka and Elfenbein (2012) have examined the emotions in eight different classes (novelty, urgency, goal conduciveness, power, self-responsibility, norm compatibility, pleasantness, other responsibility) in respect of appraisal dimensions given above, through the intensity of each emotion [25].

Emotional Stimulus and Perceptual Analysis in SER Studies

Even the data on the available databases were labelled according to the emotional states; the sound records conducted through the corpus requires to be labelled with the emotions. In the reviewed studies; reading emotional stories to participants or having participants listen to audio records has been used to create emotional stimulus.

Summary of the participant profiles (listeners) that were employed in perceptual analysis are presented in Table 1. It is apparent that majority of participants in those tests are the ordinary people who has no expertise on the area. Number of participants in the tests where "ordinary people" joined is higher when compared the tests involving expert participants. It can be concluded that experienced participants may avoid the need for participation of higher number of participants, since the consensus can be easily reached.

Table 1. List of SER studies employing perceptual analysis

Reference	**Subjects**	**Reference**	**Subjects**
[29]	30 students, 6 experts	[37]	20 listeners
[30]	64 speakers	[38]	3 labelers
[31]	3 labelers	[39]	48 listeners
[32]	87 listeners	[25]	12 judges
[33]	6 listeners	[21]	20 speakers
[34]	The listeners are divided into 3	[40]	6 annotators
[24]	2 speakers	[41]	20 judges
[35]	4 labelers	[42]	27 annotators
[15]	5 labelers	[20]	14 students
[36]	3 labelers	[43]	Non-expert labelers

IV. 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. Otherwise, misleading results can be obtained. The studies which cover the data collected by their own are as summarized in Table 2. The values in the grids of the table 2, indicates 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 emotions.

Table 2. The distribution of the publications using their own collected data

Primary Emotions Considered	**Corresponding Papers**														
	Speech language used														
	English	Japanese	German	Italian	Chinese	Hindi	Arabic	French	Greek	Israeli	Italian	Farsi	Swedish	Russian	NA
Love	1		2												
Joy	3		4	5	5	6	7	8	9	9	9	10		11	13
Surprise	7		14			15	7								16
Anger	17		18	5	5	6	7	8	9	9	9	10	12	11	19
Fear	20	21	22	5	5	7	7	23							24

[1] [44], [2] [45], [3] [26], [29], [32], [39], [41], [44], [46], [47], [48], [49], [50], [4] [22], [30], [32][50], [5] [51], [6] [32], [34], [7] [32], [8] [22], [50], [52], [9] [22], [10] [53], [54], [11] [55], [12] [37], [13] [56], [8], [16], [20], [36], [57], [14] [30], [32], [45], [15] [32], [34], [16] [8], [36], [17] [29], [32], [39], [41], [44], [46], [47], [48], [50], [58], [59], [18] [22], [30], [32], [45], [50], [19] [56], [8], [16], [20], [36], [57], [20] [29], [32], [39], [41], [46], [48], [49], [58], [60], [21] [60], [22] [30], [32], [23] [52], [24] [8], [20], [57]

There are also some studies introducing novel database. Narayanan and Potamianos have developed a database named "Children's Interactive Multimedia Project (CHIMP)" involving the interaction of the machines with the kids in the games [61]. These include acoustic modeling for automatic speech recognition (ASR), language and dialog modeling, and multimodal-multimedia user interface design. Acoustic modeling adaptation and vocal tract normalization algorithms that yielded state-of-the-art ASR performance on children's speech are described [61].

Another database was introduced by [31] to collect the data indicating excessive emotional symptoms in the life-threatening situations. With this purpose, Clavel et al. has 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). Participants were given (in five different native languages) to read short sentences covering the emotions: anger, disgust, fear, happiness, sadness, surprise and neutral.

[62] introduced two databases: BHUDES- *Beihang University Database of Emotional Speech* (containing emotions such as, happiness, anger, disgust, fear, sadness, surprise and neutral) and BHUDEP-*Beihang University Database of Emotional Points* (containing emotional points in the face in 2012).

[21] have developed a corpus containing emotions such as achievement/ triumph, amusement, sensual pleasure, and relief, obtained through two women and two men participants. List of databases available to be used in SER studies are illustrated in Table 3.

Table 3. List of databases available to be used in SER studies

Corpus	**Access**	**Language**	**Size**	**Emotions**
VAM[1] [63]	Public and free	German	1018 emotional utterances by 47 speakers	Valence, Activation, Dominance
HUMAINE[2] [64]	Public and free	English	25 Audio-visual recordings - 4 Speakers	Anger, Happiness, Sadness, relaxed
GEMEP[3] [65]	Public and free	French	10 Actors - 5 Females, 5 Males	Amusement, Anxiety, Anger, Despair, Fear, Interest, Joy, Pleasure, Pride, Relief, Sadness, Admiration, Contempt, Disgust, Surprise, Tenderness

FAU Aibo Emotion[4] [66]	Public with license fee	German	51 children interacting with a Sony toy robot	Anger, Emphatic, Neutral, Positive, Rest
Berlin Emotional Database(EMO-DB)[5] [67]	Public and free	German	800 Utterances – 10 Actors	Anger, Joy, Sadness, Fear, Disgust, Boredom, Neutral
IEMOCAP[6] [68]	Public and free	English	5 Sessions -2 Actors (12 hours of data)	Happiness, Anger, Sadness, Frustration, Neutral
SUSAS[7] [69]	Public with license fee	English	16.000 Utterances - 32 Actors(13 Females + 19 Males)	Four Stress Styles
SEMAINE[8] [70]	Public and free	English	25 Recordings - 21 Participants	Anger, Disgust, Amusement, Happiness, Sadness, Contempt
Danish emotional database[9] [71]	Public with license fee	Danish	4 Actors - 2 Words - 9 Sentences - 2 Passages	Anger, Joy, Sadness, Surprise, Neutral
eNTERFACE[10] [72]	Public and free	English	42 Subjects – 14 Different Nationalities	Anger, Disgust, Fear, Happiness, Sadness, Surprise
BHUDES[11] [73]	Private	Mandarin	7 Actors - 20 Utterances	Anger, joy, sadness, disgust, surprise

[1] Emotion Research Group at the Institut für Nachrichtentechnik of the Universität Karlsruhe, Karlsruhe, Germany.
[2] Queen's University Belfast, Belfast, Northern Ireland, United Kingdom.
[3] Centre Interfacultaire en Sciences Affectives (CISA) at the University of Geneva.
[4] Head of the Medical Image Segmentation group at the Pattern Recognition Lab of the Friedrich-Alexander University Erlangen-Nuremberg
[5] Institute for Speech and Communication, Department of Communication Science, the Technical University, Germany.
[6] Speech Analysis and Interpretation Laboratory at the University of Southern California.
[7] Linguistic Data Consortium, University of Pennsylvania, USA.
[8] The SEMAINE database was collected for the SEMAINE-project by Queen's University Belfast.
[9] Department of Electronic Systems, Aalborg University, Denmark.
[10] TCTS Lab. of Faculte Polytechnique de Mons.

