International Journal of Speech Technology (2018) 21:93–120
https://doi.org/10.1007/s10772-018-9491-z

Databases, features and classifiers for speech emotion recognition: a review

Monorama Swain[1] · **Aurobinda Routray**[2] · **P. Kabisatpathy**[3]

Received: 13 April 2017 / Accepted: 11 January 2018 / Published online: 19 January 2018

Abstract
Speech is an effective medium to express emotions and attitude through language. Finding the emotional content from a speech signal and identify the emotions from the speech utterances is an important task for the researchers. Speech emotion recognition has considered as an important research area over the last decade. Many researchers have been attracted due to the automated analysis of human affective behaviour. Therefore a number of systems, algorithms, and classifiers have been developed and outlined for the identification of emotional content of a speech from a person's speech. In this study, available literature on various databases, different features and classifiers have been taken in to consideration for speech emotion recognition from assorted languages.

Keywords Speech corpus · Excitation features · Spectral features · Prosodic features · Classifiers · Emotion recognition

1 Introduction

Expressions of emotions are carried out via various means like responses, language, behavior, body gestures, posture and movement. Many physiological processes like respiration, heart rate, temperature, skin conductivity and temperature, muscle and neural membrane potentials can also be used for expression of emotions. Use of these non-invasive parameters are effective in understanding the emotional state of the speaker without having any physical contact, and thereby can also be effectively used to interpret the status of emotions of the individual. The inherent difficulties of deciphering the speakers emotional state from voice arise due to factors such as 'what is said' (factor 1), 'how it is said' (factor 2) and 'who says it' (factor 3). In the first factor, the speech represents the information of linguistic origin and depends on the way of pronunciation of the words as representatives of the language, whereas, the second factor carries paralinguistic information related to the speakers emotional state. The third factor contains cumulative information regarding the speaker's basic features and identities like age, gender, and body size (Borden et al. 1994). That is why it is very difficult to draw any affective information and in specific emotion from voice of an individual.

Affective computing has emerged as an active and interdisciplinary field of research in the area of automatic recognition, interpretation and compilation of human emotions (Picard 1997). The present work falls in a multidisciplinary domain i.e. "Emotion Recognition", involving psychology, social science, linguistics, neurology, neurophysiology, neuropsychophysiology, anthropology, cognitive science, processing of digital signals, speech signal, natural language, and processing of artificial intelligence.

According to Salovey et al. (2004), emotional intelligence has four branches—perception of emotion, facilitation by thought, understanding and management of emotions. The speech signal has become the latest and the fastest communication system between humans involving complex signal processing systems, networking and multiple signaling units about message, speaker, and language. Extensive research has been accomplished in the last few decades related

✉ Monorama Swain
mswain@silicon.ac.in

Aurobinda Routray
aroutray@ee.iitkgp.ernet.in

P. Kabisatpathy
pkabisatpathy@gmail.com

[1] Department of Electronics and Communication Engineering, Silicon Institute of Technology, Bhubaneswar, Odisha, India

[2] Electrical Engineering, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India

[3] Department of Electronics and Communication, CV Raman College of Engineering, Bhubaneswar, Odisha, India

to conversion of human speech to sequence of words. In spite of all these, there is a huge gap between the man and machine, as a machine cannot understand the emotional state of the speaker and thereby fails to interpret the emotions of an individual. This has opened up a new research field called speech emotion recognition, having basic goals to understand and retrieve desired emotions from the speech signal. Researchers are working to make speech as the most efficient method of interaction between humans and machines but the major obstacle is that machines do not have sufficient intelligence to recognize human voices.

Speech emotion recognition has a number of applications—(1) man machine interactions on a natural basis like web movies, (2) computer movies and tutorial applications, (3) drivers safety via car on—board system in which coded message of the driver's mental state is conveyed to the operating system of car, (4) diagnostic tool for a therapist to treat disease (5) used as a tool for automatic translation system in which a speaker plays a key role in communication between parties and (6) mobile communications (Ayadi et al. 2011).

Different theories exist regarding emotion, like evolutionary theories, the James–Lange theory, the Cannon–Bard theory, Schachter and Singer's two-factor theory, and cognitive appraisal. Charles Darwin proposed that emotions evolved due to adaptive nature. Most researchers are of the view that a concept of emotion consists of an event, a perception or an interpretation, an appraisal, physiological change, an action potential, and conscious awareness. Schachter and Singer (1962) have developed a few theories on physiological arousal and the cognitive interpretation. Also psychologist Lazarus' (1991) research has shown that people's experience of emotion depends on the way they appraise or evaluate the events around them. The two emotion structure theories which strongly influenced subsequent research on vocal emotion are discrete category and Dimensional structure theories. This theory tries to categorize human emotions according to where they lie (in 1 or 2-dimensions). These models based on dimensions tend to theorize that affective states result from neuro-phyiological systems. In recent years, the two dimensional model of emotion has gained support among emotion researchers (Gomez and Danuser 2004; Schubert 1999); this is primarily due to the fact that the two dimensional circumplex model was utilized instead of an independent system (Russell 1980). Russell and Barrett (1999) proposed that all affective states arise from two independent neuro-physiological systems, namely valence and activation. Valance describes a pleasure and displeasure continuum while activation (Russell 1980) describes the activeness of the subject in the state of the emotion. But the two dimensional space does not give sufficient information regarding emotion structure so Wundt (2013) suggested emotions could be described by their position in a three dimensional space formed by the dimensions of valance (positive–negative), arousal (calm-excited), and tension (tense-relaxed). Russell and Mehrabian (1977) describe the subjects dominance or power or control over the situation which leads to her/his state of emotion. In this three dimensional space the primary emotion dimensions has been conceptualized in terms of pairs of opposites. Anger and fear are opposites in the sense that one implies attack and the other tends to fight. In the same way joy and sadness are opposites in the sense that one implies a possession or gain while the other implies loss. Acceptance and disgust are opposites in the sense that one implies a taking in and the other implies an ejection or riddance. Surprise and anticipation are opposites in the sense that one implies the unpredictable and other implies the predictable. A four-dimensional emotion space can be described by four mutually orthogonal dimensions valence, activation, potency and intensity (Fontaine et al. 2007). The evolutionary view which is closely related to the discrete emotion theory is believed to have its own particular pattern of cognitive appraisal, physiological activity, action tendency and expression (Darwin 1872/1965; Ekman 1999; Izard 1992; Tomkins 1962). A limited number of basic emotions that have evolved with pertinent life problems such as anger, fear, happiness and sadness are dealt with using discrete emotion theories (Power and Dalgleish 2000). In this paper we have presented a overall review of speech emotion recognition systems keeping in mind the three aspects—the database design, the important speech features and the classifiers used in speech emotion recognition system. Though there are already several reviews available, our literature survey gives insight regarding the features, databases and classifiers used from 2000 to 2017. The article is arranged as follows. Segment 2 gives review of databases used, segment 3 gives the features used, and the different classification techniques available in the literature are presented in Sect. 3. Finally conclusions are drawn in Sect. 4.

2 Speech corpus (emotional): a review

A formidable task for researchers is speech emotion recognition in the area of speech processing. To evaluate the performance of a speech recognition system it is essential to design a suitable database (Ayadi et al. 2011). There are four criteria necessary in preparing a database—the scope, the physical existence, contents and the actual language chosen. The scope of a database design consists of several kinds of variation in a database like number of speakers, speaker gender, type of emotions, number of dialects, type of language and age. While creating the database consideration of additional signals like speech, language, physiological signals recorded simultaneously with speech, the data collection purpose (emotional speech recognition, expressive

synthesis), the emotional states recorded and the kind of the emotions (natural, simulated, elicited) are also taken into consideration for better performance. According to Eckman (1992), the basic emotions are categorized as anger, fear, sadness, sensory pleasure, amusement, satisfaction, contentment, excitement, disgust, contempt, pride, shame, guilt, embarrassment and relief. Non-basic emotions are called "higher-level" emotions (Buck 1999), and they are rarely represented in data collections. The methods and objectives of collecting speech corpora depends according to the motivation behind the development of speech systems.

According to researchers the facial emotions are universal. But vocal signatures show variations in features like the emotion anger. Anger typically has high fundamental frequency when compared with sadness which has a low value of fundamental frequency. These variations occur due to the distinction between real and simulated data. Also some variations can arise due to the differences in culture, language, gender and particular situations. Again, from the literatures it is evident that the emotional scope of databases needs to be designed and developed carefully. Generally, representation of a natural speech from a male voice or female voice for production of a database is a difficult task. So some of the databases consists of some video and audio clips from television, radio programs or from call centers and the natural speech recorded are called spontaneous speech in real world situations. Here we have discussed some of the databases used for speech emotion recognition system. There are three kinds of databases which are used for development of speech emotion recognition systems—natural, simulated (Acted) and elicited (Induced) emotional speech database (Ververidis and Kotropoulos 2006). Natural database is a database developed on spontaneous speech of real data. It includes the data recorded from call center conversations, cockpit recordings during abnormal conditions, conversation between a patient and a doctor, conversation with emotions in public places and similar interactions Batliner et al. (2000). Simulated emotion speech database is known as acted database because the speech utterances were collected from experienced, trained and professional artists. Emotions used for simulated database are full blown emotions (Ayadi et al. 2011). Elicited speech emotion database is where emotions were induced i.e. artificial emotional situation without any knowledge about the speaker. The basic emotions used in the literature are: anger, fear, sadness, sensory pleasure, amusement, satisfaction, contentment, excitement, disgust, contempt, pride, shame, guilt, embarrassment, and relief. Also in addition some physiological signals such as heart rate, blood pressure, respiration and EGG (ELECTROGLOTTOGRAM) could be recorded during experiments (Pravena and Govind 2017). The speech emotion recognition system is also possible from biomultimodal bio-potential signals as in Takahashi (2004). The evaluation done on five emotions joy, anger, sadness, fear and relax. The recorded data are useful for repeating or augmenting the experiments. Speech from interviews with specialists such as psychologists and scientist specialized in phonetics (Douglas-Cowie et al. 2003). A conversation of parents with their children when their intension is to keep them away from dangerous objects can be taken as a real life example (Slaney and McRoberts 2003). Analysis of interviews between a doctor and a patient before and after medication was used in France et al. (2000). Speech was recorded during machine human interaction e.g. during telephone calls to automatic speech recognition (ASR) call centers as discussed in Lee and Narayanan (2005). Speech Under Simulated and Actual Stress (SUSAS) (You et al. 1997) was created by the Robust Speech Processing Laboratory at the University of Colorado-Boulder under the direction of Professor John H. L. Hansen. The simulated database in English language is partitioned into four domains, encompassing a wide variety of stresses and emotions. A total of 32 speakers (13 female, 19 male), with ages ranging from 22 to 76 years were employed to generate 16,000 utterances. SUSAS also contains several longer speech files from four Apache helicopter pilots. Those helicopter speech files were transcribed by the Linguistic Data Consortium and are available in SUSAS Transcripts. LDC Emotional Prosody Speech and Transcripts was developed by the Linguistic Data Consortium and contains audio recordings and corresponding transcripts. It is now part of a commercially available database on English language and contains seven professional actors with 15 emotions and 10 utterances per emotion. The emotions considered are panic, anxiety, hot anger, cold anger, despire, sadness, elation, joy, interest, boredom, shame, pride, contempt and neutral (University of Pennsylvania Linguistic Data Consortium 2002).

The AIBO database is a natural database which consists of recording from children while interacting with robot. The database contain 110 dialogues and 29,200 words in 11 emotion categories of anger, boredom, emphatic, helpless, ironic, joyful, reprimanding, rest, surprise and touchy. The data labeling is based on the listeners' judgment (Batliner et al. 2004). The Berlin Database of Emotional Speech is a German acted database, which consists of recordings from 10 actors (5 male, 5 female). The data consist of 10 German sentences recorded in anger, boredom, disgust, fear, happiness, sadness and neutral. The final database consists of 493 utterances after the listeners' judgment (Burkhardt et al. 2005). The Danish Emotional Speech Database is another audio database recorded from 4 actors (2 male, 2 female). The recorded data consist of 2 words, 9 sentences and 2 passages, resulting in 10 min of audio data. The recorded emotions are anger, happiness, sadness, surprise and neutral (Engberg and Hansen 1996). The RUSLANA simulated speech emotional database was used for linguistic and speech processing research on communicative and emotive

attitudinal aspects of spoken language. A total of sixty-one native speakers (12 males and 49 females) of standard Russian were recorded for this database and the emotions were considered anger, happiness, neutral, sadness, surprise and fear (Makarova and Petrushin 2002).

