Int J Speech Technol (2012) 15:99–117
DOI 10.1007/s10772-011-9125-1

Emotion recognition from speech: a review

Shashidhar G. Koolagudi · K. Sreenivasa Rao

Received: 7 July 2011 / Accepted: 17 December 2011 / Published online: 4 January 2012

Abstract Emotion recognition from speech has emerged as an important research area in the recent past. In this regard, review of existing work on emotional speech processing is useful for carrying out further research. In this paper, the recent literature on speech emotion recognition has been presented considering the issues related to emotional speech corpora, different types of speech features and models used for recognition of emotions from speech. Thirty two representative speech databases are reviewed in this work from point of view of their language, number of speakers, number of emotions, and purpose of collection. The issues related to emotional speech databases used in emotional speech recognition are also briefly discussed. Literature on different features used in the task of emotion recognition from speech is presented. The importance of choosing different classification models has been discussed along with the review. The important issues to be considered for further emotion recognition research in general and in specific to the Indian context have been highlighted where ever necessary.

Keywords Emotion recognition · Simulated emotional speech corpus · Elicited speech corpus · Natural speech corpus · Excitation source features · System features · Prosodic features · Classification models

S.G. Koolagudi (✉) · K.S. Rao
School of Information Technology, Indian Institute of Technology Kharagpur, Kharagpur 721302, West Bengal, India
e-mail: koolagudi@yahoo.com

K.S. Rao
e-mail: ksrao@iitkgp.ac.in

1 Introduction

Speech is a complex signal containing information about message, speaker, language, emotion and so on. Most existing speech systems process studio recorded, neutral speech effectively, however, their performance is poor in the case of emotional speech. This is due to the difficulty in modeling and characterization of emotions present in speech. Presence of emotions makes speech more natural. In a conversation, non-verbal communication carries an important information like intention of the speaker. In addition to the message conveyed through text, the manner in which the words are spoken, conveys essential non-linguistic information. The same textual message would be conveyed with different semantics (meaning) by incorporating appropriate emotions. Spoken text may have several interpretations, depending on how it is said. For example, the word *'OKAY'* in English, is used to express admiration, disbelief, consent, disinterest or an assertion. Therefore understanding the text alone is not sufficient to interpret the semantics of a spoken utterance. However, it is important that, speech systems should be able to process the non-linguistic information such as emotions, along with the message. Humans understand the intended message by perceiving the underlying emotions in addition to phonetic information by using multi-modal cues. Non-linguistic information may be observed through (1) facial expressions in the case of video, (2) expression of emotions in the case of speech, and (3) punctuation in the case of written text. The discussion in this paper is confined itself to emotions or expressions related to speech. Basic goals of emotional speech processing are (a) understanding emotions present in speech and (b) synthesizing desired emotions in speech according to the intended message. From machine's perspective understanding speech emotions can be viewed as classification or discrimination of emotions. Synthesis of

emotions can be viewed as incorporating emotion specific knowledge during speech synthesis.

Speech is one of the natural modalities of human machine interaction. Today's speech systems may reach human equivalent performance only when they can process underlying emotions effectively (O'Shaughnessy 1987). Purpose of sophisticated speech systems should not be limited to mere message processing, rather they should understand the underlying intentions of the speaker by detecting expressions in speech (Schroder 2001; Ververidis and Kotropoulos 2006). In the recent past, processing speech signal for recognizing underlying emotions is emerged as one of the important speech research areas. Embedding the component of *emotion processing* into existing speech systems makes them more natural and effective. Therefore, while developing speech systems (i.e., speech recognition, speaker recognition, speech synthesis and language identification), one should appropriately utilize the knowledge of emotions.

Speech emotion recognition has several applications in day-to-day life. It is particularly useful for enhancing naturalness in speech based human machine interaction (Schuller et al. 2004; Dellert et al. 1996; Koolagudi et al. 2009). Emotion recognition system may be used in an onboard car driving system, where information about mental state of a driver may be used to keep him alert during driving. This helps avoiding some accidents, caused due to stressed mental state of the driver (Schuller et al. 2004). Call center conversation may be used to analyze behavioral study of call attendants with their customers, and helps to improve quality of service of a call attendant (Lee and Narayanan 2005). Interactive movie (Nakatsu et al. 2000), story telling (Charles et al. 2009) and E-tutoring (Ververidis and Kotropoulos 2006) applications would be more practical, if they can adapt themselves to listeners' or students' emotional states. The automatic way to analyze emotions in speech is useful for indexing and retrieval of the audio/video files based on emotions (Sagar 2007). Medical doctors may use emotional contents of a patient's speech as a diagnosing tool for various disorders (France et al. 2000). Emotion analysis of telephone conversation between criminals would help crime investigation department for the investigation. Conversation with robotic pets and humanoid partners would be more realistic and enjoyable, if they are able to understand and express emotions like humans do (Oudeyer 2003). Automatic emotion analysis may be useful in automatic speech to speech translation systems, where speech in language x is translated into other language y by the machine. Here, both emotion recognition and synthesis are used. The emotions present in source speech are to be recognized, and the same emotions are to be synthesized in the target speech, as the translated speech is expected to represent the emotional state of the original speaker (Ayadi et al. 2011). In aircraft cockpits, speech recognition systems trained to recognize stressed-speech are used for better performance (Hansen and Cairns 1995). Call analysis in emergency services like ambulance and fire brigade, may help to evaluate genuineness of requests. There are also some practical emotional speech systems available www.exaudios.com.

Some important research concerns in speech emotion recognition are discussed below in brief.