Fig. 4 illustrates the distribution of the databases that were employed by the SER studies (among 81 publications). The databases that were employed by less than three studies were given under the "other" category. (In summary: HUMAINE:2, ChIMP:1, German IVR:1, VENEC:1, SUSAS:1, Danish Em.:1, eNTERFACE:2, BHUDES:1). Berlin Emotional Database has been the most widely used one. VAM, GEMEP and FAU Aibo Emotion databases followed it respectively.

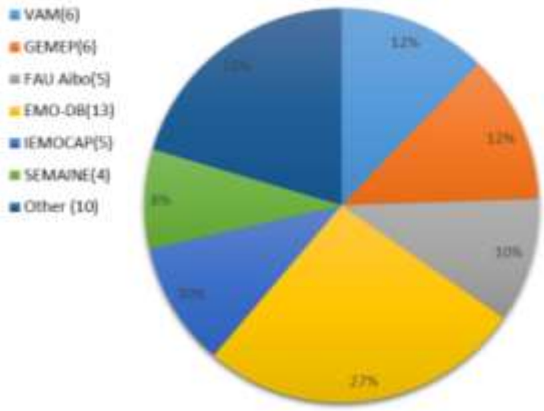

Fig.4. The distribution of the databases used.

Although there are several databases that contain the emotional speech, their accuracy of emotion labelling may be a misleading factor. Accuracy of labelling may be affected from the simulation of the emotions or the factors like recording environment, sound recorder, sound quality, and actors used for record may distort the labelling performance [74].

V. 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. Acoustic features are a set of parameters that are widely used for emotion recognition purposes. 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 4.

Table 4. The acoustic parameters used in the sound analysis.

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].	

The usage frequency of the acoustic parameters (employed in ten or more publications) is as given Fig.5. Twenty seven different types of acoustic parameters were used in the reviewed studies. Twelve of them have been very common in use as depicted in Fig.5.

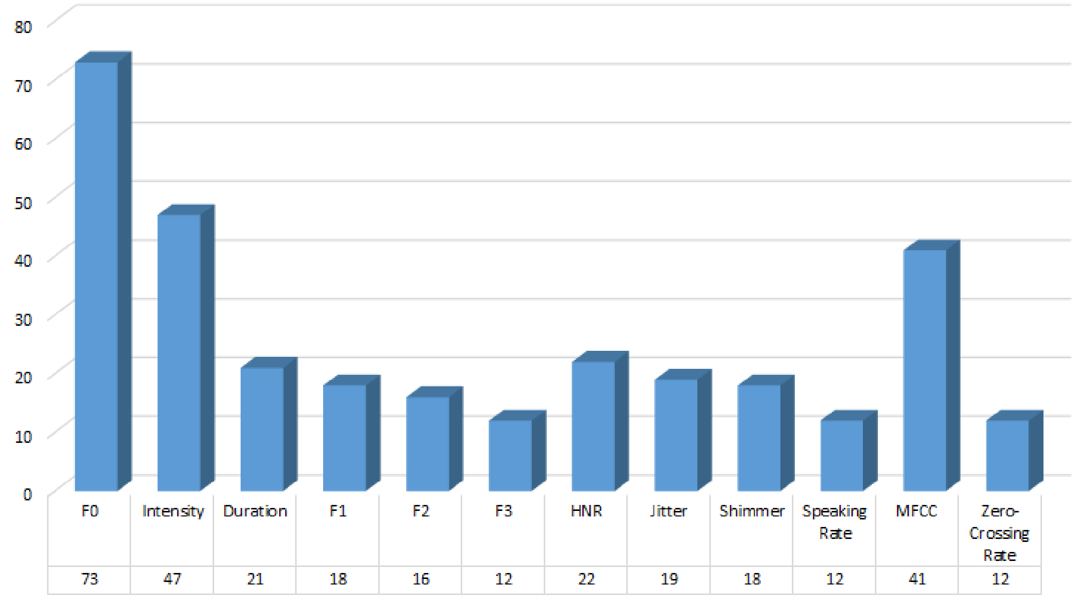

Fig.5. The frequency of the acoustic parameters used

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. It is remarkable to state that, noise removal methods should not result in a change in the main structure of the signal while it is been removed. 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 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, 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 designed tools for the calculation of acoustic parameters the pre-processing phases to be executed on the speech signals. These tools and acoustic features that can be detected by them are given in Table 5. PRAAT has been used in the vast majority of the studies (Agrawal et al., 2010; Bänziger et al., 2014; Bejani et al., 2014; Belyk and Brown, 2014; Busso and Narayanan, 2007; Coutinho and Dibben, 2013; Curtis and Bharucha, 2010; Diamond et al., 2010; Goudbeek and Scherer, 2010; Hoque et al., 2006; Jia et al., 2011; Lanjewar and Chaudhari, 2013; Laukka et al., 2011, 2005; Laukka and Elfenbein, 2012; 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; Pell and Kotz, 2011; Pérez-Espinosa et al., 2012; Ramakrishnan and El Emary, 2013; Rochman et al., 2008; Scherer, 2013; Scherer et al., 2015; Szameitat et al., 2009; Truong et al., 2012).

Table 5. Feature extraction tools.

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

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

$$x_n = \frac{x - \mu}{\sigma} \qquad (1)$$

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

Feature normalization techniques used in the studies are given in Table 6 [89]

Table 6. Summary of feature normalization approaches for speech emotion recognition [89].

Method	Normalization	Literature
z-score	Global, speaker-dependent and, speaker and corpus-dependent normalization	[6], [28], [80], [37], [38], [90], [52], [8], [40], [79], [43], [23]
Min-max	Global and speaker-dependent normalization	[31]
Min-max and z-score		-
Zero mean	Global normalization	-
Exponential transformation		-
Divide energy by energy mean	Speaker-dependent normalization	-
Divide energy by energy peak		-
Cepstral mean subtraction		-
Feature warping		-
Whitening		[91]
Divide each feature by its mean	Speaker and emotion-dependent normalization	-
Relative feature	Speaker, text and emotion-dependent normalization	[6]

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.

Principal Component Analysis (PCA)

Principal component analysis (PCA) is used to find a subspace whose basis vectors correspond to the maximum-variance directions in the original space [92]. PCA unsupervised feature reduction method has been extensively used in the context of speech emotion recognition.

It is assumed that covariance matrix representing the relationship between the feature vectors in PCA method, equals the multiplication of eigenvector and eigenvalues. Eigenvectors obtained by this way are considered to be new basic components, and new data components are calculated. If PCA is used to reduce the dimension, the dimension with the minimun deviation is removed, and if necessary, the data is converted back to its own size.

Arousal and valence ratings were not included, to ensure that any association between the resultant factors and these dimensions was not biased [93].

PCA is the mostly used method to reduce the dimension of the acoustic features set obtained after signal processing [6], [23], [28], [37], [56], [60], [80], [92], [94], [95].

In the result of the performance test of the feature sets and classifiers using three feature set (PCA, f10-10 best features and f15-15 best features) and two different classifiers (LDC and k-NN), it is seen that, the lowest classification error in men with LDC classifier has been found in PCA in the result of the experiment performed with only acoustic feature [6]. By the comparision of results obtained both via the features obtained after feature selection via PCA and via use of all the features, it is found that, the results with all of the features yielded better results than the use of PCA [94]. However, this is a special case, and in many studies, an counter case is seen.