Zhang et al. (2004) presents an approach to automatically recognize emotion which children exhibit in an intelligent tutoring system. Emotion recognition can assist the computer agent to adapt its tutorial strategies to improve the efficiency of knowledge transmission. In this study, we detect three emotional classes: confidence, puzzle, and hesitation.

Some of the authors focus on the 'Interviews corpus' also known as the Belfast database (Douglas-Cowie et al. 2003); the 'EmoTV' corpus—a set of TV interviews in French recorded in the HUMAINE project (Abrilian et al. 2006). Further speech from real life spoken dialogues from call center services can used for emotion recognition (Vidrascu and Devillers 2005). This corpus of naturally-occurring dialogs recorded in a real-life call center. The corpus contains real agent-client recordings obtained from a convention between a medical emergency call center and the LIMSI-CNRS. The transcribed corpus contains about 20 h of data. Around 404 agent caller dialogs were involved (6 different agents and 404 callers).

The MASC (Mandarin Affective Speech Corpus) (Wu et al. 2006) contains recordings of 68 native speakers (23 female and 45 male) and five kinds of emotions: neutral, anger, elation, panic and sadness. Each speaker pronounces 5 phrases, 10 sentences for three times for each emotional state and 2 paragraphs for neutral. This database can also be used for recognition of affectively stressed speakers and prosodic feature analysis; speaker recognition baseline experiments are also performed on this database. The interactive emotional dyadic motion capture database (IEMOCAP) (Busso et al. 2008), which was collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). This database was recorded from ten actors in dyadic sessions with markers on the face, head, and hands, which provide detailed information about their facial expression and hand movements during scripted and spontaneous spoken communication scenarios. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions such as happiness, anger, sadness, frustration and neutral state. The corpus contains approximately 12 h of data. The Surrey Audio-Visual Expressed Emotion (SAVEE) database (Haq and Jackson 2009) has been recorded as a pre-requisite for the development of an automatic emotion recognition system. The database consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. The sentences were chosen from the standard TIMIT corpus and phonetically-balanced for each emotion and the performance analysis was done by considering 10 subjects under audio, visual and audio-visual conditions. The SEMAINE database contains the recording conversations between humans (the user) and artificially intelligent agents (the operators). The emotion labels for four emotion dimensions: activation, expectation, power, and valence (McKeown et al. 2007).

Koolagudi et al. (2009) proposed one simulated database named IITKGP-SESC which was recorded in the Telugu language with help of some professional artists from AIR Vijayawada, India. The database contains 10 professional artists (5 male and 5 female). For analyzing the emotions they had considered 15 Telugu sentences. Each of the artists had to speak the 15 sentences in 8 basic emotions in one session. The number of sessions considered for preparing the database was 10. The total number of utterances recorded in the database was 12,000 (15 sentences × 8 emotions × 10 artists × 10 sessions). Each emotion has 1500 utterances. Eight basic emotions considered for collecting the proposed speech database were: Anger, Compassion, Disgust, Fear, Happy, Neutral, Sarcastic and Surprise. An IITKGP-SEHSC (Rao and Koolagudi 2011) Hindi dialect speech corpus was collected from five different geographical regions (central, eastern, western, northern and southern) of India, representing the five dialects of Hindi. For each dialect, speech was taken from five male and five female speakers. Speech data was collected from the speaker, by posing the questions arbitrarily so as to describe one's childhood, about the history of the home town, and past memories. Altogether, the duration of each dialect was, about 1–1.5 h. This speech database contains 10 (5 male and 5 female) professional artists for recording purpose collected from All India Radio (AIR) Varanasi, India. The eight emotions considered for recording this database are anger, disgust, fear, happy, neutral, sadness, sarcastic and surprise. Fifteen emotionally neutral, Hindi sentences are chosen as text prompts for the database. Each of the artists had to speak 15 sentences in 8 given emotions in one session. The number of sessions recorded for preparing the database was 10 and each emotion had 1500 utterances. The total number of utterances recorded in this database was 12,000 (15 sentences × 8 emotions × 10 artists × 10 sessions). An IIIT-H Semi-natural Telugu database was used for speech emotion recognition (Gangamohan et al. 2013). The database was recorded in Telugu language and it contained 7 numbers of students from IIIT Hyderabad for emotions such as anger, happy, neutral and sad. EMOVO simulated database was the first database of emotional speech for the Italian language used for speech emotion recognition system. This database contained six actors who were summoned (three males and three females) with proven expertise, and will utter fourteen sentences (assertive, interrogative, lists) based on six basic emotional states (disgust, fear, anger, joy, surprise,

sadness) plus the neutral state (Costantini et al. 2014). In Agrawal (2011) studies conducted to analyze, perceive and recognize commonly occurring emotions in Hindi speech. Also Experiments has been conducted to study and recognise emotions: anger, happiness, fear, sadness, surprise and neutral based on phonetic as well as prosodic parameters in the speech samples due to changes in emotions. The work presented by Pravena and Govind (2017) shows the development of a simulated emotion database for excitation source analysis. The study involves development of a large simulated emotion database for three emotions anger, happy and sad along with neutrally spoken utterances in three languages: Tamil, Malayalam and Indian English. Here also some other databases available in literature also listed in Table 1.

3 Speech features and classifiers

In the designing of a system that recognizes emotions from speech, the identification and extraction of different emotion related speech features is a challenging task. In the real scenario humans have the ability to interpret and detect linguistic and paralinguistic information. Proper selection of speech features affects the classification performance. There are several features like local, global, continuous speech, qualitative, spectral and Teager Energy Operator (TEO)—based, excitation source features, vocal tract features in the pattern recognition problem reviewed in the literature by Koolagudi and Rao (2012a). Since speech signal is non-stationary in nature, to make it stationary it is divided into small segments called frames. Here we present the study based on some important speech features such as excitation source features, vocal tract system, prosodic features and different combination of features. Also we have emphasized on classifiers which were used for developing speech emotion recognition systems. In the literature several classifiers are implemented and tested for developing a speech recognition system. To evaluate the performance of the classifiers, the design of databases are of paramount importance. Every speech database has been created on the basis of environmental conditions and language, but sometimes features selected to design a classifier is not robust enough for speech emotion detection. So classifiers are usually trained and tested by using the same data base. In the literature, a variety of classifiers designed and modelled for recognizing emotions from speech like single classifiers, multiple classifiers, hybrid classifiers or ensemble classifiers are available. In the following sections we present the details of the literature survey about excitation source features, spectral features and prosodic features as well as the different classifiers developed for speech emotion recognition systems.

3.1 Excitation source features

The speech features which were derived from excitation source signals are called source features. The vocal tract (VT) characteristics are suppressed to obtain the excitation source signals. The linear prediction (LP) residual contains information about the excitation source and is achieved by prediction of VT information using filter coefficients and their separation is achieved by inverse filter formulation (Makhoul 1975). Glottal volume velocity (GVV) signal is also used to represent an excitation source, and it is derived from the LP residual signal. The sub segmental analysis of speech signal is glottal pulse, pen and closed phases of glottis and strength of excitation. The correlation of excitation source information is obtained from LP residual signal and glottal volume velocity (GVV). LP residual signal contains valid information as a primary excitation to vocal tract, while speech is being produced. The extracted information about pitch from LP residual signal has been successfully used by Atal (1972). LP residual energy has been also used for vowel and speaker recognition by Wakita (1976). A novel approach of comparative analysis between two features glottal waveform based AUSEEG features and speech based AUSEES features was proposed in He et al. (2010). The study involves an English dataset containing 170 adult speakers for recognition of seven emotions—contempt, angry, anxious, dysphoric, pleasant, neutral and happy. They have also considered MFCC parameters for comparison with the proposed features. For performance evaluation GMM and KNN classifiers were used for classification of emotions. It has been observed that the new features AUSEEG representing the spectral energy distribution of the glottal waveform gives better classification rates than the AUSEES features representing the spectral energy distribution of the speech signal. The excitation component of speech can be exploited for speaker recognition studies in Prasanna et al. (2006).

Chauhan et al. (2010) explored the Linear Prediction (LP) residual of speech signal for characterizing basic emotions. The auto associative neural network (AANN) and Gaussian mixture models (GMM) were considered for emotion classification on IITKGP-Simulated Emotion Speech Corpus (IITKGP-SESC) which includes eight emotions anger, compassion, disgust, fear, happy, neutral, sarcastic and surprise. The emotion recognition performance was observed to be about 56%. Epoch-based analysis of speech helps not only to segment the speech signals based on speech production characteristics, but also helps in accurate analysis of speech (Yegnanarayana and Gangashetty 2011). It enables extraction of important acoustic–phonetic features such as glottal vibrations, formants, and instantaneous fundamental frequency, etc. Accurate estimation of epochs helps in characterizing voice quality features. Epoch extraction also

Table 1 Survey databases

Database	Language	Type of database	Size	Purpose and approach	Emotions
1. Engberg and Hansen (1996)	Danish emotional database	Simulated	Four actors, two of each gender (two isolated words, nine sentences and two passages)	Synthesis To evaluate how well the emotional state in emotional speech is identified by humans	Anger, sadness, surprise, neutral, happiness
2. Montero et al. (1999)	Spanish	Simulated	Two sessions conducted one professional actor, (3 passages, 15 sentences of neutral-content text). 2000 phonemes per emotion are considered for analysis	Synthesis Pitch, tempo, and stress are used for synthesis	Happiness, sadness, cold anger and surprise
3. Amir et al. (2000)	Hebrew	Natural	40 subjects were considered (19 males, 21 females)	Physiologic evaluations Signal analysis over sliding windows and extracting a representative feature set	Anger, fear, joy, sadness, disgust
4. Cecile Pereira (2000)	English	Simulated	2 actors (40 utterances)	Recognition Findings of emotions on the three dimensional scales arousal, pleasure and power	Happiness, sadness, anger (hot and cold), neutral
5. Marc Schroder (2000)	German	Simulated	6 native speakers (3 male, 3 female)	Recognition	Admiration, threat, disgust, elation, boredom, relief startle, worry contempt, anger
6. Iriondo et al. (2000)	Castilian Spanish	Simulated	Eight actors (4 male and 4 female), 336 discourses were recorded	Synthesis	Desire, disgust, fear, fury (anger), joy, sadness, surprise
7. Nogueiras et al. (2001)	Spanish	Simulated	Two professional actors (one male and one female)	Synthesis Emotion recognition using RAMES, the UOC's speech recognition system based on standard speech recognition technology using hidden semi-continous Markov models	Anger, disgust, fear, joy, sadness, surprise, neutral
8. New et al. (2001)	Burmese	Simulated or acted	Two Burmese language speakers, 90 emotional utterances each from two speakers	Recognition A universal codebook is constructed based on emotions	Anger, dislike, fear, happiness, sadness and surprise
9. Yu et al. (2001)	Chinese	Simulated	Native TV actors 721 short utterances per emotion are recorded	Recognition	Anger, happiness, neutral, and sad
10. Makarova and Petrushin (2002)	Russian	Simulated	61 Native speakers (12 male, 49 female), 10 sentences were recorded per emotion, total 3660 utterances	Recognition This database is a source for linguistic and speech processing research	Neutral (unemotional), surprise, anger, happiness, sadness and fear

Table 1 (continued)