- The word *emotion* is inherently uncertain and subjective. The term *emotion* has been used with different contextual meanings by different people. It is difficult to define *emotion* objectively, as it is an individual mental state that arises spontaneously rather than through conscious effort. Therefore, there is no common objective definition and agreement on the term *emotion*. This is the fundamental hurdle to proceed with scientific approach toward research (Schroder and Cowie 2006).
- There are no standard speech corpora for comparing performance of research approaches used to recognize emotions. Most emotional speech systems are developed using full blown emotions, but real life emotions are pervasive and underlying in nature. Some databases are recorded using experienced artists, whereas some other are recorded using semi-experienced or inexperienced subjects. The research on emotion recognition is limited to 5–6 emotions, as most databases do not contain wide variety of emotions (Ververidis and Kotropoulos 2006).
- Emotion recognition systems developed using various features may be influenced by the speaker and language dependent information. Ideally, speech emotion recognition systems should be speaker and language independent (Koolagudi and Rao 2010).
- An important issue in the development of a speech emotion recognition systems is identification of suitable features that efficiently characterize different emotions (Ayadi et al. 2011). Along with features, suitable models are to be identified to capture emotion specific information from extracted speech features.
- Speech emotion recognition systems should be robust enough to process real-life and noisy speech to identify emotions.

This paper provides a review of literature on speech emotion recognition, in view of different types of emotional speech corpora used to develop the emotion recognition systems, emotion specific features extracted from different aspects of speech, classification models used for recognizing the emotions. Some directions for further research on speech emotion recognition are also discussed at the end of the paper.

The paper is organized as follows: review of some important existing emotional speech corpora is given in Sect. 2. Section 3 discusses role and review of different speech

features while developing emotion recognition systems. Review of classification models used for speech emotion recognition is briefly discussed in Sect. 4. Other research issues of general importance, useful for further research are discussed in Sect. 5. Paper concludes with Sect. 6, by providing summary.

2 Databases: a review

For characterizing emotions, either for synthesis or for recognition, suitable emotional speech database is a necessary prerequisite (Ververidis and Kotropoulos 2006). An important issue to be considered in evaluating the emotional speech systems is the quality of the databases used to develop and assess the performance of the systems (Ayadi et al. 2011). The objectives and methods of collecting speech corpora, highly vary according to the motivation behind the development of speech systems. Speech corpora used for developing emotional speech systems can be divided into 3 types namely:

1. Actor (Simulated) based emotional speech database
2. Elicited (Induced) emotional speech database
3. Natural emotional speech database.

The important properties of these databases are briefly mentioned in Table 1.

Simulated emotional speech corpora are collected from experienced and trained theater or radio artists. Artists are asked to express linguistically neutral sentences in different emotions. Recording is done in different sessions to consider the variations in the degree of expressiveness and physical speech production mechanism of human beings. This is one of the easier and reliable methods of collecting expressive speech databases containing wide range of emotions. More than 60% of the databases collected, for expressive speech research are of this kind. The emotions collected through simulated means are fully developed in nature, which are typically intense, and incorporate most of the aspects considered relevant for the expression of emotions (Schroder et al. 2001). These are also known as *full blown* emotions. Generally, it is found that acted/simulated emotions tend to be more expressive than real ones (Ayadi et al. 2011; Williams and Stevens 1972).

Elicited speech corpora are collected by simulating artificial emotional situation, without knowledge of the speaker. Speakers are made to involve in emotional conversation with anchor, where different contextual situations are created by anchor through the conversation to elicit different emotions from the subject, without his/her knowledge. These databases may be more natural compared to their simulated counterparts, but subjects may not be properly expressive, if they know that they are being recorded. Sometimes these databases are recorded by asking the subjects to involve in verbal interaction with computer whose speech responses are in turn controlled by the human being without the knowledge of the subjects (Batliner et al. 2000).

Unlike full blown emotions, natural emotions are mildly expressed. Sometimes, it may be difficult to clearly recognize these emotions. They are also known as *underlying emotions*. Naturally available real world data may be recorded from call center conversations, cockpit recordings

Table 1 Different types of databases used in speech emotion recognition

Type of database	Advantages	Disadvantages
Actor(Simulated)		
Eg: LDC speech corpus (Ververidis and Kotropoulos 2006), Emo-DB (Burkhardt et al. 2005), IITKGP-SESC (Koolagudi et al. 2009).	• Most commonly used. • Standardized. • Results can be compared easily. • Complete range of emotions is available. • Wide variety of databases in most of the languages is available.	• Acted speech tells how emotions should be portrayed rather than how they are portrayed. • Contextual, individualistic and purpose dependent information is absent. • Episodic in nature, which is not true in real world situation. • Often it is a read speech, not spoken.
Elicited(Induced)		
Eg: Wizard of Oz databases, ORESTEIA (McMahon et al. 2003).	• Nearer to the natural databases. • Contextual information is present, but it is artificial.	• All emotions may not be available. • If the speakers know that they are being recoded, the quality will be artificial.
Natural		
Eg: Call center conversations (Lee and Narayanan 2005), Cockpit recordings.	• These are completely naturally expressed. • Useful for real world emotion modeling.	• All emotions may not be available. • Copyright and privacy issues. • Overlapping of utterances. • Presence of background noise. • Contains multiple and concurrent emotions. • Pervasive in nature. • Difficult to model.