Linear Discriminant Analysis (LDA)

LDA is a method decreasing the dimension by maximizing the linear discrimination of the groups, which belong to different groups in data. LDA is a supervised feature reduction method searching for the linear transformation that maximizes the ratio of the determinants of the between-class covariance matrix and the within-class covariance matrix [96].

Hoque et al. reflected the features in the low dimensional space by using PCA and Linear Discriminant Analysis (LDA) for compact and clustered representation of the features at their study in 2006. They also determined that using PCA and LDA together yielded better results than using seperately [7].

Sequential Forward Selection (SFS)

While SFS creates a specific set of features, it performs adding a feature to a subset of features at each step. The selection criteria are the success rate of the classifier algorithm. Since algorithm needs a classifier algorithm at each step and operates through all the search process, it causes to slowdown the algorithm performance. The most important advantage of the algorithm is the success rate to achieve the solution space [97].

Sequential floating forward selection (SFFS) is an improved SFS method in the sense that at each step, previously selected features are checked and can be discarded from the optimal group to overcome nesting effects. Experiments show SFFS to be superior to other methods [15].

LSBOUND, MUTINF and R2W2

Zhou and Mao proposed a new feature selection algorithm to be used in gene selection problem for DNA micro-arrays [98]. In this algorithm, an evaluation criterion similar to the use of the filter approach is proposed. These criteria are called as Least Squared Bound (LSBOUND). The method combines the speed advantage of filter method with high classifier advantage of wrapper method. Criteria used are based on the idea of obtaining LOOCV upper bound error analytically for LS-SVM (Least Squares Support Vector Machines), and using this value in the selection of attributes [99].

The mutual information (MUTINF) is a widely used information theoretic measure for the stochastic dependency of discrete random variables. In the context of a filter approach, one may employ mutual information to discard irrelevant features in order to find a small subset of features on the basis of low values of mutual information [99].

R2W2 is a state of the art feature selection algorithm especially designed for binary classification tasks using an SVM classifier [100]. It can be considered as a wrapper approach and indirectly exploits the maximal margin principle for feature selection. The idea in this approach is to find a weight vector over the features in order to minimize the objective function of an SVM [99].

Fisher Selection

Fisher selection algorithm [101], is an statistical method used frequently to obtain the information belonging to the individual attribute. Its measurement method uses the arithmetic average of the numerical data and standard deviation values of each attribute for each class.

$$FKS(x_i) = \frac{|u_i^+ - u_i^-|}{\sigma_i^+ + \sigma_i^-} \qquad (2)$$

In this equation, the + and – marks, refers to the different classes for a problem with different classes. It indicates the arithmetic average of the values calculated for each attribute and class. By this method it is possible removal of noisy data to obtain a subset of attributes with the large data sets that [102].

Clavel et al., (2008) reduced features range in two steps by the selection of optimal fourty features by using fisher selection algorithm in their study. The first election for each feature group (prosodic, voice quality, and spectral) was performed separately. Thus, 20% of nearly one hundred features including features from each feature class are selected. In the last step it is applied again to the features selected by fisher selection algorithm [31].

Chen et al. (2012), performed four comparative experiments, for which two feature demotion method (Fisher and PCA) and two classifiers (SVM and ANN) were used, and as for tested emotion recognition accuracy rates. It is found in the experimental results that Fisher was better than PCA in reducing the size. The highest accuracy rate in the all experiments performed, was achieved with the Fisher + SVM [92].

Fast Correlation Based Filter (FCBF)

The FCBF method is used for the dimension reduction and the construction of a lower-size feature space. This method selects the features that are individually informative and two-by-two weakly dependent. It is noted that the mutual information (MI) of two vectors *X* and *Y*, *I(X,Y)*, computes the statistical dependency of them in the following way [53], [54]:

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(X = x, Y = y) \log \left(\frac{p(X = x, Y = y)}{p(X = x)p(Y = y)} \right) \quad (3)$$

Where *p* is the probability function. Obviously, *I(X,Y)* is equal to 0; when *X* and *Y* are independent (*p(X = x, Y = y) = p(X = x)p(Y = y)*) and is increased when their dependency increases.

In FCBF method, Y is the vector of data labels and X_i is the vector of i_{th} feature value for all data.

Gharavian et al. (2011), performed nine experiments including FCBF feature selection, and selected 2, 8, 15, 21, 27, 32, 36, 42 and 47 features respectively, from a set of 52 features [53]. The best result was obtained at the experiment with 27 features.

Gharavian et al. (2013) have carried out experiments with six and ten features from a set of 55 features with FCBF and FCBF. Average accuracy is 79% when a set of 55 features is used. Average accuracy with 6 and 10 features by FCBF is 79.9% and 77.9% respectively. Average accuracy with 6 and 10 features by ANOVA is 79.3% and 79.4% respectively [54].

Wavelet Transform

Wavelet transform provides both low and high frequency components of the signal, using variable sized windowing technique in signal processing [103]. With this method, analysis of systems with frequency changeable by time and transient state analysis are carried out in quite sensitive way [104]. Kandali et al. (2009), in their study, used four different methods,(WPCC2-Wavelet Packet Cepstral Coefficients computed by method 2, MFCC-Mel-Frequency Cepstral Coefficients, tfWPCC2-Teager energy operated in Transform domain WPCC2 and tfMFCC), two of which is Wavelet-based. Wavelet-based WPCC2 and tfWPCC2 methods have yielded better results than MFCC and tfMFCC methods. TEO (Teager Energy Operator) has raised its average score for all use cases [33].

Time-frequency parameters obtained using Wavelet transform and continuous wavelet transform are used to ascertain the accent and the intonation of the sentence during the recognition process [105], and to provide additional features to the general acoustic features [82], [106] and to create the new wavelet filter-based feature vectors [13]. Wavelet filter-based feature vector containing twelve coefficients provides a higher classification accuracy comparing to MFCC and TAYLOR-based features [13]. classification accuracy of Wavelet Packet Cepstral Coefficients (WPCC) method is higher than MFCC [106].

Traditional Statistical Analysis

For the determination of the relationships between the features of the publications reviewed, participants, of whom perceptual analysis was performed, features, traditional statistical methods of the Pearson correlation, ANOVA, MANOVA, Kappa Statistic and Binary Logistic Regression were used.

Pearson correlation is used to measure the relationship between the two independent variables. ANOVA is used to analyze how these independent variables interact among themselves and analyze the effects of those interactions on the dependent variable. MANOVA is used to analyze the effects of two or more independent variables on more than one dependent variable. Binary Logistic Regression aims to define the relationship between dependent and independent variables with minimum variables. Kappa statistics measures the comparative compliance reliability of two evaluators [6], [24], [31], [35], [38], [94], [107].

Through ANOVA analysis performed using features of emotion and intensity as dependent variable and emotion and dimension as independent variables, the relationship between intensity and emotion was determined, on the other hand, any relationship between emotion and intensity wasn't determined for activation, valence and potency [29]. The conducted an ANOVA with repeated measures on domain (music or speech), intensity (loud or soft), rate (fast or slow), and pitch height (high or low) [108]. For valence, there were significant interactions between domain and pitch height, domain and rate, and domain, intensity, pitch height, and rate [108]. For energy arousal, there were significant interactions between domain and intensity, and domain and rate [108]. For tension arousal, there were significant interactions between domain and intensity, and domain and rate [108]. The primary acoustical measurements (mean pitch, intensity, duration) were entered in a series of one-way ANOVAs and results revealed significant differences across emotional categories for mean pitch, mean intensity, and mean duration [30]. The conducted a two-way mixed model analysis of variance (ANOVA), using order of induction as a between subjects independent variable and type of emotion as a within subjects variable [59].