Database	Language	Type of database	Size	Purpose and approach	Emotions
11. Bulut et al. (2002)	English	Simulated	1 actress	Synthesis	Anger, happiness, sadness, neutral
12. Scherer et al. (2002)	English and German	Natural	100 native speakers	Recognition	Stress and load level
13. Tato et al. (2002)	German	Elicited	14 native speaker	Synthesis	Anger, boredom, happiness, neutral, sad
14. Chuang and Wu (2002)	Chinese	Simulated	2 actors (1 Male and 1 female) From male: 558 utterances contained in 137 dialogues From female: 453 sentences in 136 dialogues	Recognition An emotional semantic network proposed to extract the schematic information related to emotion	Anger, surprise, sadness, fear, happiness and antipathy
15. Yuan et al. (2002)	Chinese	Elicited	9 native speakers	Recognition	Anger, fear, joy, neutral, sadness
16. Hozjan et al. (2002)	English, Slovenian Spanish and French	Simulated	One male and one female speaker have been recorded, for English two male and 1 female have been recorded. English interface database contain 8928 sentences, Slovenian contain 6080 sentences, French contain 5600 sentences and Spanish contain 5520 sentences	Synthesis The recorded INTERFACE database is used to develop a multilingual emotion classifier and for multilingual emotion modeling for speech synthesis	Anger, sadness, joy, fear, disgust, surprise and neutral
17. Rahurkar and Hansen (2002)	English	Ntural	6 soldiers	Recognition	5 stress levels
18. New et al. (2003)	Burmese and Mandarin	Simulated	Tweleve speakers Sixty different utterances ten each for each emotion for each speaker was constructed In Burmese six speakers (3 male and 3 female), for Mandarin language (3 male and 3 female) speakers were employed to generate 720 utterances	Recognition Log frequency power coefficients are used for emotion recognition using HMM classifier	Anger, disgust, fear, joy, sadness, and surprise
19. Schroder and Grice (2003)	German	Simulated			
20. Slaney and McRoberts (2003)	English and German	Natural	12 native speakers (six fathers and six mothers)	Recognition A multidimensional Gaussian mixture-model discriminator classified adult-directed and infant-directed speech using pitch and broad spectral-shapes measures	Approval, attentation and prohibition

Table 1 (continued)

Database	Language	Type of database	Size	Purpose and approach	Emotions
21. Lee and Narayanan (2003)	English	Natural	Unknown	Recognition Call center application	Negative (anger, frustration, boredom) and positive emotions (neutral, happiness, others)
22. Yamagishi et al. (2003)	Japanese	Simulated	I male speaker Phonetically balanced 503 sentences of ATR Japanese database	Speech recognition and synthesis An approach to realizing various emotional expressions and speaking styles in synthetic speech using HMM based synthesis	Joyful and sad
23. Schuller et al. (2003)	English, German	Natural and Simulated	5 speakers Total 5250 samples taken for analysis	Recognition Two different methods propagated for various feature analysis and comparison between two classifies such as GMM and HMM	Anger, disgust, fear, surprise, joy, neutral and sadness
24. Hozjan and Kacic (2003)	English, Slovenian Spanish and French	Simulated	Total 9 speakers for each language Total 23,000 sentences were recorded. For English language two male and one female speakers were recorded and for Slovenian, Spanish, French language one male and one female speaker were recorded	Recognition Analysis of various acoustic and large set of statistical features	Anger, sadness, joy, fear, disgust, surprise and neutral
25. Lida et al. (2003)	Japanese	Simulated	Two native speakers (one male and one female)	Synthesis To synthesizing emotional speech by a corpus based concatenative speech synthesis system (ATR CHATR) using speech corpora of emotional speech	Anger, joy and sadness
26. Fernandez and Picard (2003)	English	Natural	Four drivers	Recognition Use of features derived from multi resolution analysis of speech and TEO for classification of driver's speech under stressed conditions	Stress

Table 1 (continued)

Database	Language	Type of database	Size	Purpose and approach	Emotions
27. Jovičić et al. (2004)	Serbian	Simulated	Six actors (3 female, 3 male) GEES database contains 32 isolated words, 30 short semantically neutral sentences, 30 long semantically neutral sentences and one passage with 79 words in size. Total database contains 2790 recordings and duration of speech around 3 h	Recognition Designing, processing and evaluation of Serbian emotional speech database	Neutral, anger, happiness, sadness and fear
28. Schuller et al. (2004)	German and English	Simulated	German and English sentences of 13 speakers, one female were assembled German database contains 2829 emotional recorded samples for training and evaluation in the prosodic and for linguistic analysis English database contains 700 selected utterances in automotive infotainment speech interaction dialogs recorded for the evaluation of the fusion	Recognition Combination of acoustic features and language information for a most robust automatic recognition of a speakers emotion	Anger, disgust, fear, joy, neutral, sadness, surprise
29. Batliner et al. (2004)	German, English	Elicited	51 children	Recognition Recorded at the university of Maribor, in German and English	Anger, Boredom, joy and surprise
30. Yildirim et al. (2004)	English	Simulated	One actress 112 utterances per emotion are recorded	Recognition Main aim was how speech is modulated when speakers emotion changes to a certain emotional state. Speech prosody, vowel articulation and spectral energy distribution are to analyze 4 emotions	Sadness, anger, happiness and neutral
31. Nordstrand et al. (2004)	Swedish	Simulated Multimodal corpus	One native speaker	Synthesis Variations in articulatory parameters are used for recording of Swedish vowels in two emotions	Happiness and neutral

Table 1 (continued)

Database	Language	Type of database	Size	Purpose and approach	Emotions
32. Caldognetto et al. (2004)	Italian	Simulated	Single native speaker	Synthesis Here analysis on the interaction between the articulatory lip targets of the Italian vowel and consonants defined by phonetic-phonological rules and labil configurations peculiar to each emotion	Anger, disgust, fear, joy, sadness and surprise
33. Jiang and Cai (2004)	Chinese	Simulated	Single amateur actress 200 Chinese utterances for each emotion	Recognition Combination of statistic features and temporal features	Anger, fear, happiness, sadness, surprise and neutral
34. Ververidis et al. (2004)	Danish	Simulated	Four actors (two male and two female) Danish emotional speech database Total amount of data used in the experiment was 500 speech segments (with no silence interruptions)	Recognition Feature analysis and classification	Anger, happiness, sadness, surprise and neutral
35. Jiang et al. (2005)	Mandarin	Natural	One female speaker Total 216 sad sentences, 143 happy sentences and 10 sentences per each emotion	Synthesis Analysis and modeling the emotional prosody features	Sadness and happiness
36. Cichosz and Slot (2005)	Polish	Simulated	Four actors and four actresses Total 240 utterances uttered by four actors and four actresses	Recognition To determine a set of low dimensional feature spaces that provides high recognition rates	Anger, fear, sadness, boredom, joy and neutral (no emotion)
37. Lin and Wei (2005)	Danish	Simulated	Four actors (two male and two female) familiar with radio theatre	Recognition Gender dependent and gender independent speech emotion recognition	Anger, happiness, sadness, surprise and a neutral state
38. Luengo et al. (2005)	Basque	Simulated	One actress Total 97 recordings for each emotion were done Database contains numbers, isolated words and sentences of different length	Emotion identification Analysis of prosodic features and spectral features with classifiers GMM and SVM for emotion identification	Anger, fear, surprise, disgust, joy and sadness
39. Lee and Narayanan (2005)	English	Natural	Customers and call attendants	Recognition Call center conversations are recorded	Negative and positive

Table 1 (continued)

Database	Language	Type of database	Size	Purpose and approach	Emotions
40. Pao et al. (2005)	Mandarin	Simulated	Eighteen male and sixteen female uttered 20 different utterances. Total 3400 sentences were recorded	Recognition Evaluation of Mandarin speech using weighted D-KNN classification	Anger, happiness, sadness, boredom and neutral
41. Batliner et al. (2006)	English	Elicited	51 school children (21 male and 30 Female)	Recognition Children are asked to spontaneously react with Sony AIBO pet robot. Around 9.5 h of effective emotional expressions of children were recorded	Different elicited emotions are recorded
42. Wu et al. (2006)	Chinese	Simulated	Non –broadcasting speakers Total 25 male, 25 female speakers were involved in the recording process	Recognition Study on GMM-UBM based speaker verification system on emotional speech	Anger, fear, happiness, sadness, neutral
43. Grimm et al. (2006)	English	Simulated	EMA (Electromagnetic articulography) Database contains 680 emotional speech utterances, generated by one female professional and two non-professional (one male and one female) speakers. Female speaker produce 10 sentences and male speaker produce 14 sentences each for 4 different emotions	Recognition Feature based categorical classification and primitives-based dynamic emotion estimation	Happy, angry, sad, neutral
44. Morrison et al. (2007)	Mandarin and Burmese	Natural and simulated	1. Natural database contains 11 speakers with 388 numbers of utterances for two emotion classes 2. ESMBS database contains 12 emotional speeches of Mandarin and Burmese speakers with 720 utterances for six emotions. Six Mandarin and six Burmese speakers were used. 10 different sentences uttered by the speakers	Recognition Call center applications	1. Anger, neutral 2. Anger, happiness, sadness, disgust, fear, surprise

Table 1 (continued)

Database	Language	Type of database	Size	Purpose and approach	Emotions
45. Kandali et al. (2008a)	Assamese	Simulated	MESDNEI (multilingual emotional speech database of North East India) database contains short sentences of six full blown basic emotions with neutral Total 140 simulated utterances per speaker was collected for 5 native language of Assam. Specifically students and faculty members from educational institutions were chosen for the recording. 30 subjects (3 male and 3 female per language) were chosen for recording	Recognition Vocal emotion recognition	Anger, disgust, fear, happiness, sadness, surprise and 1 no emotion (neutral)
46. Grimm et al. (2008)	German	Natural	104 Native speakers (44 male and 60 female)	Recognition 12 h of audio visual recording is done using TV talk show Vera am Mittang in German. Emotion annotation is done based on activation, valence, and dominance dimensions	Two emotions for each emotional dimensions are recorded 1. Activation (calm-excited) 2. Valence (positive–negative) 3. Dominance (weak-strong)
47. Koolagudi et al. (2009)	Telugu	Simulated	The database contains 10 professional artists (5 male and 5 female) from All India Radio (AIR) Vijaywada. Total number of utterances recorded in the database was 12,000 (15 sentences, 8 emotions, 10 artists and 10 sessions). Each emotion contains 1500 utterances	Recognition Design, acquisition, post processing and evaluation of IITKGP-SESC database	Anger, disgust, fear, happy, compassion, neutral, sarcastic, surprise
48. Mohanty and Swain (2010)	Oriya language	Elicited	Database contains 35 speakers (Male 23 and Female 12), reading text fragments taken from various Oriya drama scripts	Recognition Creation of Odiya database and emotion recognition from Odiya speech	Anger, sadness, astonish, fear, happiness, neutral

Table 1 (continued)

Database	Language	Type of database	Size	Purpose and approach	Emotions
49. Rao and Koolagudi (2011)	Hindi	Natural and simulated	1. Hindi dialect speech corpus used for dialect identification. It contains 5 females and 5 males, sentences uttered based on their past memories 2. IITKGP-SEHSC corpus used for the speech emotion recognition. It contains 10 professional artists from All India Radio Varanasi, India. Total 12,000 utterances recorded and each emotion has 1500 sentences	Recognition and Identification Dialect identification, emotion recognition and feature analysis	Anger, disgust, fear, happy, neutral, sadness, surprise, sarcastic
50. Koolagudi et al. (2012)	Hindi	Semi natural	Utterances taken for the database were the recording dialogues delivered by Hindi film actor and actress	Recognition Proposed a Semi natural database (Graphic Era University semi natural speech emotion corpus) for emotion recognition	Sad, anger, happy, neutral
51. Caballero-Morales (2013)	Mexican Spanish	Simulated	Total 6 speakers from the local cultural center of the city of Huajuapan de Leon in Oaxaca, Mexico took part in the recording process. The database contained total 40 utterances and 233 words (vocabulary of 140 unique words)	Recognition Acoustic modeling of emotion-specific vowels	Anger, happiness, neutral, sadness
52. Quiros-Ramirez et al. (2014)	Latin-American, Japanese	Natural	Total 57 participants were involved for the recording process, 30 (12 females and 18 males) participants from Latin-American and 27 (10 females and 17 males) from Japan	Recognition Spontaneous cross-cultural emotion database	Negative, positive
53. Esmaileyan and Marvi (2014)	Persian	Simulated	Persian Drama Radio Emotional Corpus (PDREC) contains emotional utterances taken from radio programs. Total 748 utterances were recorded by 33 (15 females and 18 males) native speakers of Persian language	Recognition Design of database for Automatic Persian speech emotion recognition	Anger, boredom, disgust, fear, neutral, sadness, surprise, happiness

Table 1 (continued)