Table 2 Literature survey of speech databases used for emotion processing

S.No.	Emotions	Number of speakers	Type of database	Purpose and approach	Ref.
English emotional speech corpora					
01	Depression and neutral (02)	22 patients and 19 healthy persons	Simulated	Recognition. Prosody variations are analyzed with respect to the speech samples of depressed and healthy people.	Ambrus (2000)
02	Anger, disgust, fear, joy, neutral, sadness and surprise (07)	8 actors (2 per language)	Simulated	Synthesis. Emotional speech is recorded in 4 languages (English, Slovenian, Spanish, and French).	Alpert et al. (2001)
03	Anger, boredom, joy, and surprise (04)	51 children	Elicited	Recognition. Recorded at the university of Maribor, in German and English.	Batliner et al. (2004)
04	Anger, fear, happiness, neutral, and sadness (05)	40 native speakers	Natural	Recognition. Two broad domains of emotions are proposed based on prosodic features.	Cowie and Douglas-Cowie (1996)
05	Different natural emotions	125 TV artists	Natural	Recognition. It is known as Belfast natural database and is used for several emotion processing applications.	Cowie and Cornelius (2003)
06	Anger, boredom, fear, happiness, neutral, and sadness (06).	Single actor	Simulated	Synthesis. F_0, duration and energy are modeled for synthesizing the emotions.	Edgington (1997)
07	Depression and neutral (02)	70 patients 40 healthy persons	Natural	Recognition. F_0, amplitude modulation, formants, power distribution are used to analyze depressed and suicidal speech	France et al. (2000)
08	Depression and neutral (02)	Different native speakers	Elicited	Recognition.	Gonzalez (1999)
09	Negative and positive (02)	Customers and call attendants	Natural	Recognition. Call center conversations are recorded.	Lee and Narayanan (2005)
10	Annoyance, shock and stress (03)	29 Native speakers	Elicited	Recognition.	McMahon et al. (2003)
11	Hot anger, cold anger, happiness, neutral, and sad (05), 40 utterances per emotion are recorded.	29 native speakers	Elicited	Recognition. Dimensional analysis of emotions is performed using F0 parameters.	Pereira (2000)
12	Anger, fear, neutral, and sad (04)	Different native speakers	Simulated	Recognition. Prosodic, spectral and verbal cues are used for emotion recognition.	Polzin and Waibel (2000)
13	5 Stress levels (05)	6 Soldiers	Natural	Recognition.	Rahurkar and Hansen (2002)
14	2 Task load stress conditions and 2 normal stress conditions (02)	100 Native speakers	Natural	Recognition. Effects of stress and load on speech rate, F0, energy, and spectral parameters are studied. The databases are recorded in English and German	Scherer et al. (2002)

Table 2 (*Continued*)

S.No.	Emotions	Number of speakers	Type of database	Purpose and approach	Ref.
15	Approval, attention, and prohibition (03)	12 Native speakers	Natural	Recognition. Pitch and broad spectral shapes are used to classify adult-directed and infant-directed emotional speech (BabyEars). The databases are recorded in English and German	Slaney and McRoberts (2003)
16	Anger, happiness, neutral, sad (04), 112 utterances per emotion are recorded.	Single actress	Simulated	Recognition. Speech prosody, vowel articulation and spectral energy distribution are used to analyze 4 emotions.	Yildirim et al. (2004)
German emotional speech corpora					
17	Anger, Boredom, disgust, fear, joy, neutral, and sad (07)	10 Actors	Simulated	Synthesis.	Burkhardt and Sendlmeier (2000)
18	Different elicited emotions are recorded.	51 School children (21M+30F)	Elicited	Recognition. Children are asked to spontaneously react with Sony AIBO pet robot. Around 9.5 hours of effective emotional expressions of children are recorded	Batliner et al. (2006)

during abnormal conditions, a dialog between patient and a doctor, emotional conversations in public places and so on. But, it is difficult to find wide range of emotions in this category. Annotation of these emotions is also highly subjective (expert opinion based) and categorization is always debatable. There are also some legal issues such as, privacy and copyright while using natural speech databases (Batliner et al. 2000; Ayadi et al. 2011). Table 1 briefly explains advantages and drawbacks of the given three types of emotional speech databases.

Design and collection of emotional speech corpora mainly depends on the research goals. For example: single speaker emotional speech corpus would be enough for the purpose of emotional speech synthesis, whereas, for recognizing emotions needs database with multiple speakers and various styles of expressing the emotions. The survey presented in this section gives the information about the emotional speech databases based on the language, number of emotions and the method of collection. The general issues to be considered while recording the speech corpus are as follows.

- The scope of emotion database both in terms of number of subjects contributing for recording and number of emotions to be recorded is to be decided properly (Douglas-Cowie et al. 2003).
- The decision about the nature of the speech to be recorded as natural or acted, helps to decide the quality and applications of the database.
- Proper contextual information is essential, as naturalness of expressions mainly depends upon the linguistic content and its context.
- Labeling of soft emotions present in the speech databases is highly subjective and utmost care has to be taken while labeling. Getting the data annotated using multiple experts and choosing the majority decision would be an acceptable approach.
- Size of the database used for speech emotion recognition plays an important role in deciding the properties such as scalability, generalisability, and reliability of the developed systems. Most of the existing emotional speech databases used for developing emotion systems are too small in size (Douglas-Cowie et al. 2003).

The properties of some important and representative emotional speech corpora being used for emotional speech research are briefly discussed in Tables 2 and 3. From tables, it may be observed that, there is a huge disparity among the databases, in terms of language, number of emotions, number of subjects, purpose and methods of database collection.

The set of emotional speech databases, given in Tables 2 and 3, is dominated by English language, followed by German and Chinese. Very few databases are collected in languages such as: Russian, Dutch, Slovenian, Swedish,