The conducted separate within-participant ANOVA to determine whether the speech samples differed according to the intended emotion of the speaker for each acoustic parameter [47]. The differences in these measures due to emotion were tested using individual repeated measures ANOVAs for each acoustic parameter [80]. ANOVAs used the stimulus as the unit-of-analysis (averaging across the listeners' mean ratings), and compared values across the 15 emotions, with separate ANOVA models for each appraisal dimension [25]. The significant variability on valence and arousal scales across emotion categories was confirmed by two ANOVAs [21]. An ANOVA of the intensity levels for the different emotion alternatives shows that speakers indicated being significantly happier and more satisfied in the positive induction condition than in the negative one [50]. The average voice ratings used had yielded significant effects for an Emotion factor in repeated measures ANOVAs [43]. They removed the two acoustic parameters that did not show significant effects for the Emotion factor in repeated measures ANOVAS [43]. To reduce the number of features, a feature selection method, based on the

ANOVA, was used [84]. This feature selection process resulted in a range of 40–60 features for each binary classifier per cross validation fold [84].

Pearson examined the speech and facial expressions via correlation, aiming to examine articulation and emotions based on the interaction between the face gestures and speech, and the results indicated that there is a strong relationship between face and acoustic features [56]. In the another study, Pearson correlation method was used to examine the relationship between twelve parameters obtained through speech samples and the parameter consistency between each speaker and the other speakers [52].

Diamond et al., (2010) examined the utility of 2 emotion-focused interventions relational reframes and empty-chair enactments in terms of arousing primary sadness associated with loss and longing among individuals suffering from unresolved anger. A multivariate analysis of variance (MANOVA) with stage of the session serving as the independent repeated measure and vocal acoustic parameters serving as dependent measures was conducted to determine whether participants' vocal acoustical profiles varied across baseline, relational reframe, and empty-chair enactment [58].

Lee et al., (2011) perform feature selection on the 384 features using the standard statistics software SPSS to obtain a reduced feature set. They used binary logistic regression in SPSS with step-wise forward selection [90].

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 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 8 [109].

Table 7. The characteristic of acoustic features for five emotions

	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

As seen in Table 7, 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 8 [110].

Table 8. Acoustic features characteristics for six emotion states.

	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 9 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 9 [5]

Table 9. Summary of the effects of several emotion states on selected acoustic features [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	<	<	<	↗	<	<	$>_{M}, <_{F}$	>

Explanation of symbols: >: increases, <: decreases, =: no change from neutral, ↗: inclines, ↘: declines. Double symbols indicate a change of increased predicted strength. The subscripts refer to gender information: M stands for males and F stands for females.

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 disgus, 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 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 the combination of acoustic, lexical and contextual informations have yielded better results. 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 rages and men [24]. Peak performance has been obtained over MFCC feature set [24]. When the difficulty of detecting State of emotions is examined, it is found that the detection of anger and sadness emotions is the easiest and that of fear is the most difficult [62]. In the study where ANOVA analysis of acoustic parameters and emotional states are included, it is seen that, the increase in F0 is meaningful only for joy, and the increase in F1 is meaningful for F1, and differences for F2 is not found meaningful for F2. The changes in the duration are meaningless [55]. The first syllable of the sentence including fear emotion is expressed in a diffrent way comparing online learning based study performed with the emotional speech recognition [16].

Anger: In the studies performed under anger emotion state, it is found that F0 is high, mean F0 is very high, max F0 is very high, min F0 is high, standard deviation of F0 is medium, and F0 range is very high [9], [21], [32], [34], [36], [41], [45]–[48], [51], [52], [55], [59], [60]; voice quality is high [11], [21], [46], [48], [60]; duration is very low in some studies [36], [45], [48] and high in some studies [21], [47]; mean intensity and standard deviation intensity are high [9], [21], [36], [46]–[48], [51]; pause and speaking rate are medium [32], [51]; jitter and shimmer are high [52]; F1, standard deviation is F1, F2 and standard deviation of F2 is very high [55]. The men speaking rate value is lower than women [9].

Fear: In the studies performed under fear emotion state, it is found that max F0 is very low, min F0 is medium, mean F0 is high, standard deviation of F0 is medium, F0 range is medium [9], [21], [32], [41], [46], [48], [49], [51], [52], [58]; voice quality is low [21], 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]; jitter and shimmer parameters are low [52].

Happiness: In the studies performed under Happiness emotion state, it is found that mean F0 is so high, standard deviation of F0 is high, range F0 is medium, max F0 is so high, min F0 is high [32], [34], [36], [41], [46]–[49]; voice quality is low [46]; 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 studies performed under Happiness emotion state, it is found that mean F0 is very low and high, max F0 is high, min F0 is medium, standard deviation of F0 is medium, range 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]; jitter and shimmer parameters are high [52].

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].

Surprise: In the studies performed under Surprise emotion state, it is found that mean F0 is so high, max F0 is so high, min F0 is high, standard deviation of F0 is high, range F0 is so high [32], [34], [36]; duration and intensity are medium [36]; speaking rate is so high [32].

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 studies performed under Joy emotion state, it is found that mean F0 is very low, max F0 is very low, range 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].

Table 10 is given to improve intelligibility of relationship between emotion status and the acoustic parameters.

Table 10. The relationship between emotional states and acoustic parameters.

Acoustic Parameter	**Anger**	**Fear**	**Happiness**	**Sadness**	**Neutral**	**Surprise**	**Joy**	**Disgust**
MF0	>>	>	>>	<< >	<<	>>	>	<<
sdF0	O	O	>	O	<O	>	>	>
MxF0	>>	<<	>>	>	NA	>>	>	<<
MnF0	>	O	>	O	NA	>	NA	NA
RF0	>>	O	O	>	<<	>>	NA	<
VQ	>	<O>	<	< >	NA	NA	< >	<O
D	<< >	<	O	< >	NA	O	O>	O
Mint	>	>	>	>	NA	O	>	O>
sdInt	>	>	O	<O	NA	NA	NA	O>
P	O	O	NA	>	NA	NA	O	NA
SR	O	O	>	<O	>	>	>	<
F1	>>	NA	NA	NA	O	NA	< >	NA
sdF1	>>	NA	NA	NA	O	NA	>>	NA

F2	>>	NA	NA	NA	O	NA	< >	NA
sdF2	>>	NA	NA	NA	O	NA	>	NA
Jt	>	<	NA	>	O	NA	<	NA
Sh	>	<	NA	>	O	NA	<	NA

MF0: mean F0, sdF0: standard deviation F0, MxF0: max F0, MnF0: Min F0, RF0: range F0, VQ: voice quality, D: duration, Mint: mean intensity, sdInt: standard deviation intensity, P: pause, SR: speaking rate, sdF1: standard deviation F1, sdF2: standard deviation F2, Jt: Jitter, Sh: shimmer, “<<” very low, “<” low, “O” , “>” high, “>>” very high

When Table 10 is considered, it is apparent that the average value of “base frequency” has affect over all emotions. Voice quality value may not be used effectively for emotion recognition because of its unpredictability. Duration, speaking rate, jitter and shimmer parameters can be used as decisive, whereas other parameters can be used as supportive. F1 and F2 parameters can be used for diagnosing emotion. Anger is the emotion with high intensity and F0 value. Disgust is the emotion with medium-high intensity and low mean F0 value. Fear is associated with a high level of F0 and the intensity level rises. Fear is higher than disgust in terms of speaking rate value. Joy is the emotion with high mean F0 and intensity and increased speaking rate. Sadness is the emotion with high medium intensity and mean, very low and with high F0 level.