Database	Language	Type of database	Size	Purpose and approach	Emotions
54. Ooi et al. (2014)	German, English, Mandarin, urdu, Punjabi, Persian, and Italian	Simulated	1. EMO-DB (Berlin emotional database) contains 10 speakers (5 male and 5 female), 10 sentences were chosen for recording, total 840 recorded utterances were used. 2. eNTERFACE'05 audio-visual emotion database contains 1170 utterances with 42 subjects (34 male and 8 female) chosen from different nations. 3. RML (audio-visual emotion database) contains 720 videos from 8 subjects.	Recognition A new architecture of intelligent audio emotion recognition is introduced and analysis of different prosodic and spectral features was done.	1. Anger, boredom, disgust, fear, happiness, neutral, sadness 2. Happy, angry, disgust, sad, surprise, fear 3. Anger, disgust, fear, happiness, surprise, sadness, neutral
55. Mencattini et al. (2014)	Italian	Simulated	EMOVO Italian speech corpus. It contains 588 recordings: 14 Italian sentences by 6 professional actors (3 male and 3 female)	Recognition PLS regression model was introduced, new speech features related to speech amplitude modulation parameters was discussed	Disgust, joy, fear, anger, sadness, surprise, neutral
56. Kadiri et al. (2015)	Telugu and German	Semi natural and simulated	1. Students (two females and five males) were involved in the recording process and the utterances recorded based on past memories. Total 200 utterances recorded for experiment (IIIT-H Telugu emotion database) 2. EMO-DB Berlin emotion database contains 10 professional native German actors (5 males and 5 females) were asked to speak 10 sentences in different emotions. Total 535 utterances were recorded. 339 utterances were taken for final experiment	Recognition Excitation source feature analysis for speech emotion recognition	Anger, happy, neutral, sad

Table 1 (continued)

Database	Language	Type of database	Size	Purpose and approach	Emotions
57. Song et al. (2016)	German and English	Simulated	Berlindataset: emotional utterances recorded by ten actors (5 males and 5 females) in German language. Total 494 utterances were used for experiment eNTERFACE (Audio-visual database) 42 speakers were allotted for recording (34 males and 8 females).Total 1170 video samples were collected	Recognition A novel transfer non-negative matrix factorization (TNMF) method is presented for cross-corpus Speech emotion recognition	1. Anger, boredom, disgust, fear, happiness, sadness and neutral 2. Anger, disgust, fear, happiness, sadness and surprise
58. Brester et al. (2016)	German, English, Japanese	Simulated and natural	Four emotional databases 1. EMO-DB (GERMAN database) recorded at the Technical University of Berlin. It consists of labeled emotional German utterances spoken by 10 actors 2. SAVEE (Surrey Audio-Visual Expressed Emotion) corpus in (English). It contains four native English male speakers 3. LEGO emotion database (English). It comprises of non-acted American English utterances extracted from an automated bus information system of the Carnegie Mellon University at Pittsburg USA 4. UUDB (The Utsunomiya University Spoken Dialouge Database for paralinguistic information studies) (Japanese) consists of spontaneous human –human speech	Recognition Evolutionary feature selection technique based on the two criterion optimization model	1. Neutral, anger, fear, joy, sadness, boredom or disgust 2. Anger, disgust, fear, happiness, sadness, surprise, and neutral 3. Angry, slightly angry, very angry, neutral, friendly, and non speech (critical noisy recordings or just silence) 4. Happy-exciting, angry-anxious, sad-bored, relaxed-serene
59. Pravena and Govind (2017)	Tamil, Malayalam and Indian English	Simulated	10 speakers Emotionally biased utterances	Recognition Development of a simulated emotion database for excitation source analysis	Anger, happy, sad

helps in speech enhancement and multispeaker separation. Prasanna and Govind (2010) examined the effect of emotions on the excitation source of speech production. Five emotions neutral, angry, happy, boredom and fear were considered for the study. Initially the electroglottogram (EGG) and its derivative signals are compared across different emotions. The mean, standard deviation and contour of instantaneous pitch, and strength of excitation parameters are derived by processing the derivative of the EGG and also speech using zero-frequency filtering (ZFF) approach. The work presented by Pravena and Govind (2017) explored the effectiveness of the excitation source parameters-strength of excitation and instantaneous fundamental frequency (F0) for emotion recognition from speech and electroglottographic (EGG) signals. They have considered GMM classifier for performance evaluation. Nandi et al. (2017) shows the linear prediction (LP) residual signal has been parameterized to capture the excitation source information for language identification (LID) study. LP residual signal has been processed at three different levels: sub-segmental, segmental and supra-segmental levels to demonstrate different aspects of language-specific excitation source information. The proposed excitation source features have been evaluated on 27 Indian languages from Indian Institute of Technology Kharagpur-Multi Lingual Indian Language Speech Corpus (IITKGP-MLILSC), Oregon Graduate Institute Multi-Language Telephone-based Speech (OGI-MLTS) and National Institute of Standards and Technology Language Recognition Evaluation (NIST LRE) 2011 corpora. LID systems were developed using Gaussian mixture model (GMM) and i-vector based approaches. Experimental results have shown that segmental level parametric features provide better identification accuracy (62%), compared to sub-segmental (40%) and supra-segmental level (34%) features. Here Table 2 represents the excitation source features used in the available literature.

3.2 Prosodic feature

Prosody or supra segmental information structures the flow of speech and consists of duration, intensity, intonation and sound units. The acoustic correlates of prosodic features include pitch, energy, duration and their derivatives (Rao and Yegnanarayana 2006). The four levels of computing prosody are (a) linguistc intention level, (b) articulatory level, (c) acoustic realization level, (d) perceptual level (Werner and Keller 1994). Different linguistic elements are related in an utterance by bringing semantic emphasis on an element. The articulatory movements include physical movements of the muscles in the throat. The fundamental frequency intensity and duration help in analysis of prosody factors according to Rao and Yegnanarayana (2006). It is also expressed in form of pauses, length and melody to subjective listeners. The acoustic properties are therefore used to analyze prosody. Prosodic features pitch such as mean. maximum, minimum, variance and standard deviation of energy are used by Dellert et al. (1996b). They have considered maximum likelihood Bayes classification, kernel regression, and a k-nearest neighborhood methods. The steepness of fundamental frequency, articulation duration, rise and fall durations are also measured. Fear, anger, sadness and joy are expressed as peaks and troughs of F0. Minimum maximum and medium values of F0 are emotion salient features. Emotions are analyzed by using short time supra segmental features like pitch, energy, formant, locations and their bandwidths, dynamics of pitch, energy, formant contour, speaking rate (Ververidis and Kotropoulos 2006).

Nogueiras et al. (2001) investigated the classification of emotions in speech with a Hidden Markov model (HMM). They tested different HMM configurations and found that increasing the number of states from one to 64 monotonically improved the recognition accuracy. The best reported recognition accuracy was 82.5% which was obtained using HMMs with 64 states and all 11 features based on Prosodic features such as energy and pitch. Ramamohan and Dandapat (2006) computed sinusoidal model based features consisting of 10 frequencies (sorted in ascending order) and 10 phase angles corresponding to 10 most significant peaks of amplitudes at which slope of amplitude changes from positive to negative. A trained Vector Quantizer (VQ) based classifier and Hidden Markov Model with discrete state output probability density function (VQ-HMM) were used. They reported a score of success of 80% in case of Telugu and 70% in case of English. Agrawal et al. (2009) collected 5 short sentences of Hindi language uttered by 6 male graduate drama club students aged between 20 and 23 years with 4 repetitions in *Anger, Fear, Happiness, Sadness*, and *Neutral* emotions. They computed pitch using *Praat* software and applied Neural Network (NN) and Fisher's Linear Discriminant Analysis (FLDA) classifiers and achieved average recognition success scores as 64.3% and 56.2% in Hindi language. Koolagudi et al. (2009) created the IITKGP-SESC database which consists of 10 repetitions of 15 portrayed utterances in *Anger, Compassion, Disgust, Fear, Happy, Neutral, Sarcastic* and *Surprise* emotions uttered by 10 professional actors (5 males and 5 females) in the *Telugu* language. They computed mean duration of each utterance, mean and standard deviation of pitch values and mean energy across each utterance of 1 male and 1 female speaker. The achieved average recognition success score was 75 and 69% for male and female speakers respectively, using an Euclidean distance classifier. Lee et al. (2011) computed zero crossing rate, root mean square energy, harmonics-to noise ratio and 12 mel frequency cepstral coefficients and their deltas for emotion analysis. The study involves a hierarchical binary decision tree approach and SVM classifier for analysis of emotions anger,

Table 2 Excitation source features

Author	Feature	Approach and classifier
Wakita (1976)	LP residual energy	Vowel and speaker recognition
Rao et al. (2007)	LP residual	Detection of instants of significant excitation
Yegnanarayana et al. (2009)	LP residual	Speech enhancement in multi-speaker environment
Bapineedu et al. (2009)	LP residual	Characterizing loudness, lombard effect, speaking rate, and laughter segments
Cummings and Clements (1998)	Glottal excitation signal	Analyzing the relation between emotional state of the speaker and glottal activity
Gangamohan et al. (2014)	Excitation source signal	Two features strength of excitation and Spectral band energy ratio related to excitation component of speech were examined for discrimination between anger and happy emotions. Zero frequency filtering method and spectral band magnitude energies from short-time spectral analysis were used
Koolagudi et al. (2010)	Epoch parameters LP residual	Epoch parameters extracted from LP (linear prediction) residual and zero frequency filtered speech signal for recognizing the six emotions anger, disgust, fear, happy, neutral and sadness in speech. Gaussian mixture models and support vector machines were in the experiment and the average emotion recognition observed was of 61% for GMM and 58% for SVM classifier
Krothapalli and Koolagudi (2013)	Excitation source features, sequence of LP residual samples and their phase information, parameters of epochs and their dynamics at syllable and utterance levels, samples of GVV signal and its parameters	Characterizing and recognizing the emotions from the speech signal. Auto-associative neural networks (AANN) and support vector machines (SVM) were used. Anger, disgust, fear, happy, neutral and sadness are the six emotions were considered for the experiment
Kadiri et al. (2015)	Excitation source signal	Speech emotion recognition Hierarchical binary decision Tree was used for classification. Four basic emotions were used anger, happy, neutral and sad
Koolagudi and Rao (2012b)	LP residual	Recognition of emotions from speech. GMM was used for classification
Gangamohan et al. (2013)	Excitation source	KL distance values taken as consideration. The emotion recognition system for the 4-class (anger, happy, neutral and sad) problem gives an accuracy of 76% for IIIT-H Telugu emotion database and 69% for Berlin EMO-DB database
Sun and Moore (2011)	Glottal waveform parameters and TEO	Performance analysis of glottal waveform parameters and TEO in distinguishing binary classes of four emotion dimensions (activation, expectation, power, and valence) using authentic emotional speech SVM classifier taken for performance evaluation and used SEMAINE database for the feature analysis
Pravena and Govind (2017)	Instantaneous F0 and strength of excitation parameters	Development of simulated emotion speech database for excitation source analysis GMM model was considered for classification of three emotions anger, happy and sad on German (EmoDb) and IITKGP-SESC Telugu speech emotion databases

happy, sad, neutral, emphatic. Chen et al. (2012) worked on BHUDES database which contains mandarin utterances of six emotions like sadness (sad), anger (ang), fear (fea), happiness (hap), and disgust (dis). The database consists of 5400 utterances which were performed by 15 speakers (actors and actress) whose ages were between 20 and 25. They conducted experiments on a speaker independent system. They computed some instantaneous features with SVM and ANN. Mirsamadi et al. (2017) utilized the deep learning to automatically discover emotionally relevant features from

speech. The study involves classification of four emotions happy, sad, neutral and angry from IEMOCAP dataset. A speaker independent speech emotion recognition was performed using a deep recurrent neural network classifier. Jin and Wang (2005) studied the distribution of the seven emotions in spoken Chinese, including joy, anger, surprise, fear, disgust, sadness and neutral, in the two dimensional space of valence and arousal, and analyzed the relationship between the dimensional ratings and the prosodic characteristics in terms of F0 maximum, minimum, range and mean. Thirty-two different acoustic features F0, energy, duration, and tune had studied for the classification of five emotion states in McGilloway et al. (2000).

Table 3 represents the some of the prosodic features with classifiers used in the literature for speech emotion recognition.