Table 3 Literature survey of speech databases used for emotion processing

S.No.	Emotions	Number of speakers	Type of database	Purpose and approach	Ref.
19	Anger, Boredom, disgust, fear, joy, neutral, and sad (07)	10 Actors (5M+5F)	Simulated	Recognition. About 800 utterances are recorded using 10 neutral German sentences.	Burkhardt et al. (2005)
20	Soft, modal, and loud (03)	Single actor	Simulated	Synthesis. Di-phone based approach is used for emotional speech synthesis.	Schroder and Grice (2003)
21	Anger, Boredom, disgust, and worry (04)	6 Native speakers	Simulated	Recognition. Affective bursts and short emotional non-speech segments are analyzed for discriminating the emotions.	Schroder (2003)
22	Two emotions for each emotional dimension are recorded. (1) Activation (calm-excited), (2) Valence (positive-negative), and (3) Dominance (weak-strong)	104 Native speakers (44M+60F)	Natural	Recognition. 12 hours 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.	Grimm et al. (2008)
Chinese emotional speech corpora					
23	Antipathy, anger, fear, happiness, sad, and surprise (06).	Two actors	Simulated	Recognition.	Wu et al. (2006)
24	Anger, disgust, fear, joy, sad, and surprise (06), 60 Utterances per emotion per speaker are recorded	12 Actors	Simulated	Recognition. Log frequency power coefficients are used for emotion recognition using HMMs.	Nwe et al. (2003)
25	Anger, happiness, neutral, and sad (04), 721 short utterances per emotion are recorded	Native TV actors	Simulated	Recognition.	Yu et al. (2001a)
26	Anger, fear, joy, neutral and sad (05), 288 sentences per emotion are recorded	9 Native speakers	Elicited	Recognition. Phonation, articulation and prosody are used to classify 4 emotions.	Yuan et al. (2002)
Spanish emotional speech corpora					
27	Desire, disgust, fear, fury (anger), joy, sadness, and surprise (07)	8 Actors (4M+4F)	Simulated	Synthesis. Acoustic modeling of Spanish emotions is studied. Rules are used to identify significant behavior of emotional parameters.	Iriondo et al. (2000)
28	Anger, disgust, happiness, and sadness (04), 2000 phones per emotion are considered	Single actor	Simulated	Synthesis. Pitch, tempo, and stress are used for emotion synthesis.	Montro et al. (1999)
Japanese emotional speech corpus					
29	Anger, joy, and sadness (03)	2 Native speakers	Simulated	Synthesis. Concatenative synthesis approach is used.	Iida et al. (2003)

Table 3 (*Continued*)

S.No.	Emotions	Number of speakers	Type of database	Purpose and approach	Ref.
Russian emotional speech corpus					
30	Anger, fear, happiness, neutral, sad, and surprise (06), 10 sentences are recorded per emotion in different sessions	61 Native speakers	Simulated	Recognition. This database is used for both language and speech processing applications (RUSSLANA).	Makarova and Petrushin (2002)
Swedish emotional speech corpus					
31	Happiness and neutral (02)	Single native speaker	Simulated	Synthesis. Variations in articulatory parameters are used for recording Swedish vowels in 2 emotions	Nordstrand et al. (2004)
Italian emotional speech corpus					
32	Anger, disgust, fear, joy, sad, and surprise (06)	Single native speaker	Simulated	Synthesis.	Caldognetto et al. (2004)

Japanese and Spanish. There is no reported reference of an emotional speech database in any of the Indian languages. Among emotional speech databases given in Tables 2 and 3, 24 speech corpora are collected for the purpose of recognition and 8 are collected with the intention of synthesis. Subjective listening tests confirm that, average emotion recognition rate in the case of any database has not crossed beyond 80%. For full blown emotions subjective listening tests have shown more than 90% of recognition performance. Most automatic emotion recognition systems have achieved recognition performance close to the results of subjective listening tests. About 70% of databases contain only 4–5 basic emotions. Few emotional speech databases contain 7–8 emotions. Most existing databases rarely contain the uncommon emotions like: antipathy, approval, attention, prohibition, etc. Majority of the databases contain clearly distinguishable emotions such as anger, sad, happy and neutral. Since, actor based simulated database collection is a straight forward and comparatively easy process, more than half of the databases mentioned in Tables 2 and 3 belong to the category of simulated databases. Sometimes depending upon the need, emotional speech conversations are also recorded from TV shows, and later annotation of emotions is performed by expert artists. From the available emotional speech databases, it is observed that, there is no standard, internationally approved database available for emotion processing. Recently COCOSDA, The International Committee for the Coordination and Standardization of Speech Databases and Assessment Techniques, which promotes collaboration and information exchange in speech research, has adopted emotional speech as a future priority theme www.slt.atr.co.jp/cocosda. 'HUMAINE', a group of researchers dedicated to the speech emotion recognition, has started *INTERSPEECH emotion challenge* since 2009, to facilitate feature, classifier, and performance comparison for non-prototypical spontaneous emotion recognition. In Indian context, some organizations such as Linguistic Data Consortium- Indian Languages (LDC-IL), Center for Development of Advanced Computing (CDAC), Tata Institute of Fundamental Research (TIFR), Department of Information Technology (DIT-Technology Development for Indian Languages) are contributing toward speech data collection. However, they are specifically confined to collect speech corpora in different Indian languages for the purpose of speech recognition/synthesis and speaker recognition tasks.

From above mentioned survey, it is observed that, there is an acute need of generic emotional speech databases to research community rather than purpose driven corpora. There is a necessity of properly designed, phonetically balanced natural emotional speech databases covering wide range of emotions. These databases may be internationally standardized and used for both emotion recognition and synthesis. This effort of creating standard databases should be extended to all the major languages, to promote cross lingual

and language specific speech emotion recognition. Different databases are to be designed and collected for analyzing the influence of textual information on expression of emotions (Database with neutral text, database containing emotionally salient words).

3 Features: a review

Choosing suitable features for developing any of the speech systems is a crucial decision. The features are to be chosen to represent intended information. Different speech features represent different speech information (speaker, speech, emotion and so on) in highly overlapped manner. Therefore in speech research, very often features are selected on experimental basis, and sometimes using the mathematical approach like PCA (Principal component analysis). The following subsections present the literature on three important speech features namely: excitation source, vocal tract system, and prosodic features.