VI. 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 11.

Table 11. Distribution of the classifiers used in the reviewed publications

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

It is seen that at Table 11 that; SVM, GMM, HMM, K-NN and Bayesian Classifier are the most common classifiers.

The Classifiers Performance

When emotional speech recognition studies are examined, it is found that, various studies have come up using with different classifiers, different extraction methods and their combinations. In this section, data collection methods of most commonly used classifiers such as SVM, GMM, HMM, k-NN and Bayesian, and comparison of their performances with data collection method used and feature extraction methods, are given in Table 12.

Table 12. Classification performance of popular classifiers for speech emotion recognition.

Classifier	Data Collection	Feature Extraction/Selection	Acoustic Parameter(s)	Emotion(s)	Average Classification Accuracy (%)
SVM	EMO-DB	SFS	F, I, M	ES, EH, EN, ES	83.4% [99]
		LSBOUND			83.7% [99]
		MUTINF			82.5% [99]
		R2W2			83.7% [99]
		Wavelet Transform	Zc, F, M	EA, EH, EN, ES	90.9% [105]
		NA	F, Ft, M	EA, EF, EH, EN, ES, ESr, EJ, ED, EB	90.5% [9]
		NA	Nf	EA, EB, EH, ES, EN, ED, EF	86.3% [113]
	FAU Aibo	SFFS	P, I, D, F, H, M, J, S, Ft	EA, EN	63.0% [15]
	German IVR	IGR (information gain ratio)	F, M, Ft, I	EA	77.7% [38]
	English IVR				77.5% [38]
	German WoZ				73.3% [38]
	IEMOCAP	LFFS (Linear Floating Forward Selection)	F, I, H, M	EA, EH, EN, ES	82.8% [81]
		Binary Logistic Regression	F, I, Zc, H, M	EA, EH, EN, ES	51.0% [90]
		Correlation Feature Selection	I, Zc, F, J, S	EA, EF, EH, EN, ES, ESr	72.0% [40]
		NA	M, I, Zc, F, J	EA, EF, EH, EN	56.3% [91]
	BHUDES	Fisher	F, I, Zc, F1, M	EA, EF, EH, ES, ESr, ED	50.3% [62], 50.1% [92]
		PCA			43.2% [92]
	DAS	NA	F, Ft, M	EA, EF, EH, EN, ES, ESr, EJ, ED, EB	78.5% [9]
GMM	Data Collection	FFT spectral entropy	-	EA, EH, EN, ES	77.9% [114]
	MESDNEI	NA	WPCC2, MFCC, tfWPCC2, tfMFCC	EA, EF, EH, EN, ES, ESr, ED	90.5%, 83.3%, 100.0%, 88.1% [33]
	Data Collection	NA	F, I, D, M	EA, EN, EP, EFr	73.2% [35]
	EMO-DB	NA	M, I, F	EA, EF, EN, ES,EJ,ED,EB	63.5% [12]
		NA	Nf	EA, EB, EH, ES, EN, ED, EF	92.5% [113]
		NA	F, M, W	EA, EF, EH, EN, ES, ESr	66.0% [82]
		NA	Nf, M, T	EA, EF, EH, EN, ES, ED, EB	68.1% (Nf), 61.3% (M), 50.4% (T) [13]
	SUSAS	NA	Nf, M		64.0% (Nf), 61.0% (M), 54.3% (T) [13]
	VAM	NA	Nf	EA, EB, EH, ES, EN, ED, EF	60.2% [113]
	Data Collection	MIC (Maximal information coefficient)	F, I, Zc, Sr, J, S, D, M, H, F1, F2, F3	EA, EH, EN, ES	68.6% [16]

	Farsi emotional speech	FCBF, ANOVA	F1, F2, F3, F, M	EA, EH, EN	92.2%, 92.3% [54]
HMM	Spanish Emotional Speech	NA	M, F	EA, EH, EN, ES, ESr	86.8% [116]
	IEMOCAP	Binary Logistic Regression	F, I, Zc, H, M	EA, EH, EN, ES	50.7% [90]
	EMO-DB	NA	M, I, F	EA, EF, EN, ES, EJ, ED, EB	68.6% [12]
		PCA	M, F, H, W	EA, EF, EH, EN, ES, ESr	92.2% [95]
		NA	F, Ft, M	EA, EF, EH, EN, ES, ESr, EJ, ED, EB	84.5% [9]
		NA	F, M, W	EA, EF, EH, EN, ES, ESr	67.8% [82]
	DAS	NA	F, Ft, M	EA, EF, EH, EN, ES, ESr, EJ, ED, EB	68.5% [9]
k-NN	ChIMP	NA	F, I, Zc, H, M	EA, EF, EH, EN, ES, ED, EB	77.3% [10]
	EMA	PCA	F	EA, EH, EN, ES	83.5% [94]
	Data Collection	FS (Forward Selection)	F, I, D, F1, F2	E+, E-	80.6% [6]
	EMO-DB	NA	F, M, W	EA, EF, EH, EN, ES, ESr	53.2% [82]
Bayesian	GEMEP	NA	F, D, M, I, P	EA, ES, EJ	57.1% [22]
	EMO-DB	SFFS	F	EA, EH, EN, ES, EB	73.5% [11]
	IEMOCAP	Binary Logistic Regression	F, I, Zc, H, M	EA, EH, EN, ES	58.5% [90]
	Data Collection	Correlation Feature Selection	F, M, J, S	EA, EH, EN, ES	51.8% [42]

P: pauses, I:intensity, D: duration, F:F0, H:HNR, M: MFCC, J: Jitter, S: shimmer, Ft: formants, Zc: zero-crossing rate, Sr: speaking rate, F1: first formant, F2: second formant, F3: third formant, W: wavelet, Nf: New feature set, T: TEO, EA: anger, EH: happiness, EN: neutral, ES: sadness, EF: fear, ESr: surprise, EJ: joy, ED: disgust, EB: boredom, EP: polite, EFr: frustrated, E+: positive, E-:negative

As it can be concluded from Table 12;

- As a result of the use of the same classifier and different methods of feature selection on the same database, the highest success for feature extraction has been achieved with wavelet transform [99], [105].
- Fundamental frequency has been used effectively in all the results.
- The database used affects classification accuracy [38].
- The most precise classification accuracy has been obtained with the GMM over EMO-DB [113].
- Fisher feature selection method is superior than the PCA [92].
- Binary Decision Tree method based on Bayesian Logistic Regression with IEMOCAP database has provided success higher than traditional Bayesian Logistic Regression, SVM and HMM [90].
- GMM classifier accuracy rate is higher than the HMM's and SVM's [12], [113].
- SVM classifier accuracy rate is higher than HMM's [9].
- SVM-RBF classifier accuracy rate is higher than Bayesian's [42].
- When surprise emotion included, GMM performance declines and HMM provides higher accuracy rate. Both fear and surprise emotions decrease the K-NN accuracy rate [82].
- When the SVM and ANN performance compared on BHUDES database, it's seen that SVM accuracy rate higher than ANN's [92].