3.3 Vocal tract feature

Vocal tract features are basically known as spectral features or segmental features. To extract the vocal tract system features a speech segment of length 20–30 ms is generally required. For the analyses of different speech features formants, bandwidth spectral energy and slope of the signal from the spectrum, it is necessary to find the Fourier transform of a speech frame. The Fourier transform on log magnitude spectrum gives the cestrum (Rabiner and Juang 1993). The cepstral domain represents the MFCCs (Mel frequency spectral frequency coefficients), PLPCs (perpetual linear prediction coefficients) and the LPCCs (linear predictions cepstral coefficients) which are also called segmental or system features (Ververidis and Kotropoulos 2006). In the literature spectral features have been successfully used for various speech applications in the area of development of speech recognition and speaker recognition systems. Here we have discussed some of the studies on speech emotion recognition using spectral features.

The utterances of 10 sentences regarding anger, fear, joy, sadness, disgust and surprise emotions in Mandarin and Burnese, were examined by New et al. (2003). The values of LFPC (Log-frequency power coefficient) were computed for each frame and discrete Markov models were developed. Success was at 79.9% for Burmese and 76.4% for Mandarin. 1134 utterances (5 repeated utterances of 7 short sentences (1 sentence per emotion) uttered by each speaker and 1 long sentence per emotion uttered by each speaker from his own thought) of 6 full-blown basic emotions (Anger, Disgust, Fear, Happiness, Sadness, Surprise) and 1 'no-emotion' i.e. Neutral were collected by Kandali et al. (2008a) from 27 native speakers (14 Males and 13 Females) of the Assamese language. This work reports a success score of 76.5% in the text-independent but speaker-dependent experiment using the spectral feature MFCC (Mel-Frequency Cepstral Coefficient) feature set and GMM (Gaussian Mixture Model) classifier. Tests were conducted with varying contents of utterances and speakers, compared with those during training conditions. i.e. the experiments are text-and-speaker-independent Kandali et al. (2008a). Firoz Shah et al. (2009) analyzed an elicited database consisting of 700 utterances for emotions neutral, happy, sad and anger in Malayalam (One of the south Indian languages) using both Discrete Wavelet Transforms (DWTs) and Mel Frequency Cepstral Coefficients (MFCCs). The overall recognition accuracy was observed 68.5 and 55% using Artificial Neural Network classifier for speech emotion recognition. Kandali et al. (2008b) collected 4200 utterances (140 utterances of different short sentences (20 sentence per emotion) uttered by each speaker) of 6 *full-blown* basic emotions (*Anger, Disgust, Fear, Happiness, Sadness, Surprise*) and 1 'no-emotion' i.e. *Neutral* from 30 native speakers (3 Males and 3 Females per language) of 5 native languages of Assam (India). They performed text-and-speaker-independent experiments using utterances of all the languages taken together. They reported a success score of 73.1, 75.1, 60.9, 61.1, 95.1, 93.8, 93.7 and 94.29% using feature sets MFCC, tfMFCC (Teager-Energy-Operated-in-Transform-Domain MFCC), LFPC, tfLFPC, WPCC (Wavelet Packet Cepstral Coefficient), tfWPCC, WPCC2 (Wavelet Packet Cepstral Coefficient computed by method 2) and tfWPCC2 respectively with a GMM classifier. Rao et al. (2012) worked on spectral features for both acted and real databases with a GMM classifier. Ververidis et al. (2004) they have considered 87 features calculated over 500 utterances of the Danish Emotional Speech database. The study involves The Sequential Forward Selection method (SFS) has been used in order to discover the 5–10 features which are able to classify the samples in the best way for each gender. Bayes classifier was considered for classification of emotions: anger, happiness, neutral, sadness and surprise. A correct classification rate of 61.1% is obtained for male subjects and a corresponding rate of 57.1% for female ones. In the same experiment, a random classification would result in a correct classification rate of 20%. When gender information is not considered a correct classification score of 50.6% was obtained. In Lanjewar et al. (2015) Mel frequency cepstrum coefficients (MFCC), wavelet features of speech, pitch of vocal traces were considered for speech emotion recognition. Gaussian mixture model (GMM), k-Nearest Neighbor (k-NN) models considered for classification and recognition of six emotions: happy, angry, neutral, surprised, fearful, and sad in Berlin emotion database (BES). Table 4 represents the some of the vocal tract features with classifiers used in the literature for speech emotion recognition.

Table 3 Literature survey of prosodic features used for speech emotion recognition

References	Feature	Approach and classifier
Banse and Scherer (1996)	Fundamental frequency/pitch (F0), energy, speech rate, and spectral information in voiced and unvoiced portions	Acoustic profiles or vocal cues for emotion expression using actors voices for fourteen emotion categories
Nicholson et al. (2006)	Prosodic features	Recognizing emotions using a large database of phoneme balanced words, speaker and contest-independent. One-class-in-one Neural Networks used for classifications of emotions. Around 50% recognition rate was achieved
Picard et al. (2001)	Statistical features (mean, standard deviation of raw signals, absolute values of first and second differences of raw signals), sequential forward selection search, fischer projection	Electromyogram blood volume pulse skin conductance respiration Hybrid linear discriminant analysis used for classification of emotions neutral, anger hate, grief platonic love, romantic love, joy, reverence
Park and Sim (2003)	Prosodic features	Recognition of four emotions neutral, anger, laugh, surprise and feature analysis DRNN (dynamic recurrent neural network) for classification
Tao and Kang (2005)	Prosodic features intonation, speaking rate, intensity	CART model and a weight decay neural network model considered for the classification performance analysis of emotions
Koolagudi and Krothapalli (2012)	Prosodic features (1) duration patterns, (2) average pitch, (3) variation of pitch with respect to mean [standard deviation (SD)], and (4) average energy	Performance evaluation of eight emotions anger, compassion, disgust, fear, happy, neutral, sarcastic and surprise of IITKGP-SESC: speech database Euclidian distance measure and subjective evaluation considered for the classification performance analysis of emotions
Lee et al. (2001)	Prosodic features fundamental frequency (F0), energy, duration, the first and second formant frequencies	Detection of negative and non-negative emotions using spoken language data obtained from a call center application
Petrushin (1999)	Prosodic features pitch, the first and second formants, energy and the speaking rate	Real-time emotion recognizer using an ensemble of neural networks for a call center application. Classification accuracy was achieved 77% for two emotional states, agitation and calm, with eight features chosen by a feature selection method
Luengo et al. (2005)	Total 86 prosodic features were used, best features were chosen from feature selection	Identification of four different emotions in Basque language.GMM classifier was used for performance analysis. 92% recognition accuracy was achieved
Iliou and Anagnostopoulos (2009)	35 dimensional prosodic feature vectors including pitch, energy, and duration	Classification of seven emotions of Berlin emotional speech corpus. For speaker independent cases using neural networks the emotion recognition accuracy was achieved around 51%
Kao and Lee (2006)	Pich and power based features are extracted from frame, syllable, and word levels	Recognizing four emotions in Mandarin. Combination of features from frame, syllable and word level yielded 90% emotion recognition performance
Zhu and Luo (2007)	Duration, energy and pitch based features	Recognizing emotions in Mandarin Language. Sequential forward selection was used for best feature selection among the all prosodic features. Emotion classification studies are conducted on multi-speaker multi-lingual database. Modular neural networks were used as classifiers
Lugger and Yang (2007)	Eight static prosodic features and voice quality features	Classification of six emotions: anger, anxiety, boredom, happiness, Neutral and sadness from Berlin emotional speech corpus. Speaker independent emotion classification is performed using Bayesian classifiers
Wang et al. (2008)	Energy, pitch and duration based features	Classification of six emotions from Mandarin language. Around 88% of average emotion recognition rate is reported using SVM and genetic algorithm
Zhang et al. (2008)	Prosody and voice quality based features	Classification of four emotions anger, joy, neutral, and sadness from Chinese natural emotional speech corpus. Around 76% emotion recognition performance was achieved using support vector machine

3.4 Combination of features

Dellert et al. (1996b) was among the first to consider the problem of classifying utterances of human speech according to their emotional content. The Danish emotional speech database was examined by Ververidis et al. (2004); it consisted of two male and two female persons uttering sentences related to happiness, sadness, anger, surprise and neutral emotions. Out of the initially chosen 17 spectral, 36 pitch and 34 intensity (energy) features, computed as average over all the frames, 5 features by sequential forward selection method based on correct classification rate were selected by the Bayes (using Gaussian probability distribution functions) classifier. They reported a success-score of 54% using all data for training and testing and a success rate of 51.6% with cross-validation. Vogt and André (2006) used utterances of 10 neutral sentences of German language uttered by 5 male and 5 female persons, expressed in Anger, Disgust, Fear, Joy, Sad, Boredom and Neutral emotions taken from "Berlin Emotional Speech Database" recorded at Technical University, Berlin. They computed initially a total of 1280 features consisting of pitch, energy, MFCC, pauses, duration related to speaking rate which were finally reduced to 90–160 by the correlation-based feature selection method. A success of 69.1% using the full feature set and 77.4% using the reduced feature set was reported using a Native Bayes classifier.

Bitouk et al. (2010) worked on English and Emo-DB databases, computed MFCC features from consonants, stressed and unstressed vowels. Wu et al. (2011) computed modulation spectral features and prosodic features respectively with SVM and LDA classifiers. They computed 13 MFCCs, deltas and double deltas features, PLP features and prosodic features. For feature selection they had implemented SFS algorithm and FDR techniques. Kwon et al. (2003) explored log energy, formants, F0 information, mel based energy, MFCCs with their velocity and acceleration coefficient for speech emotion recognition. Four different classifiers—quadratic discriminant analysis (QDA), Support vector machine (SVM), Hidden Markov Model (HMM), and Linear discriminant analysis were considered for performance evaluation in the SUSAS database. Neutral, angry, Lombard, loud emotions were considered for the experiment. The best accuracy of 96.3% was achieved for stressed/neutral style classification, and 70.1% for 4 class speaking style classification using Gaussian SVM. For speaker-independent AIBO database 42.3% was achieved for 5 class emotions angry, bored, happy, neutral, and sad recognition. In Wang and Guan (2004) 25 prosodic features, 24 MFCCs, 6 formant frequency features were considered for recognizing human emotional state by using audiovisual signals. They have considered maximum likelihood classifier (MLC), Gaussian mixture model (GMM), neural network (NN), K-nearest neighbors (K-NN), and Fisher's linear discriminant analysis (FLDA) classifier s for speech emotion recognition system. Among all classifiers FLDA shows better recognition accuracy been observed. Kim et al. (2011) conducted experiments using the IEMOCAP database for prosodic and spectral features applied with SVM and deep belief network. Esmaileyan and Marvi (2014) worked on Persian emotional corpus where the utterances were cut from the radio programs of emotions anger, fear, joy, sadness and neutral. For prosodic and spectral features applied with LDA classifier, an accuracy rate of 55.74% for female and 47.28% for male was shown. They also conducted experiments using the Berlin database where the accuracy was found 78.64% for female and 73.40% for male. Amer et al. (2014) proposed a novel staged hybrid model for speech emotion detection. The proposed hybrid setup consisted of a generative model used for unsupervised representation of learning of short term temporal phenomenon, and a discriminative model which was used for event detection and classification of long range temporal dynamics. For evaluation they had taken three audio-visual datasets like AVEC, VAM and SPD and for classification deep networks were implemented. Rao et al. (2013) proposed local prosodic and global prosodic features for speech emotion analysis. In this study the duration, pitch and energy values carries prosodic information and global features represent mean, minimum, maximum, standard deviation and slope of the prosodic contours. The SVM classifier was used for emotion classification on IIT-KGP-SESC Telugu emotional corpus. In Alonso et al. (2015) a SVM classifier was established to get the best classification performance results for emotion classification in German corpus, the English corpus, and the Polish corpus which includes emotions like happy, boredom, neutral, anger and sadness. The classifier was used by the author for extraction of two kinds of features, two prosodic features and four paralinguistic features related to pitch and spectral energy balance. The results indicate different classification accuracies for three databases with lower complexity as compared to other approaches in real time applications. The analysis was done separately and different accuracies found like 94.9% for EMO-DB, 88.32% for LDC, and 90% for Polish database. Wang et al. (2015) formulated a new model called Fourier parameter model. The perceptual content of voice quality and first and second order differences were used in this model, for Speaker-Independent speech emotion recognition. Evaluation of the parameters was done by using SVM and Bayesian classifier in German database (EMODB), Chinese language database (CASIA) and Chinese elderly emotion database (EESDB) containing emotions like happy, surprise, neutral, sadness, anger and fear. The recognition rate was found out by using FP features over MFCC features was 16.2, 6.8, 16.6 points but the combination of FP and MFCC features showed better accuracy like 17, 5, 10, 10.5 points. Lin and Wei (2005) examined two classification