3.1 Excitation source features: a review

Speech features derived from excitation source signal are known as source features. Excitation source signal is obtained from speech, after suppressing vocal tract (VT) characteristics. This is achieved by, first predicting the VT information using filter coefficients (linear prediction coefficients (LPCs)) from speech signal, and then separating it by inverse filter formulation. The resulting signal is known as *linear prediction residual*, and it contains mostly the information about the excitation source (Makhoul 1975). In this paper, features derived from LP residual are referred to as excitation source, sub-segmental, or simply source features. The sub-segmental analysis of speech signal is aimed at studying characteristics of glottal pulse, open and closed phases of glottis, strength of the excitation and so on. The characteristics of the glottal activity, specific to the emotions may be estimated using the excitation source features. The LP residual signal and the glottal volume velocity (GVV) signal are explored in literature as the correlates of excitation source information (Kodukula 2009). In literature, very few attempts have been made to explore the excitation source information for developing any of the speech systems. The reasons may be

1. Popularity of the spectral features.
2. The excitation signal (LP residual) obtained from the LP analysis is viewed mostly as an error signal (Ananthapadmanabha and Yegnanarayana 1979) due to unpredictable component of the speech signal.
3. The LP residual basically contains higher order relations, and capturing these higher order relations is not well known (Yegnanarayana et al. 2002).

It may be difficult to parameterize LP residual signal, but it contains valid information as it provides primary excitation to the vocal tract system, while producing speech. LP residual signal basically contains the higher order correlations among its samples (Bajpai and Yegnanarayana 2008), as the first and the second order correlations are filtered out during LP analysis. These higher order correlations may be captured to some extent, by using the features like strength of excitation, characteristics of glottal volume velocity waveform, shapes of the glottal pulse, characteristics of open and closed phases of glottis and so on.

The existing studies based on excitation source features of speech have clearly demonstrated that excitation source information contains all flavors of speech such as message, speaker, language, and emotion specific information. However, the available excitation source features may not compete with well established spectral and prosodic features. Some of the important references regarding the use of excitation information in developing different speech systems are given below. Pitch information extracted from LP residual signal is successfully used in Atal (1972), for speaker recognition. LP residual energy is used in Wakita (1976), for vowel and speaker recognition. Cepstral features derived from LP residual signal are used in Thevenaz and Hugli (1995), for capturing the speaker specific information. The combination of features derived from LP residual and LP residual cepstrum has been used to minimize the equal error rate in case of speaker recognition (Liu and Palm 1997). By processing LP residual signal using Hilbert envelope and group delay function, the instants of significant excitation are accurately determined (Rao et al. 2007b).

The higher order relations among samples of the LP residual are also used for categorizing different audio documents like: sports, news, cartoons, music in noisy and clean environments (Bajpai and Yegnanarayana 2004). The instants of significant excitation obtained from LP residual signal during the production of voiced speech are used to determine the relative delays between the speech segments of different speakers in multi-speaker environment, and they are further used to enhance the speech of individual speakers (Yegnanarayana et al. 2009). The epoch (instants of glottal closure) properties of LP residual are exploited in Yegnanarayana et al. (1998), for enhancing the reverberant speech. The parameters extracted from the excitation source signal at the epoch locations, are exploited for analyzing loudness, lombard effect, speaking rate and characteristics of laughter segments (Kumar et al. 2009; Bapineedu et al. 2009; Seshadri and Yegnanarayana 2009). Table 4 briefs out some of the important achievements in speech research using excitation source information.

From available literature, it is observed that, excitation source information is equally important to develop speech

Table 4 Literature review on different speech tasks using excitation source features

Sl.No	Features	Purpose and approach	Ref.
01	LP residual energy	Vowel and speaker recognition	Wakita (1976)
02	LP residual	Detection of instants of significant excitation.	Rao et al. (2007b)
03	Higher order relations among LP residual samples	Categorizing audio documents	Bajpai and Yegnanarayana (2004)
04	LP residual	Speech enhancement in multi-speaker environment	Yegnanarayana et al. (2009)
05	LP residual	Characterizing loudness, lombard effect, speaking rate, and laughter segments	Bapineedu et al. (2009)
06	Glottal excitation signal	Analyzing the relation between emotional state of the speaker and glottal activity	Cummings and Clements (1995)
07	Glottal excitation signal	To analyze emotion related disorders	Cummings and Clements (1995)
08	Excitation source signal	To discriminate emotions in continuous speech	Hua et al. (2005)

systems compared to spectral and prosodic features. Excitation source information is not exhaustively and systematically explored for speech emotion recognition. The excitation source signal may also contain the emotion specific information, in the form of higher order relations among linear prediction (LP) residual samples, parameters of instants of significant excitation, parameters of glottal pulse and so on. There is very little work done on emotion recognition using excitation source information (I and Scordilis 2011; Chauhan et al. 2010; Koolagudi et al. 2010). Hence, there is a scope for conducting the detailed and systematic study on excitation source information for characterizing the emotions.

3.2 Vocal tract features: a review

Generally, a speech segment of length 20–30 ms is used to extract vocal tract system features. It is known that, vocal tract characteristics are well reflected in frequency domain analysis of speech signal. The Fourier transform of a speech frame gives short time spectrum. Features like formants, their bandwidths, spectral energy and slope may be observed from spectrum. The cepstrum of a speech frame is obtained by taking the Fourier transform on log magnitude spectrum (Rabiner and Juang 1993). The MFCCs (Mel frequency cepstral coefficients) and the LPCCs (Linear prediction cepstral coefficients) are the common features derived from the cepstral domain that represent vocal tract information. These vocal tract features are also known as segmental, spectral or system features. The emotion specific information present in the sequence of shapes of vocal tract may be responsible for producing different sound units in different emotions. MFCCs, LPCCs, perceptual linear prediction coefficients (PLPCs), and formant features are some of the widely known system features used in the literature (Ververidis and Kotropoulos 2006). In general spectral features are treated as the strong correlates of varying shapes of the vocal tract and the rate of change in the articulator movements (Benesty et al. 2008).