In the light of this information, GMM and SVM classifiers are used heavily in the studies. Since HMM classifier provides higher success than GMM and SVM in terms of emotional state and emotion number, it should be taken into consideration for classifier selection in emotional states.

VII. 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 (labelled) the emotions by perceptual analysis and listening tests before the acoustic analysis.
- In some of the SER studies, existing databases were used and in some of the other studies; data was 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 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, MUTINF 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].
- A part 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 acoustic 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.
- The most common classifiers are SVM, GMM, HMM, K-NN and Bayesian classifiers. In the studies examined, the most precise classification was obtained through Berlin Emotional Database and GMM classifiers [113].

The fundamental problem encountered in improvement of the success rate in SER, has been about 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 participants increases, the number of the participants decreases.

It is also very interesting to state that, the emotions included by the studies may affect the performance of the classifiers. [82] finds that if the emotion "surprise" is included in a study, performance of the GMM classifier declines, and HMM gets higher accuracy.

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. One another future research focus can be extended by the addition of novel feature sets by the combination of different acoustic parameters since recent trends in research of audio emotion recognition emphasized the use of combination of different features to achieve improvement in the recognition performance [117].

Compliance with Ethical Standards

Conflict of Interest Turgut Özseven, Muharrem Düğenci, and Alptekin Durmuşoğlu declare that they have no conflict of interest.

Informed Consent Informed consent was not required as no human or animals were involved.

Human and Animal Rights This article does not contain any studies with human or animal subjects performed by any of the authors.

VIII. 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] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, “New approach in quantification of emotional intensity from the speech signal: emotional temperature,” *Expert Syst. Appl.*, vol. 42, no. 24, pp. 9554–9564, Dec. 2015.

[3] K. M. Okur E., “CSL ve Dr. Speech ile ölçülen temel frekans ve pertürbasyon değerlerinin karşılaştırılması,” *KBB Ihtis. Derg.*, vol. 8, pp. 152–157, 2001.

[4] R. T. Sataloff, *Treatment of Voice Disorders*. San Diego: Plural Publishing, 2005.

[5] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, Sep. 2006.

[6] Chul Min Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[7] M. E. Hoque, M. Yeasin, and M. M. Louwerse, “Robust recognition of emotion from speech,” in *Intelligent Virtual Agents*, 2006, pp. 42–53.

[8] K. P. Truong, D. A. van Leeuwen, and F. M. G. de Jong, “Speech-based recognition of self-reported and observed emotion in a dimensional space,” *Speech Commun.*, vol. 54, no. 9, pp. 1049–1063, Nov. 2012.

[9] S. Ramakrishnan and I. M. M. El Emary, “Speech emotion recognition approaches in human computer interaction,” *Telecommun. Syst.*, vol. 52, no. 3, pp. 1467–1478, Mar. 2013.

[10] A. Milton and S. Tamil Selvi, “Class-specific multiple classifiers scheme to recognize emotions from speech signals,” *Comput. Speech Lang.*, vol. 28, no. 3, pp. 727–742, May 2014.

[11] B. Yang and M. Lugger, “Emotion recognition from speech signals using new harmony features,” *Signal Process.*, vol. 90, no. 5, pp. 1415–1423, May 2010.

[12] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, “Spoken emotion recognition using hierarchical classifiers,” *Comput. Speech Lang.*, vol. 25, no. 3, pp. 556–570, Jul. 2011.

[13] L. Zao, D. Cavalcante, and R. Coelho, “Time-Frequency Feature and AMS-GMM Mask for Acoustic Emotion Classification,” *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 620–624, May 2014.

[14] I. Siegert, D. Philippou-Hübner, K. Hartmann, R. Böck, and A. Wendemuth, “Investigation of Speaker Group-Dependent Modelling for Recognition of Affective States from Speech,” *Cogn. Comput.*, vol. 6, no. 4, pp. 892–913, Dec. 2014.

[15] A. Batliner *et al.*, “Whodunnit – Searching for the most important feature types signalling emotion-related user states in speech,” *Comput. Speech Lang.*, vol. 25, no. 1, pp. 4–28, Jan. 2011.

[16] C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha, and L. Zhao, “Practical Speech Emotion Recognition Based on Online Learning: From Acted Data to Elicited Data,” *Math. Probl. Eng.*, vol. 2013, pp. 1–9, 2013.

[17] M. Kockmann, L. Burget, and J. “Honza” Černocký, “Application of speaker- and language identification state-of-the-art techniques for emotion recognition,” *Speech Commun.*, vol. 53, no. 9–10, pp. 1172–1185, Nov. 2011.

[18] S. Planet and I. Iriondo, “Children’s Emotion Recognition from Spontaneous Speech Using a Reduced Set of Acoustic and Linguistic Features,” *Cogn. Comput.*, vol. 5, no. 4, pp. 526–532, Dec. 2013.

[19] 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.