Table 4 Literature survey of vocal tract features used for speech emotion recognition

New et al. (2001)	LFPC (log-frequency power coefficient), MFCC (mel-frequency cepstral coefficients) and LPCC (linear prediction cepstral coefficients)	Text independent method of emotion classification	The discrete HMM classifier was considered for the classification of emotions: anger, fear, joy, sadness, disgust and surprise emotions in Mandarin and Burmese language
Lee et al. (2004)	MFCC (mel-frequency cepstral coefficients)	Phoneme-level modeling for the classification of emotional states from speech	Hidden Markov models (HMM) based on short-term spectral features are used for the experiment, two sets of HMM classifiers were discussed: a generic set of "emotional speech" HMMs (one for each emotion) and a set of broad phonetic-class based HMMs for each emotion type considered. The emotions considered: anger, happiness, neutral, and sadness
Bitouk et al. (2010)	Mel-frequency cepstral coefficients computed over three phoneme type classes of interest—stressed vowels, unstressed vowels and consonants in the utterance	Class level spectral features analysis for speech emotion recognition from an English emotional speech database from Linguistic Data Consortium (LDC) and Berlin database of German emotional speech	SVM classifiers with radial basis kernels
Koolagudi et al. (2012)	MFCCS (Mel frequency cepstral coefficients)	Feature extraction Semi natural database (Graphic Era University Semi Natural Emotion Speech Corpus) was considered	Emotion classification using GMM classifier: anger, happy, neutral, sad and surprise
Neiberg et al. (2006)	Combination of MFCCs and MFCC low features	Emotion classification using Swedish and English emotional speech databases	GMM classifier was considered for classification purpose
Lotfian and Busso (2015)	WPCC	Odia database containing 117 speakers	Polynomial classifier was used for speech emotion recognition
Dellaert et al. (1996a)	MFCC, RASTA feature, RMS energy, voice quality feature	A novel approach to create lexical dependent models at utterance-level using synthetic speech	SVM Classifier was used for performance evaluation, in Semaine database for speech emotion recognition system

methods, the hidden Markov model (HMM) and the support vector machine (SVM), to classify five emotional states anger, happiness, sadness, surprise and neutral. In the HMM method, 39 candidate instantaneous features were extracted, and the Sequential Forward Selection (SFS) method was used to find the best feature subset. The classification performance of the selected feature subset was then compared with that of the Mel frequency cepstrum coefficients (MFCC). Within the method based on SVM, a new vector measuring the difference between Mel frequency scale subbands energies is proposed. The performance of the K-nearest Neighbors (KNN) classifier using the proposed vector was also investigated. Both gender dependent and gender independent experiments were conducted on the Danish Emotional Speech (DES) Database. The recognition rates by the HMM classifier were 98.9% for female subjects, 100% for male subjects, and 99.5% for gender independent cases. When the SVM classifier and the proposed feature vector were employed, correct classification rates of 89.4, 93.6 and 88.9% were obtained for male, female and gender independent cases respectively. The work presented by Luengo et al. (2005) shows an experiment based on Basque speech emotional database for automatic speech emotion recognition. They built three classifiers for their experiment. The first classifier was using a spectral feature and GMM, second classifier using prosodic feature and SVM and third classifier using prosodic feature and GMM. Total 86 prosodic features were evaluated and the accuracy was achieved after feature selection was 92.3%. In Rong et al. (2009) acoustic features were used for speech emotion recognition system using actual and acted dataset in Chinese language. Two classification methods C4.5 decision tree algorithm and Random forest algorithm were applied to evaluate the quality of the feature pitch, intensity, zero crossing rate, spectral features Mel-scale frequency cepstral coefficients. It has been observed that 72.25% is the best recognition accuracy achieved by Random forest classifier. Sheikhan et al. (2013) proposed a modular neural support vector machine (SVM) classifier for speech emotion recognition. The study shows the performance of the proposed classifier compared with other classifiers—GMM, Multilayer perceptron neural network and C5.0-based classifiers. The proposed neural-SVM classifier achieved 8% improved recognition accuracy over other classifiers. The Farsi neutral speech corpus with 6000 utterances from 300 speakers with various accents had been collected for the experiment. Each speaker uttered 202 sentences in three emotional states: neutral, happiness, anger. The features used for the experiment 12 MFCCS, log energy (LE), the first three formant frequencies and pitch frequency. For feature selection combination of ANOVA and Tukey methods were considered. Rao and Koolagudi (2011) explored speech features to identify Hindi dialects and emotions. In this work, they have considered five prominent dialects of Hindi: Chhattisgarhi (spoken in central India), Bengali (Bengali accented Hindi spoken in Eastern region), Marathi (Marathi accented Hindi spoken in Western region), General (Hindi spoken in Northern region) and Telugu (Telugu accented Hindi spoken in Southern region) for the identification task. The work shows the performance of both dialect identification and emotion identification from speech. Speech database considered for the dialect identification task consists of spontaneous speech spoken by male and female speakers. Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) was used for conducting the emotion recognition studies. The different types of emotion considered in this study were anger, disgust, fear, happy, neutral and sad. Spectral and Prosodic features educed from speech are used for discriminating the dialects and emotions. Spectral features are represented by Mel frequency cepstral coefficients (MFCC) and prosodic features are represented by durations of syllables, pitch and energy contours. Auto-associative neural network (AANN) models and Support Vector Machines (SVM) are explored for capturing the dialect specific and emotion specific information from the above specified features. Classification systems were developed separately for dialect classification and emotion classification. Recognition performance of the dialect identification and emotion recognition systems was found to be 81 and 78% respectively. Table 5 shows list of combination of features with classifiers.

4 Conclusions

In the recent past extensive effort have been expended by researchers in the area of emotion recognition via speech. In this study a significant number of research papers were surveyed based on three parameters—database, feature extraction and classifiers. This paper summarizes some of the research work carried out by various workers between 2000 and 2017 on speech emotion recognition system. A bulk of the current research also focuses on feature extraction and selection in order to select the best features and improve the performance accuracy. It has been noted from analyses of the data that to improve the system performance and identify the correct emotions, classifier selection is a challenging task. Many classifiers have been chosen for speech emotion recognition system but it is very difficult to conclude which performs better-there is no clear winner. Recent works focus mostly on deep neural network and architecture, hybrid classifiers and fusion methods for emotion recognition system. It is evident from this review that enough scope exists for the development of new protocols along with their proper design in the area of speech emotion recognition. Higher accuracy,

Table 5 Literature survey of combination of features used for speech emotion recognition

References	Features	Purpose	Classifier
Bozkurt et al. (2009)	Spectral, prosodic and HMM based features	Classification of five emotions of INTERSPEECH 2009 challenge	Emotion classification accuracy was found around 63%
Nakatsu et al. (2000)	Combination of LPCCs and pitch related features	100 phonetically balanced words were recorded using 100 native speakers (50 male and 50 female)	Artificial neural network was used for classification and performance analysis of eight emotions for speech emotion recognition system
Iliev et al. (2010)	Glottal symmetry and MFCC features	Speech emotion classification	Classification of four emotions using optimum path forest classifier
Jeon et al. (2013)	Spectral and prosody	Chinese: ACC database (CA), EMO-DB (simulated parallel), English: EMA database (EA), German: EMO-DB database (GA), English: IEMOCAP database (ES), four emotions (anger, happiness, sadness and neutral). Cross lingual/corpus effect of emotion recognition from speech	Support vector machines (SVM), implemented in WEKA, a three-order polynomial kernel and multi-class (4-way classification) discrimination was used
Yeh and Chi (2010)	Spectral and prosodic features 13 MFCCs, their deltas, double-deltas were computed as 156 features per utterance and 30 prosodic features	EMO-DB (simulated parallel), seven emotions (anger, happiness, fear, disgust, boredom, sadness and neutral)	SVM used as the classifier for emotion analysis
Espinosa et al. (2010)	Spectral, prosody and voice quality features	VAM (natural), German spontaneous emotional speech, three categories (arousal, valence and dominance)	SVM and Pace Regression classifiers chosen for best recognition accuracy. Bagging ensemble of Pace Regression classifiers shows estimations reached an average correlation of 0.6885 and a mean error of 0.1433
Rozgic et al. (2012)	Spectral, prosodic and lexical	USC-IEMOCAP (semi-natural), four emotions (anger, happiness, sadness and neutral)	Support vector machine and Gaussian mixture model was chosen for classification. The fusion of acoustic and lexical features delivers an emotion recognition accuracy of 65.7%
Atassi and Esposito (2008)	Prosody and voice quality (emotion selective)	EMO-DB (simulated parallel), six emotions (anger, happiness, fear, disgust, boredom and sadness)	Gaussian mixture model classifier was used and SFFS feature selection algorithms was chosen for selecting best features

specificity and reproducibility are some of the criteria that should be considered in developing such new protocols.

Acknowledgements The authors are grateful for the valuable input given by Prof. J. Talukdar, Silicon Institute of Technology, Bhubaneswar, Odisha.

References

Abrilian, S., Devillers, L., & Martin, J. C. (2006). Annotation of emotions in real-life video interviews: Variability between coders. In *5th international conference on language resources and evaluation (LREC 06)*, Genoa, pp. 2004–2009.

Agrawal, S. S. (2011). Emotions in Hindi speech-analysis, perception and recognition. In *International conference on speech database and assessments (Oriental COCOSDA)*.

Agrawal, S. S., Jain, A., & Arora, S. (2009). Acoustic and perceptual features of intonation patterns in Hindi speech. *In International workshop on spoken language prosody (IWSLPR-09)*, Kolkata, pp. 25–27.

Alonso, J. B., Cabrera, J., Medina, M., & Travieso, C. M. (2015). New approach in quantification of emotional intensity from the speech signal: Emotional temperature. *Experts Systems with Applications, 42*, 9554–9564.

Amer, M. R., Siddiquie, B., Richey, C., & Divakaran, A. (2014). Emotion detection in speech using deep networks. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 3724–3728.

Amir, N., Ron, S., & Laor, N. (2000). Analysis of an emotional speech corpus in Hebrew based on objective criteria. In *Proceedings of ISCA workshop speech and emotion*, Belfast, Vol. 1, pp. 29–33.

Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *The Journal of the Acoustical Society of America, 52*(6), 1687–1697.

Atassi, H., & Esposito, A. (2008). A speaker independent approach to the classification of emotional vocal expressions. In *IEEE international conference on tools with artificial intelligence (ICTAI'08)*, Dayton, Ohio, USA, Vol 2, pp 147–152.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70*(3), 614–636.

Bapineedu, G., Avinash, B., Gangashetty, S. V., & Yegnanarayana, B. (2009). Analysis of Lombard speech using excitation source information. In *INTERSPEECH-09*, Brighton, UK, pp. 1091–1094.

Batliner, A., Biersack, S., & Steidl, S. (2006). The prosody of pet robot directed speech: Evidence from children. In *Speech prosody*, Dresden, pp. 1–4.

Batliner, A., Hacker, C., Steidl, S., Noth, E., D'Arcy, S., Russell, M., & Wong, M. (2004). You stupid tin box—children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proceedings of language resources and evaluation (LREC 04)*, Lisbon.

Batliner, A., Huber, R., Niemann, H., Nöth, E., Spilker, J., & Fischer, K. (2000). The recognition of emotion. In *Verbmobil: Foundations of speech-to-speech translation*, pp. 122–130.

Bitouk, D., Verma, R., & Nenkova, A. (2010). Class-level spectral features for emotion recognition. *Speech Communication*, *52*(7–8), 613–625.

Borden, G., Harris, K., & Raphael, L. (1994). *Speech science primer: Physiology, acoustics, and perception of speech* (3rd ed.). Baltimore: Williams and Wilkins.