Generally, spectral features have been successfully used for various speech tasks including development of speech and speaker recognition systems. Some of the important works on emotion recognition using spectral features are discussed below. The MFCC features are used in Mubarak et al. (2005), to distinguish speech and non-speech (music) information. It has been observed that the lower order MFCC features carry phonetic (speech) information, whereas higher order features contain non-speech (music) information. Combination of MFCCs, LPCCs, RASTA PLP coefficients and log frequency power coefficients (LFPCs) is used as the feature set, to classify anger, boredom, happy, neutral and sad emotions in Mandarin (Pao et al. 2005, 2007). Log frequency power coefficients (LFPC) are used to represent the emotion specific information in Williams and Stevens (1981), for classifying six emotions. A four stage ergodic hidden Markov model (HMM) is used as a classifier to accomplish this task. Performance of LFPC parameters is compared with conventional LPCC and MFCC features, and observed that LFPCs perform slightly better (Williams and Stevens 1981; Kamaruddin and Wahab 2009). The MFCC features extracted from lower frequency components (20 Hz to 300 Hz) of speech signal are proposed to model pitch variation. These are known as MFCC-low features and used to recognize emotions in Swedish and English emotional speech databases. MFCC-low features are reported to perform better than pitch features in case of emotion recognition (Neiberg et al. 2006). The mel-frequency cepstral coefficients computed over three phoneme classes namely: stressed vowels, unstressed vowels and consonants are used for speaker-independent emotion recognition. These features are referred to as class-level spectral features. Classification accuracies are observed to be consistently higher for class-level spectral features compared to prosodic or utterance-level spectral features. The combination of class-

Table 5 Literature review on emotion recognition using vocal tract features

Sl.No	Features	Purpose and approach	Ref.
01	MFCC features	Discrimination of speech and music. Higher order MFCCs contain more music specific information and lower number of MFCCs contain more speech specific information.	Mubarak et al. (2005)
02	MFCCs, LPCCs RASTA PLP coefficients, log frequency power coefficients	Classification of 4 emotions in Mandarin language. Anger, happy, neutral and sad emotions are considered in this study.	Pao et al. (2005, 2007)
03	Combination of MFCCs and MFCC-low features	Emotion classification using Swedish and English emotional speech databases.	Neiberg et al. (2006)
04	MFCC features from consonant, stressed and unstressed vowels (class-level MFCCs)	Emotion classification on English LDC and Emo-DB databases.	Bitouk et al. (2010)
05	Spectral features obtained using Fourier and Chirp transformations	Modeling human emotional states under stress.	Sigmund (2007)

level features with prosodic features improved the emotion recognition performance. Further, results showed that, spectral features computed from consonant regions contain more emotion specific information than either stressed or unstressed vowel features. It is also reported in this work that, the average emotion recognition performance is proportional to the length of the utterance (Bitouk et al. 2010). In Sigmund (2007), spectra of vowel segments obtained using Fourier and Chirp transforms are analyzed for emotion classification and observed that, the higher frequency regions of speech are suitable for characterizing stressed speech. These features are used to model the emotional state of a stressed person. Some of the efforts on the use of system features for speech emotion recognition are given in Table 5. From the references mentioned in Table 5, it is observed that, in most of the cases, spectral features are extracted through conventional block processing approach, wherein, entire speech signal is processed frame by frame, considering the frame size of 20 ms, and a shift of 10 ms. In reality, emotion specific information may be more prominent either in some emotion salient words or in some sub-syllabic regions like vowels and consonants. Different portions of the utterance carry different amount of emotion specific information depending upon the emotion expression pattern. Manifestation of emotions is a gradual process and may be observed clearly in finer spectral variations. Therefore, extension of spectral analysis of speech signal to the sub-utterance levels with smaller frame size may be useful study while characterizing and recognizing the emotions.

3.3 Prosodic features: a review

Human beings impose duration, intonation, and intensity patterns on the sequence of sound units, while producing speech. Incorporation of these prosody constraints (duration, intonation, and intensity), makes human speech natural. Prosody can be viewed as speech features associated with larger units such as syllables, words, phrases and sentences. Consequently, prosody is often considered as supra-segmental information. The prosody appears to structure the flow of speech. The prosody is represented acoustically by the patterns of duration, intonation (F_0 contour), and energy. They normally represent the perceptual speech properties, which are normally used by human beings to perform various speech tasks (Rao and Yegnanarayana 2006; Werner and Keller 1994). In the literature, mainly, pitch, energy, duration and their derivatives are used as the acoustic correlates of prosodic features (Banziger and Scherer 2005; Cowie and Cornelius 2003). Human emotional expressiveness (i.e. emotionally excited behavior of articulators) can be captured through prosodic features. The prosody can be distinguished at four principal levels of manifestation (Werner and Keller 1994). They are at (a) Linguistic intention level, (b) articulatory level, (c) acoustic realization level and (d) perceptual level.

At the linguistic level, prosody refers to relating different linguistic elements of an utterance to bring out required naturalness. For example, the linguistic distinctions that can be communicated through distinction between question and statement, or the semantic emphasis on an element. At the articulatory level, prosody is physically manifested as a series of articulatory movements. Thus, prosody manifestations typically include variations in the amplitudes of articulatory movements as well as the variations in air pressure. Muscle activity in the respiratory system as well as along the vocal tract, leads to radiation of sound waves. The acoustic realization of prosody can be observed and quantified using the analysis of acoustic parameters such as fundamental frequency (F_0), intensity, and duration. For example, stressed syllables have higher fundamental frequency, greater amplitude and longer duration than unstressed syllables. At the