[20] S. R. Livingstone, D. H. Choi, and F. A. Russo, “The influence of vocal training and acting experience on measures of voice quality and emotional genuineness,” *Front. Psychol.*, vol. 5, Mar. 2014.
[21] C. F. Lima, S. L. Castro, and S. K. Scott, “When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing,” *Behav. Res. Methods*, vol. 45, no. 4, pp. 1234–1245, Dec. 2013.
[22] L. Kessous, G. Castellano, and G. Caridakis, “Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis,” *J. Multimodal User Interfaces*, vol. 3, no. 1–2, pp. 33–48, Mar. 2010.
[23] K. R. Scherer, J. Sundberg, L. Tamarit, and G. L. Salomão, “Comparing the acoustic expression of emotion in the speaking and the singing voice,” *Comput. Speech Lang.*, vol. 29, no. 1, pp. 218–235, Jan. 2015.
[24] S. Yildirim, S. Narayanan, and A. Potamianos, “Detecting emotional state of a child in a conversational computer game,” *Comput. Speech Lang.*, vol. 25, no. 1, pp. 29–44, Jan. 2011.
[25] P. Laukka and H. A. Elfenbein, “Emotion Appraisal Dimensions can be Inferred From Vocal Expressions,” *Soc. Psychol. Personal. Sci.*, vol. 3, no. 5, pp. 529–536, Sep. 2012.
[26] D. P. Szameitat, K. Alter, A. J. Szameitat, D. Wildgruber, A. Sterr, and C. J. Darwin, “Acoustic profiles of distinct emotional expressions in laughter,” *J. Acoust. Soc. Am.*, vol. 126, no. 1, p. 354, 2009.
[27] N. Kamaruddin, A. Wahab, and C. Quek, “Cultural dependency analysis for understanding speech emotion,” *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5115–5133, Apr. 2012.
[28] M. Goudbeek and K. Scherer, “Beyond arousal: Valence and potency/control cues in the vocal expression of emotion,” *J. Acoust. Soc. Am.*, vol. 128, no. 3, p. 1322, 2010.
[29] P. Laukka, P. Juslin, and R. Bresin, “A dimensional approach to vocal expression of emotion,” *Cogn. Emot.*, vol. 19, no. 5, pp. 633–653, Aug. 2005.
[30] S. Paulmann, M. D. Pell, and S. A. Kotz, “How aging affects the recognition of emotional speech,” *Brain Lang.*, vol. 104, no. 3, pp. 262–269, Mar. 2008.
[31] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, “Fear-type emotion recognition for future audio-based surveillance systems,” *Speech Commun.*, vol. 50, no. 6, pp. 487–503, Jun. 2008.
[32] M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, “Factors in the recognition of vocally expressed emotions: A comparison of four languages,” *J. Phon.*, vol. 37, no. 4, pp. 417–435, Oct. 2009.
[33] 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.
[34] S. S. Agrawal, N. Prakash, and A. Jain, “Transformation of emotion based on acoustic features of intonation patterns for Hindi speech,” *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 10, no. 9, pp. 198–205, 2010.
[35] R. López-Cózar, J. Silovsky, and M. Kroul, “Enhancement of emotion detection in spoken dialogue systems by combining several information sources,” *Speech Commun.*, vol. 53, no. 9–10, pp. 1210–1228, Nov. 2011.
[36] 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.
[37] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, and K. Elenius, “Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation,” *Comput. Speech Lang.*, vol. 25, no. 1, pp. 84–104, Jan. 2011.
[38] T. Polzehl, A. Schmitt, F. Metze, and M. Wagner, “Anger recognition in speech using acoustic and linguistic cues,” *Speech Commun.*, vol. 53, no. 9–10, pp. 1198–1209, Nov. 2011.
[39] M. D. Pell and S. A. Kotz, “On the Time Course of Vocal Emotion Recognition,” *PLoS ONE*, vol. 6, no. 11, p. e27256, Nov. 2011.
[40] S. Mariooryad and C. Busso, “Exploring Cross-Modality Affective Reactions for Audiovisual Emotion Recognition,” *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 183–196, Apr. 2013.
[41] E. Coutinho and N. Dibben, “Psychoacoustic cues to emotion in speech prosody and music,” *Cogn. Emot.*, vol. 27, no. 4, pp. 658–684, Jun. 2013.
[42] C. Oflazoglu and S. Yildirim, “Recognizing emotion from Turkish speech using acoustic features,” *EURASIP J. Audio Speech Music Process.*, vol. 2013, no. 1, pp. 1–11, 2013.
[43] T. Bänziger, S. Patel, and K. R. Scherer, “The Role of Perceived Voice and Speech Characteristics in Vocal Emotion Communication,” *J. Nonverbal Behav.*, vol. 38, no. 1, pp. 31–52, Mar. 2014.
[44] M. A. Guzman *et al.*, “Influence of Emotional Expression, Loudness, and Gender on the Acoustic Parameters of Vibrato in Classical Singers,” *J. Voice*, vol. 26, no. 5, p. 675.e5-675.e11, Sep. 2012.

[45] K. Hammerschmidt and U. Jürgens, "Acoustical Correlates of Affective Prosody," *J. Voice*, vol. 21, no. 5, pp. 531–540, Sep. 2007.

[46] D. I. Leitman, P. Laukka, P. N. Juslin, E. Saccente, P. Butler, and D. C. Javitt, "Getting the Cue: Sensory Contributions to Auditory Emotion Recognition Impairments in Schizophrenia," *Schizophr. Bull.*, vol. 36, no. 3, pp. 545–556, May 2010.

[47] M. E. Curtis and J. J. Bharucha, "The minor third communicates sadness in speech, mirroring its use in music.," *Emotion*, vol. 10, no. 3, pp. 335–348, 2010.

[48] Jianhua Tao, Yongguo Kang, and Aijun Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006.

[49] H. Pihan, M. Tabert, S. Assuras, and J. Borod, "Unattended emotional intonations modulate linguistic prosody processing," *Brain Lang.*, vol. 105, no. 2, pp. 141–147, May 2008.

[50] K. R. Scherer, "Vocal markers of emotion: Comparing induction and acting elicitation," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 40–58, Jan. 2013.

[51] L. Anolli, Lei Wang, F. Mantovani, and A. De Toni, "The Voice of Emotion in Chinese and Italian Young Adults," *J. Cross-Cult. Psychol.*, vol. 39, no. 5, pp. 565–598, Sep. 2008.

[52] J. Sundberg, S. Patel, E. Bjorkner, and K. R. Scherer, "Interdependencies among Voice Source Parameters in Emotional Speech," *IEEE Trans. Affect. Comput.*, vol. 2, no. 3, pp. 162–174, Jul. 2011.

[53] D. Gharavian, M. Sheikhan, A. Nazerieh, and S. Garoucy, "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network," *Neural Comput. Appl.*, vol. 21, no. 8, pp. 2115–2126, 2011.

[54] D. Gharavian, M. Sheikhan, and F. Ashoftedel, "Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model," *Neural Comput. Appl.*, vol. 22, no. 6, pp. 1181–1191, May 2013.

[55] E. S. Dmitrieva, V. Y. Gel'man, K. A. Zaitseva, and A. M. Orlov, "Perception of the Emotional Intonation of Short Pseudowords," *Neurosci. Behav. Physiol.*, vol. 43, no. 6, pp. 663–669, 2013.

[56] C. Busso and S. S. Narayanan, "Interrelation Between Speech and Facial Gestures in Emotional Utterances: A Single Subject Study," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2331–2347, Nov. 2007.

[57] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Inf. Process. Manag.*, vol. 45, no. 3, pp. 315–328, May 2009.

[58] G. M. Diamond, D. Rochman, and O. Amir, "Arousing primary vulnerable emotions in the context of unresolved anger: 'Speaking about' versus 'speaking to'.," *J. Couns. Psychol.*, vol. 57, no. 4, pp. 402–410, 2010.

[59] D. Rochman, G. M. Diamond, and O. Amir, "Unresolved anger and sadness: Identifying vocal acoustical correlates.," *J. Couns. Psychol.*, vol. 55, no. 4, pp. 505–517, 2008.

[60] D. Erickson, K. Yoshida, C. Menezes, A. Fujino, T. Mochida, and Y. Shibuya, "Exploratory Study of Some Acoustic and Articulatory Characteristics of Sad Speech," *Phonetica*, vol. 63, no. 1, pp. 1–25, 2006.

[61] S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *Speech Audio Process. IEEE Trans. On*, vol. 10, no. 2, pp. 65–78, 2002.

[62] L. Chen, X. Mao, P. Wei, Y. Xue, and M. Ishizuka, "Mandarin emotion recognition combining acoustic and emotional point information," *Appl. Intell.*, vol. 37, no. 4, pp. 602–612, Dec. 2012.

[63] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*, 2008, pp. 865–868.

[64] E. Douglas-Cowie *et al.*, "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *Affective computing and intelligent interaction*, Springer, 2007, pp. 488–500.

[65] T. Bänziger and K. R. Scherer, "Introducing the geneva multimodal emotion portrayal (gemep) corpus," *Bluepr. Affect. Comput. Sourceb.*, pp. 271–294, 2010.