Bozkurt, E., Erzin, E., & Erdem, A. T. (2009). Improving automatic emotion recognition from speech signals. In *10th annual conference of the international speech communication association (INTERSPEECH)*, Brighton, UK, pp. 324–327.

Brester, C., Semenkin, E., & Sidorov, M. (2016). Multi-objective heuristic feature selection for speech-based multilingual emotion recognition. *JAISCR, 6*(4), 243–253.

Buck, R. (1999). The biological affects, a typology. *Psychological Review, 106*(2), 301–336.

Bulut, M., Narayanan, S. S., & Syrdal, A. K. (2002). Expressive speech synthesis using a concatenative synthesizer. In *Proceedings of international conference on spoken language processing (ICSLP'02)*, Vol. 2, pp. 1265–1268.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. In *Proceedings of the INTERSPEECH 2005*, Lissabon, Portugal, pp. 1517–1520.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S. et al. (2008) IEMOCAP: Interactive emotional dyadic motion capture database. In: *Language resources and evaluation*.

Caballero-Morales, S. O. (2013) Recognition of emotions in Mexican Spanish speech: An approach based on acoustic modelling of emotion-specific vowels. *The Scientific World Journal*, *2013*, 1–13.

Caldognetto, E. M., Cosi, P., Drioli, C., Tisato, G., & Cavicchio, F. (2004). Modifications of phonetic labial targets in emotive speech: Effects of the co-production of speech and emotions. *Speech Communication, 44*, 173–185.

Chauhan, A., Koolagudi, S. G., Kafley, S. & Rao, K. S. (2010). Emotion recognition using LP residual. In *Proceedings of the 2010 IEEE students' technology symposium*, IIT Kharagpu.

Chen, L., Mao, X., Xue, Y., & Lung, L. (2012). Speech emotion recognition: Features and classification models. *Digital Signal Processing, 22*(6), 1154–1160.

Chuang, Z.-J., & Wu, C.-H. (2002). Emotion recognition from textual input using an emotional semantic network. In *Proceedings of international conference on spoken language processing (ICSLP'02)*, Vol. 3, pp. 2033–2036.

Cichosz, J., & Slot, K. (2005). Low-dimensional feature space derivation for emotion recognition. In *INTERSPEECH'05*, Lisbon, Portugal, pp. 477–480.

Costantini, G., Iaderola, I., Paoloni, A., & Todisco, M. (2014). EMOVO Corpus: An Italian emotional speech database. In *Proceedings of the 9th international conference on language resources and evaluation—LREC 14*, pp. 3501–3504.

Cummings, K. E., & Clements, M. A. (1998). Analysis of the glottal excitation of emotionally styled and stressed speech. *The Journal of the Acoustical Society of America*, *98*, 88–98.

Darwin, C. (1872/1965). *The expression of the emotions in man and animals*. Chicago University Press, Chicago.

Dellaert, F., Polzin, T., & Waibel, A. (1996a). Recognising emotions in speech. In *ICSLP 96*.

Dellert, F., Polzin, T., & Waibel, A. (1996b). Recognizing emotion in speech. In *4th international conference on spoken language processing*, Philadelphia, PA, USA, pp. 1970–1973.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication, 40*, 33–60.

Eckman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*, 169–200.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion*. Sussex: Wiley.

EI Ayadi M, Kamel MS, Karray F (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 1*(44), 572–587.

Engberg, I., & Hansen, A. (1996). "Documentation of the Danish emotional speech database" des. Retrieved from http://cpk.auc.dk/tb/speech/Emotions/.

Esmaileyan, Z., & Marvi, H. (2014). A database for automatic Persian speech emotion recognition: Collection, processing and evaluation. *IJE Transactions A: Bascis, 27*(1), 79–90.

Espinosa, H. P., Garcia, J. O., & Pineda, L. V. (2010). Features selection for primitives estimation on emotional speech. In *ICASSP*, Florence, Italy, pp. 5138–5141

Fernandez, R., & Picard, R. W. (2003). Modeling driver's speech under stress. *Speech Communication, 40*, 145–159.

Shah, A. F., Vimal Krishnan, V. R., Sukumar, A. R., Jayakumar, A., & Anto, P. B. (2009). Speaker independent automatic emotion recognition in speech: A comparison of MFCCs and discrete wavelet transforms. In *International conference on advances in recent technologies in communication and computing, ARTCom '09*.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotion is not two dimensional. *Psychological Science, 13*, 1050–1057.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering, 7*, 829–837.

Gangamohan, P., Kadiri, S. R., Gangashetty, S. V., & Yegnanarayana, B. (2014). Excitation source features for discrimination of anger and happy emotions. In: *INTERSPEECH*, Singapore, pp. 1253–1257.

Gangamohan, P., Kadiri, S. R., & Yegnanarayana, B. (2013). Analysis of emotional speech at sub segmental level. In *Interspeech*, Lyon, France, pp. 1916–1920.

Gomez, P., & Danuser, B. (2004). Relationships between musical structure and physiological measures of emotion. *Emotion, 7*(2), 377–387.

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera Ammittag German audio-visual emotional speech database. In *International conference on multimedia and expo*, pp. 865–868.

Grimm, M., Mower, E., Kroschel, K., & Narayanan, S. (2006). Combining categorical and primitives-based emotion recognition. In *14th European signal processing conference (EUSIPCO 2006)*, Florence, Italy.

Haq, S., & Jackson, P. J. B. (2009). Speaker-dependent audio-visual emotion recognition. In *Proceedings of international conference on auditory-visual speech processing*, pp. 53–58.

He, L., Lech, M., & Allen, N. (2010). On the importance of glottal flow spectral energy for the recognition of emotions in speech. In *INTERSPEECH 2010*, Makuhari, Chiba, Japan, pp. 26–30.

Hozjan, V., & Kacic, Z. (2003). *Improved emotion recognition with large set of stastical features*. Geneva: Eurospecch.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., & Nogueiras, A. (2002). Interface databases: Design and collection of a multilingual emotional speech database. In *Proceedings of the 3rd international conference on language (LREC'02) Las Palmas de Gran Canaria*, Spain, pp. 2019–2023.

Iliev, A. I., Scordilis, M. S., Papa, J. P., & Falco, A. X. (2010). Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech and Language, 24*(3), 445–460.

Iliou, T., & Anagnostopoulos, C.-N. (2009). Statistical evaluation of speech features for emotion recognition. In *Fourth international conference on digital telecommunications*, Colmar, France, pp. 121–126.

Iriondo, I., Guaus, R., & Rodriguez, A. (2000). Validation of an acoustical modeling of emotional expression in Spanish using speech synthesis techniques. In *Proceedings of ISCA workshop speech and emotion*, Belfast, Vol. 1, pp. 161–166.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review, 99*, 561–565.

Jeon, J. H., Le, D., Xia, R., & Liu, Y. (2013). A preliminary study of cross-lingual emotion recognition from speech: Automatic classification versus human perception. In *Interspeech*, Layon, France, pp. 2837–2840.

Jiang, D.-N., & Cai, L. H. (2004). Classifying emotion in Chinese speech by decomposing prosodic features. In *International conference on speech and language processing (ICSLP)*, Jeju, Korea.

Jiang, D.-N., Zhang, W., Shen, L.-Q., & Cai, L.-H. (2005). Prosody analysis and modelling for emotional speech synthesis. In *IEEE proceedings of ICASSP 2005*, pp. 281–284.

Jin, X., & Wang, Z. (2005). *An emotion space model for recognition of emotions in spoken Chinese* (pp. 397–402). Berlin: Springer.

Jovičić, S. T., Kašić, Z., Đorđević, M., & Rajković, M. (2004). Serbian emotional speech database: Design, processing and evaluation. In *SPECOM 9th conference speech and computer*, St. Petersburg, Russia.

Kadiri, S. R., Gangamohan, P., Gangashetty, S. V., & Yegnanarayana, B. (2015). Analysis of excitation source features of speech for emotion recognition. In *INTERSPEECH 2015*, Dresden, pp. 1324–1328.

Kandali, A. B., Routray, A., & Basu, T. K. (2008a). Emotion recognition from Assamese speeches using MFCC features and GMM classifier. In *Proceedings of IEEE region 10 conference on TENCHON*.

Kandali, A. B., Routray, A., & Basu, T. K. (2008b). Emotion recognition from speeches of some native languages of ASSAM independent of text and speaker. In *National seminar on Devices, Circuits, and Communications*, B. I. T. Mesra, Ranchi, pp. 6–7.

Kao, Y.-H., & Lee, L.-S. (2006). Feature analysis for emotion recognition from Mandarin speech considering the special characteristics of Chinese language. In *INTERSPEECH-ICSLP*, Pittsburgh, Pennsylvania, pp. 1814–1817.

Kim, J. B., Park, J. S., Oh, Y. H. (2011). On-line speaker adaptation based emotion recognition using incremental emotional information. In *ICASSP*, Prague, Czech Republic, pp. 4948–4951.

Koolagudi, S. G., Devliyal, S., Chawla, B., Barthwal, A., & Rao, K. S. (2012). Recognition of emotions from speech using excitation source features. *Procedia Engineering, 38*, 3409–3417.

Koolagudi, S. G., & Krothapalli, S. R. (2012). Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. *International Journal of Speech Technology, 15*(4), 495–511.

Koolagudi, S. G., Maity, S., Kumar, V. A., Chakrabati, S., & Rao, K. S. (2009). *IITKGP-SESC: Speech database for emotion analysis. Communications in computer and information science, LNCS* (pp. 485–492). Berlin: Springer.

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

Koolagudi, S. G., & Rao, K. S. (2012b). Emotion recognition from speech using source, system, and prosodic features. *International Journal of Speech Technology, 15*(2), 265–289.

Koolagudi, S. G., Reddy, R., & Rao, K. S. (2010). Emotion recognition from speech signal using epoch parameters. In *International conference on signal processing and communications (SPCOM)*.

Krothapalli, S. R., & Koolagudi, S. G. (2013). Characterization and recognition of emotions from speech using excitation source information. *International Journal of Speech Technology, 16*(2), 181–201.

Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion recognition by speech signals. In *EUROSPEECH*, pp. 125–128,.

Lanjewar, R. B., Mauhurkar, S., & Patel, N. (2015). Implementation and comparison of speech emotion recognition system using

Gaussian mixture model and K-nearest neighbor techniques. *Procedia Computer Science, 49*, 50–57.

Lazarus, R. S. (1991). *Emotion & adaptation*. New York: Oxford University Press.

Lee, C. M., & Narayanan, S. (2003). Emotion recognition using a data-driven fuzzy inference system. In *European conference on speech and language processing (EUROSPEECH)*, Geneva, Switzerland, pp. 157–160.

Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing, 13*(2), 293–303.

Lee, C. M., Narayanan, S., & Pieraccini, R. (2001). Recognition of negative emotion in the human speech signals. In *Workshop on auto, speech recognition and understanding*.

Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z. et al. (2004). Emotion recognition based on phoneme classes. In *8th international conference on spoken language processing, INTERSPEECH 2004*, Korea.

Lee, C.-C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication, 53*, 1162–1171.

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

Lin, Y.-L., & Wei, G. (2005). Speech emotion recognition based on HMM and SVM. In: *Fourth International conference on machine learning and cybernetics*, Guangzhou, pp. 4898–4901.

Lotfian, R., & Busso, C. (2015). Emotion recognition using synthetic speech as neutral reference. In *IEEE International conference on ICASSP*, pp. 4759–4763.

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

Lugger, M., & Yang, B. (2007). The relevance of voice quality features in speaker independent emotion recognition. In *ICASSP*, Honolulu, Hawaii, pp. IV17–IV20.

Makarova, V., & Petrushin, V. A. (2002). RUSLANA: A database of Russian emotional utterances. In *7th International conference on spoken language processing (ICSLP 02)*, pp. 2041–2044.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4), 561–580.

McGilloway, S., Cowie, R., Douglas-Cowie, E., Gielen, S., Westerdijk, M., & Stroeve, S. (2000) Approaching automatic recognition of emotion from voice: A rough benchmark. In *Proceedings of ISCA workshop speech emotion*, pp. 207–212.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2007). The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *Journal of LATEX Class Files, 6*(1), 1–14.

Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., Bozzali, M., & Di Natale, C. (2014). Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowledge-Based Systems, 63*, 68–81.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *Proceedings of IEEE conference on ICASSP*, pp. 2227–2231.