Table 6 Literature review on emotion recognition using prosodic features

Sl.No	Features	Purpose and approach	Ref.
01	Initially 86 prosodic features are used, later best 6 features are chosen from the list	Identification of 4 emotions in Basque language. Around 92% Emotion recognition performance is achieved using GMMs.	Luengo et al. (2005)
02	35 dimensional prosodic feature vectors including pitch, energy, and duration are used	Classification of seven emotions of Berlin emotional speech corpus. Around 51% emotion recognition results are obtained for speaker independent cases using neural networks.	Iliou and Anagnostopoulos (2009)
03	Pitch and power based features are extracted from frame, syllable, and word levels	Recognizing 4 emotions in Mandarin. Combination of features from, frame, syllable and word level yielded 90% emotion recognition performance.	Kao and Lee (2006)
04	Duration, energy, and pitch based features	Recognizing emotions in Mandarin language. Sequential forward selection (SFS) is used to select best features from the pool of prosodic features. Emotion classification studies are conducted on multi-speaker multi-lingual database. Modular neural networks are used as classifiers.	Zhu and Luo (2007)
05	Eight static prosodic features and voice quality features	Classification of 6 emotions (anger, anxiety, boredom, happiness, neutral, and sadness) from Berlin emotional speech corpus. Speaker independent emotion classification is performed using Bayesian classifiers.	Lugger and Yang (2007)
06	Energy, pitch and duration based features	Classification of 6 emotions from Mandarin language. Around 88% of average emotion recognition rate is reported using SVM and genetic algorithms.	Wang et al. (2008)
07	Prosody and voice quality based features	Classification of 4 emotions namely anger, joy, neutral, and sadness from Mandarin language. Around 76% emotion recognition performance is reported using support vector machines (SVMs).	Zhang (2008)

perception level, speech sound waves enter ears of the listener who derives the linguistic and paralinguistic information from prosody via perceptual processing. During perception, prosody can be expressed in terms of subjective experience of the listener, such as pauses, length, melody and loudness of the perceived speech. It is difficult to process or analyze the prosody through speech production or perception mechanisms. Hence the acoustic properties of speech are exploited for analyzing the prosody.

In literature, prosodic features such as energy, duration, pitch and their derivatives are treated as high correlates of emotions (Dellaert et al. 1996; Lee and Narayanan 2005; Nwe et al. 2003; Schroder and Cowie 2006). Features such as, minimum, maximum, mean, variance, range and standard deviation of energy, and similar features of pitch are used as important prosodic information sources for discriminating the emotions (Schroder 2001; Murray and Arnott 1995). Some studies (Cahn 1990; Murray and Arnott 1995) have also tried to measure steepness of the F0 contour during rise and falls , articulation rate, number and duration of pauses for characterizing the emotions. Prosodic features extracted from the smaller linguistic units like syllables and at the level of consonants and vowels are also used for analyzing the emotions (Murray and Arnott 1995). The importance of prosodic contour trends in the context of different emotions is discussed in Murray et al. (1996), Scherer (2003). Peaks and troughs in the profiles of fundamental frequency and intensity, durations of pauses and bursts are proposed for identifying four emotions namely fear, anger, sadness and joy. Around 55% of average emotion recognition performance is reported using discriminant analysis (McGilloway et al. 2000). The sequences of frame wise prosodic features, extracted from longer speech segments such as words and phrases are also used to categorize the emotions present in the speech (Nwe et al. 2003). F_0 information is analyzed for emotion classification and it is reported that minimum, maximum and median values of F_0 and slopes of F_0 contours are emotion salient features. Around 80% of emotion recognition accuracy is achieved, using proposed F_0 features with K-nearest neighbor classifier (Dellert et al. 1996). Short time supra-segmental features such as pitch, energy, formant locations and their bandwidths, dynamics of pitch, energy and formant contours, speaking rate are used for analyzing the emotions (Ververidis and Kotropoulos 2006). The complex relations between pitch, duration and energy parameters are exploited in Iida et al. (2003) for detecting the speech emotions. Table 6 briefs out some of the other important and recent works on speech emotion recognition using prosodic features.

From literature, it is observed that, most speech emotion recognition studies are carried out using utterance level static (global) prosodic features (Nwe et al. 2003; Schroder and Cowie 2006; Dellaert et al. 1996; Koolagudi et al. 2009; Ververidis et al. 2004; Iida et al. 2003). Very few attempts have explored the dynamic behavior of prosodic patterns (local) for analyzing speech emotions (McGilloway et al. 2000; Rao et al. 2010). Elementary prosodic analysis of speech utterances is carried out in Rao et al. (2007a), at sentence, word, and syllable levels, using only the first order statistics of basic prosodic parameters. In this context, it is important to study the contribution of static and dynamic (i.e. global and local) prosodic features extracted from sentence, word and syllable segments toward emotion recognition. None of the existing studies has explored the speech segments with respect to their positional information for identifying the emotions. The approach of recognizing emotions from the shorter speech segments may further be helpful for real time emotion verification.

3.4 Combination of features: a review

Recent trends in research of speech emotion recognition, emphasized the use of combination of different features to achieve improvement in the recognition performance. Source, system, and prosodic features discussed in the previous subsections represent mostly mutually exclusive information of the speech signal. Therefore, these features are complementary in nature to each other. Intelligent combination of complementary features is expected to improve the intended performance of the system. Several studies on combination of features, proved to perform better emotion classification, compared to the systems developed using individual features. Some of the important works using the combination of different features for speech emotion recognition are discussed below. The role of voice quality in conveying the emotions, moods, and attitudes is studied in Gobl and Chasaide (2003) using spectral and prosodic features. The voice qualities considered in the study are: harsh voice, tense voice, modal voice, breathy voice, whisper, creaky voice and lax-creaky voice. The study reported that, these voice quality indicators are more effective in indicating underlying (mild) emotions than the full blown emotions. It is observed from the studies that, there is no one-to-one mapping between voice quality and an emotion; rather a given voice quality tends to be associated with multiple emotions (Gobl and Chasaide 2003). Along with F_0 information, log energy, formants, mel based energy, MFCCs with their velocity and acceleration coefficients are explored for emotion classification (Kwon et al. 2003). Language, speaker, and context independent speech emotion recognition is carried out in Wang and Guan (2004) using prosodic, mel-frequency cepstral coefficients (MFCCs), and formant frequency features (25 prosodic, 24 MFCCs and 6 formant frequencies) to distinguish 6 discrete emotions (Wang and Guan 2004). Prosodic (energy and pitch) and spectral features (12 LPCCs and corresponding Δ features) are used as emotion specific features in Nicholson et al. (1999) to discriminate anger, disgust, fear, joy, neutral, sadness, surprise, and teasing emotions collected from 50 male and 50 female native Japanese subjects. In the above study, around 50% recognition rate is reported using neural network classifiers (Nicholson et al. 1999). GMM super vectors computed on spectral and prosodic features are used to recognize 5 primary emotions (anger, happy, neutral, sad, and surprise) recorded in Mandarin language. The combination of features is reported to reduce an error rate compared to the error rate obtained using prosodic features alone (Zhou et al. 2009). Articulatory features in combination with spectral features are proposed for identifying the emotions in Mandarin language (Zhou et al. 2009). Long-term spectro-temporal speech features are proposed in Wu et al. (2009), to recognize 7 emotions of Berlin emotional speech corpus (Emo-DB). Their performance is found to be better compared to short-term spectral features and prosodic features.