[66] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. University of Erlangen-Nuremberg Germany, 2009.

[67] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech.," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[68] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[69] J. H. Hansen, S. E. Bou-Ghazale, R. Sarikaya, and B. Pellom, "Getting started with SUSAS: a speech under simulated and actual stress database.," in *Eurospeech*, 1997, vol. 97, pp. 1743–46.

[70] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, 2010, pp. 1079–1084.

[71] I. S. Engberg and A. V. Hansen, "Documentation of the danish emotional speech database des," *Intern. AAU Rep. Cent. Pers. Kommun. Den.*, p. 22, 1996.

[72] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, 2006, pp. 8–8.

[73] L. Fu, X. Mao, and L. Chen, "Speaker independent emotion recognition based on svm/hmms fusion system," in *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*, 2008, pp. 61–65.

[74] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011.

[75] B. Zupan, D. Neumann, D. R. Babbage, and B. Willer, "The importance of vocal affect to bimodal processing of emotion: implications for individuals with traumatic brain injury," *J. Commun. Disord.*, vol. 42, no. 1, pp. 1–17, 2009.

[76] N. Rezaei and A. Salehi, "An Introduction to Speech Sciences (Acoustic Analysis of Speech)," *Iran. Rehabil. J.*, vol. 4, no. 4, pp. 5–14, 2006.

[77] S. V. Vaseghi, *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

[78] T. Virtanen, R. Singh, and B. Raj, *Techniques for noise robustness in automatic speech recognition*. John Wiley & Sons, 2012.

[79] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," *Speech Prosody Dresd.*, pp. 276–289, 2006.

[80] S. Patel, K. R. Scherer, E. Björkner, and J. Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biol. Psychol.*, vol. 87, no. 1, pp. 93–98, Apr. 2011.

[81] H. Pérez-Espinosa, C. A. Reyes-García, and L. Villaseñor-Pineda, "Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model," *Biomed. Signal Process. Control*, vol. 7, no. 1, pp. 79–87, Jan. 2012.

[82] R. B. Lanjewar and D. S. Chaudhari, "COMPARATIVE ANALYSIS OF SPEECH EMOTION RECOGNITION SYSTEM USING DIFFERENT CLASSIFIERS ON BERLIN EMOTIONAL SPEECH DATABASE," 2013.

[83] A. Origlia, F. Cutugno, and V. Galatà, "Continuous emotion recognition with phonetic syllables," *Speech Commun.*, vol. 57, pp. 155–169, Feb. 2014.

[84] M. Bejani, D. Gharavian, and N. M. Charkari, "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks," *Neural Comput. Appl.*, vol. 24, no. 2, pp. 399–412, Feb. 2014.

[85] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program], Version 5.1. 44*. 2010.

[86] "Computerized Speech Lab, Kay Elemetrics," *Kay Elemetrics*. [Online]. Available: http://www.kaypentax.com/index.php?option=com_product&controller=product&Itemid=3&cid%5B%5D=11&task=pro_details. [Accessed: 02-Apr-2015].

[87] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, 2009, pp. 1–6.

[88] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*, 2010, pp. 1459–1462.

[89] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative Feature Normalization Scheme for Automatic Emotion Detection from Speech," *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 386–397, Oct. 2013.

[90] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Commun.*, vol. 53, no. 9–10, pp. 1162–1171, Nov. 2011.

[91] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Commun.*, vol. 57, pp. 1–12, Feb. 2014.

[92] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digit. Signal Process.*, vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[93] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *Q. J. Exp. Psychol.*, vol. 63, no. 11, pp. 2251–2272, Nov. 2010.

[94] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Commun.*, vol. 49, no. 10–11, pp. 787–800, Oct. 2007.

[95] S. Ntalampiras and N. Fakotakis, "Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 116–125, Jan. 2012.

[96] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, no. 9–10, pp. 1062–1087, Nov. 2011.

[97] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[98] X. Zhou and K. Z. Mao, "LS bound based gene selection for DNA microarray data," *Bioinformatics*, vol. 21, no. 8, pp. 1559–1564, 2005.

[99] H. Altun and G. Polat, "Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8197–8203, May 2009.

[100] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *NIPS*, 2000, vol. 12, pp. 668–674.

[101] R. O. Duda and P. E. Hart, *Pattern Classification and Science Analysis*. New York: Wiley-Interscience, 1973.

[102] T. H. Dat and C. Guan, "Feature selection based on fisher ratio and mutual information analyses for robust brain computer interface," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 1, p. I–337.

[103] L. Tawade and H. Warpe, "Detection of epilepsy disorder using discrete wavelet transforms using MATLABs," *Int. J. Adv. Sci. Technol.*, vol. 28, pp. 17–24, 2011.

[104] F. Mintzer, "Filters for distortion-free two-band multirate filter banks," *Acoust. Speech Signal Process. IEEE Trans. On*, vol. 33, no. 3, pp. 626–630, 1985.

[105] W. Tarng, Y.-Y. Chen, C.-L. Li, K.-R. Hsie, and M. Chen, "Applications of support vector machines on smart phone systems for emotional speech recognition," *World Acad. Sci. Eng. Technol.*, vol. 72, pp. 106–113, 2010.

[106] Y. Huang, G. Zhang, Y. Li, and A. Wu, "Improved Emotion Recognition with Novel Task-Oriented Wavelet Packet Features," in *Intelligent Computing Theory*, Springer, 2014, pp. 706–714.

[107] M. Wollmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening," *IEEE J. Sel. Top. Signal Process.*, vol. 4, no. 5, pp. 867–881, Oct. 2010.

[108] G. Ilie and W. F. Thompson, "A comparison of acoustic cues in music and speech for three dimensions of affect," 2005.

[109] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Am.*, vol. 93, no. 2, pp. 1097–1108, 1993.

[110] C. Drioli, G. Tisato, P. Cosi, and F. Tesser, "Emotions and voice quality: experiments with sinusoidal modeling," in *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003.

[111] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *J. Multimodal User Interfaces*, vol. 3, no. 1–2, pp. 7–19, Mar. 2010.

[112] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput.*, vol. 31, no. 2, pp. 153–163, Feb. 2013.

[113] M. C. Sezgin, B. Gunsel, and G. K. Kurt, "Perceptual audio features for emotion detection," *EURASIP J. Audio Speech Music Process.*, vol. 2012, no. 1, pp. 1–21, 2012.

[114] Y.-W. Roh, D.-J. Kim, W.-S. Lee, and K.-S. Hong, "Novel acoustic features for speech emotion recognition," *Sci. China Ser. E Technol. Sci.*, vol. 52, no. 7, pp. 1838–1848, Jul. 2009.

[115] M. Song, M. You, N. Li, and C. Chen, "A robust multimodal approach for emotion recognition," *Neurocomputing*, vol. 71, no. 10–12, pp. 1913–1920, Jun. 2008.

[116] R. San-Segundo *et al.*, “Speech technology at home: enhanced interfaces for people with disabilities,” *Intell. Autom. Soft Comput.*, vol. 15, no. 4, pp. 647–666, 2009.

[117] C. S. Ooi, K. P. Seng, L.-M. Ang, and L. W. Chew, “A new approach of audio emotion recognition,” *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5858–5869, Oct. 2014.