Mohanty, S., & Swain, B. K. (2010). Emotion recognition using fuzzy K-means from Oriya speech. In *International Conference [ACCTA-2010] on Special Issue of IJCCT*, Vol. 1 Issue 2–4.

Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., & Pardo, J. M. (1999). Analysis and modeling of emotional speech in Spanish. In *Proceedings of international conference on phonetic sciences*, pp. 957–960.

Morrison, D., Wang, R., & De Silva, L. C. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication, 49*, 98–112.

Nakatsu, R., Nicholson, J., & Tosa, N. (2000). Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Knowledge-Based Systems, 13*, 497–504.

Nandi, D., Pati, D., & Rao, K. S. (2017). Parametric representation of excitation source information for language identification. *Computer Speech and Language, 41*, 88–115.

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

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

New, T. L., Wei, F. S., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication, 41*, 603–623.

Nicholson, J., Takahashi, K., & Nakatsu, R. (2006). Emotion recognition in speech using neural networks. *Neural Computing & Applications, 11*, 290–296.

Nogueiras, A., Marino, J. B., Moreno, A., & Bonafonte, A. (2001). Speech emotion recognition using hidden Markov models. In *Proceedings of European conference on speech communication and technology (Eurospeech'01)*, Denmark.

Nordstrand, L., Svanfeld, G., Granstrom, B., & House, D. (2004). Measurements of ariculatory variation in expressive speech for a set of Swedish vowels. *Speech Communication, 44*, 187–196.

Ooi, C. S., Seng, K. P., Ang, L.-M., & Chew, L. W. (2014). A new approach of audio emotion recognition. *Experts Systems with Applications, 41*, 5858–5869.

Pao, T.-L., Chen, Y.-T., Yeh, J.-H., & Liao, W.-Y. (2005). Combining acoustic features for improved emotion recognition in Mandarin speech. In *International conference on affective computing and intelligent interaction*, pp. 279–285.

Park, C.-H., & Sim, K.-B. (2003). Emotion recognition and acoustic analysis from speech signal. In *Proceedings of the international joint conference on neural networks*, pp. 2594–2598.

Pereira, C. (2000). Dimensions of emotional meaning in speech. In *Proceedings of ISCA workshop speech and emotion*, Belfast, Vol. 1, pp. 25–28.

Petrushin, V. A. (1999). Emotion in speech: Recognition and application to call centers. In *Proceedings of the 1999 conference on artificial neural networks in engineering (ANNIE 99)*.

Picard, R. W. (1997). *Affective computing*. Cambridge: The MIT Press.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*, 1175–1191.

Power, M., & Dalgleish, T. (2000). *Cognition and emotion from order to disorder*. New York: Psychology Press.

Prasanna, S. R. M., & Govind, D. (2010). Analysis of excitation source information in emotional speech. In *INTERSPEECH 2010*, Makuhari, Chiba, Japan, pp. 781–784.

Prasanna, S. R. M., Gupta, C. S., & Yegnanarayana, B. (2006). Extraction of speaker-specific excitation information from linear prediction residual of speech. *Speech Communication, 48*, 1243–1261.

Pravena, D., & Govind, D. (2017). Significance of incorporating excitation source parameters for improved emotion recognition from speech and electroglottographic signals. *International Journal of Speech Technology, 20*(4), 787–797.

Pravena, D., & Govind, D. (2017). Development of simulated emotion speech database for excitation source analysis. *International Journal of Speech Technology, 20*, 327–338.

Quiros-Ramirez, M. A., Polikovsky, S., Kameda, Y., & Onisawa, T. (2014). A spontaneous cross-cultural emotion database: Latin-America vs. Japan. In *International conference on Kansei Engineering and emotion research*, pp. 1127–1134.

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

Rahurkar, M. A., & Hansen, J. H. (2002). Frequency band analysis for stress detection using a Teager energy operator based feature. *Proceedings of International Conference on Spoken Language Processing* (*ICSLP'*), Vol. 3, issue 02, pp. 2021–2024.

Ramamohan, S., & Dandapat, S. (2006). Sinusoidal model-based analysis and classification of stressed speech. In *IEEE transactions on audio, speech and language processing*, Vol. 14, p. 3.

Rao, K. S., & Koolagudi, S. G. (2011). Identification of Hindi dialects and emotions using spectral and prosodic features of speech. *Systemics, Cybernetics, and Informatics, 9*(4), 24–33.

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

Rao, K. S., Kumar, T. P., Anusha, K., Leela, B., Bhavana, I., & Gowtham, S. V. S. K. (2012). Emotion recognition from speech. *International Journal of Computer Science and Information Technologies, 3*, 3603–3607.

Rao, K. S., Prasanna, S. R. M., & Yegnanarayana, B. (2007). Determination of instants of significant excitation in speech using Hilbert envelope and group delay function. *IEEE Signal Processing Letters, 14*, 762–765.

Rao, K. S., & Yegnanarayana, B. (2006). Prosody modification using instants of significant excitation. In *IEEE transactions on audio and speech*, pp. 972–980.

Rong, J., Li, G., & Chen, Y. P. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. *Information Processing and Management, 45*, 315–328.

Rozgic, V., Ananthakrishnan, S., Saleem, S., Kumar, R., Vembu, A. N., & Prasad, R. (2012). Emotion recognition using acoustic and lexical features. In *INTERSPEECH*, Portland, USA.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*, 805–819.

Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality, 11*, 273–294.

Salovey, P., Kokkonen, M., Lopes, P., & Mayer, J. (2004). Emotional Intelligence: What do we know? In ASR Manstead, N. H. Frijda & A. H. Fischer (Eds.), *Feelings and emotions: The Amsterdam symposium* (pp. 321–340). Cambridge: Cambridge University Press.

Schachter, S., & Singer, J. (1962). Cognitive, social, and physiological determinants of emotional state. *Psychological Review, 69*, 379–399.

Scherer, K. R., Grandjean, D., Johnstone, T., Klasmeyer, G., & Banziger, T. (2002). Acoustic correlates of task load and stress. In *Proceedings of international conference on spoken language processing (ICSLP'02)*, Colorado, Vol. 3, pp. 2017–2020.

Schroder, M. (2000). Experimental study of affect bursts. In *Proceedings of ISCA workshop speech and emotion*, Vol. 1, pp. 132–137.

Schroder, M., & Grice, M. (2003). Expressing vocal effort in concatenative synthesis. In *Proceedings of international conference on phonetic sciences (ICPhS'03)*, Barcelona, pp. 2589–2592.

Schubert, E. (1999). Measurement and time series analysis of emotion in music, Ph.D dissertation, school of Music education, University of New South Wales, Sydeny, Australia.

Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model based speech emotion recognition. In *Proceedings of the International conference on multimedia and Expo, ICME*.

Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistis information in a hybrid support vector machine-belief network architecture. In *Proceedings of international conference on acoustics, speech and signal processing (ICASSP'04)*, Vol. 1, pp. 557–560.

Sheikhan, M., Bejani, M., & Gharavian, D. (2013). Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method. *Neural Computing and Applications, 23*(1), 215–227.

Slaney, M., & McRoberts, G. (2003). Babyears: A recognition system for affective vocalizations. *Speech Comunnication, 39*, 367–384.

Song, P., Ou, S., Zheng, W., Jin, Y., & Zhao, L. (2016). Speech emotion recognition using transfer non-negative matrix factorization. In *Proceedings of IEEE international conference ICASSP*, pp. 5180–5184.

Sun, R., & Moore, E. (2011). Investigating glottal parameters and teager energy operators in emotion recognition. In *Affective Computing and Intelligent Interaction*, pp. 425–434.

Takahashi, K. (2004). Remarks on SVM-based emotion recognition from multi-modal bio-potential signals. In *13th IEEE international workshop on robot and human interactive communication*, Roman.

Tao, J., & Kang, Y. (2005). Features importance analysis for emotional speech classification. In *Affective Computing and Intelligent Interaction*, pp. 449–457.

Tato, R., Santos, R., Kompe, R., & Pardo, J. M. (2002). Emotional space improves emotion recognition. In *Proceedings of international conference on spoken language processing (ICSLP'02)*, Colorado, Vol. 3, pp. 2029–2032.

Tomkins, S. (1962). *Affect imagery and consciousness: The positive affects*, Vol. 1. New York: Springer.

University of Pennsylvania Linguistic Data Consortium. (2002). *Emotional prosody speech and transcripts*. Retrieved from http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?CatalogId=LDC2002S28.

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

Ververidis, D., Kotropoulos, C., & Pitas, I. (2004). Automatic emotional speech classification. In *Proceedings of international conference on acoustics, speech and signal processing (ICASSP'04)*, Montreal, Vol. 1, pp. 593–596.

Vidrascu, L., & Devillers, L. (2005). Detection of real-life emotions in call centers. In *INTERSPEECH*, Lisbon, Portugal, pp. 1841–1844.

Vogt, T., & André, E. (2006). Improving automatic from speech via gender differentiation. In *Proceedings of language resources and evaluation conference (LREC 2006)*, Genoa.

Wakita, H. (1976). Residual energy of linear prediction to vowel and speaker recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 24*, 270–271.

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

Wang, Y., Du, S., & Zhan, Y. (2008). Adaptive and optimal classification of speech emotion recognition. In *Fourth international conference on natural computation*, pp. 407–411.

Wang, Y., & Guan, L. (2004). An investigation of speech based human emotion recognition. In *IEEE 6th workshop on multimedia signal processing*.

Werner, S., & Keller, E. (1994). Prosodic aspects of speech. In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition:*

Basic concepts, state of the art, the future challenges (pp. 23–40). Chichester: Wiley.

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

Wu, T., Yang, Y., Wu, Z., & Li, D. (2006). MASC: a speech corpus in mandarin for emotion analysis and affective speaker recognition. In *Speaker and language recognition workshop*.

Wu, W., Zheng, T. F., Xu, M.-X., & Bao, H.-J. (2006). Study on speaker verification on emotional speech. In *INTERSPEECH'06*, Piisburgh, Pennsylvania, pp. 2102–2105.

Wundt, W. (2013). *An introduction to psychology*. Read Books Ltd.

Yamagishi, J., Onishi, K., Maskko, T., & Kobayashi, T. (2003). *Emotion recognition using a data-driven fuzzy inference system*. Geneva: Eurospeech.

Yegnanarayana, B., & Gangashetty, S. (2011). Epoch-based analysis of speech signals. *S¯adhan¯a, 36*(5), 651–697.

Yegnanarayana, B., Swamy, R. K., & Murty, K. S. R. (2009). Determining mixing parameters from multispeaker data using speech-specific information. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(6), 1196–1207.

Yeh, L., & Chi, T. (2010). Spectro-temporal modulations for robust speech emotion recognition. In *INTERSPEECH*, Chiba, Japan, pp. 789–792.

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., & Narayanan, S. (2004). An acoustic study of emotions expressed in speech. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'04)*, Korea, Vol. 1, pp. 2193–2196.

You, M., Chen, C., Bu, J., Liu, J., & Tao, J. (1997). Getting started with susas: a speech under simulated and actual stress database. *Eurospeech, 4*, 1743–1746.

Yu, F., Chang, E., Xu, Y.-Q., & Shum, H.-Y. (2001). Emotion detection from speech to enrich multimedia content. In: *Proceedings of IEEE Pacific-Rim Conference on Multimedia*, Beijing, Vol. 1, pp. 550–557.

Yuan, J., Shen, L., & Chen, F. (2002). The acoustic realization of anger, fear, joy and sadness in Chinese. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'02)*, Vol. 3, pp. 2025–2028.

Zhang, S. (2008). Emotion recognition in Chinese natural speech by combining prosody and voice quality features. In Sun et al. (Ed.), *Advances in neural networks. Lecture notes in computer science* (pp. 457–464). Berlin: Springer.

Zhang, T., Hasegawa-Johnson, M., & Levinson, S. E. (2004). Children's emotion recognition in an intelligent tutoring scenario. In *Proceeding of the eighth European Conference on Speech Communication and Technology, INTERSPEECH*.

Zhu, A., & Luo, Q. (2007). Study on speech emotion recognition system in E-learning. In J. Jacko (Ed.), *Human computer interaction, Part III, HCII* (pp. 544–552). Berlin: Springer.