An average emotion recognition accuracy of 88.6% is achieved by using a combined long term spectro-temporal and prosodic features for classifying 7 discrete emotions (Wu et al. 2009). A novel approach of combining acoustic features and linguistic information is proposed in Schuller et al. (2004), for discriminating seven discrete emotional states. Belief networks are used to spot the emotional phrases from the spoken words. Further, acoustic and linguistic information are combined by soft decision fusion using neural network classifiers. Emotion recognition rates of 26%, 40%, and 58% are reported, using acoustic, linguistic and combined information respectively (Schuller et al. 2004). The combination of language and discourse information is proposed in Lee and Narayanan (2005), for improving the discrimination between the positive and negative emotions, in the context of call center applications (Lee and Narayanan 2005). The Teager energy values and MFCC features are combined in Zhou et al. (2001), for classifying neutral and stressed speech. Some of the other important works on speech emotion recognition using combination of different features are mentioned in Table 7.

In literature discussed above, it is reported that, combining complimentary evidence either at the feature level or at the score level would show considerable gain in the performance of speech emotion recognition systems.

4 Classification models: a review

In literature, several pattern classifiers are explored for developing speech systems like, speech recognition, speaker recognition, emotion classification, speaker verification and

Table 7 Literature review on emotion recognition using combination of different features

Sl.No	Features	Purpose and approach	Ref.
Emotional speech research using the combination of system and prosodic features			
01	Combination of features related to spectral energy, speech prosody, and articulator activities	Classification of anger, happy, neutral and sad in English language. It is reported that, anger-happy and sad-neutral share similar acoustic properties. About 75% of average emotion recognition is achieved on 4 emotions	Yildirim et al. (2004)
02	Combination of LPCCs and pitch related features	Classification of 8 emotions. 100 phonetically balanced words are recorded using 50 male and 50 female native speakers. Around 50% average, speaker independent emotion classification is reported using artificial neural network.	Nakatsu et al. (2000)
03	Pitch, formant energy, and speaking rate features	Classification of anger, fear, happy, neutral, and sad emotions, portrayed by 30 non-professional artists. Average emotion classification of 70% is achieved using artificial neural networks.	Petrushin (1999)
04	Spectral, prosodic and HMM based features	Classification of five emotions of INTERSPEECH-2009 emotional speech corpus. Average emotion classification performance reported is about 63%.	Bozkurt et al. (2009)
05	Combination of 39 spectral and prosodic features	Characterization of 15 discrete emotions. Shorter utterances carry better emotion specific characteristics. Specific words in longer utterances carry more emotion specific information.	Tischer (1995)
06	Combination of spectral and prosodic features	Classification of 5 emotional states present in Danish emotional speech corpus (DES). Emotion recognition performance of around 52% is reported.	Ververidis et al. (2004)
07	Spectral and prosodic features	Classification of positive and negative emotions from the DES speech corpus. Around 83% of average emotion recognition using different classifiers is reported.	Hoque et al. (2006)
08	Spectral, prosodic, disfluence (pauses) and paralinguistic (crying, laughter) features	Classification of real life blended emotions, recorded from call center conversations. Around 80% discrimination is reported between negative and neutral emotions on 20 Hrs. French database.	Vidrascu and Devillers (2005)
Emotional speech research using the combination of source and system features			
01	Glottal symmetry and MFCC features	Emotion classification. Optimum path forest classifier is used to classify 4 emotions.	Iliev et al. (2010)
02	Excitation source signal and spectral features	Stress classification. Combination of glottal spectral slope and non-linear Teager energy operator is used.	Iliev et al. (2010)

so on. However justification for choosing a particular classifier to the specific speech task is not provided in many instances. Most of the times suitable classifiers are chosen based on either thumb rule or some past references. Few times a particular one is chosen among the available alternatives based on experimental evaluation. Wang et al. have conducted the studies on the performance of various classification tools as applied to speech emotion recognition (Wang and Guan 2004). In general, pattern recognizers used for speech emotion classification can be categorized into two broad types namely

1. Linear classifiers and
2. Non-linear classifiers.

A linear classifier performs the classification by making a classification decision based on the value of a linear combination of the object characteristics. These characteristics are also known as feature values and are typically presented to the classifier in the form of an array called as a feature vector. If the input feature vector to the classifier is a real vector $\vec{x}$, then the output score is given by $y = f(\vec{w} \cdot \vec{x}) = f(\sum_j w_j x_j)$, where $\vec{w}$ is a real vector of weights and f is a function that converts the dot product of the two vectors into the desired output. The weight vector $\vec{w}$ is learned from a set of labeled training samples. j is the dimension of the feature vectors. Often f is a simple function that maps all values above a certain threshold to the first class and all other values to the second class. A more complex f might give the probability that an item belongs to a certain class.

Non-linear weighted combination of object characteristics is used to develop non-linear classifiers. During imple-

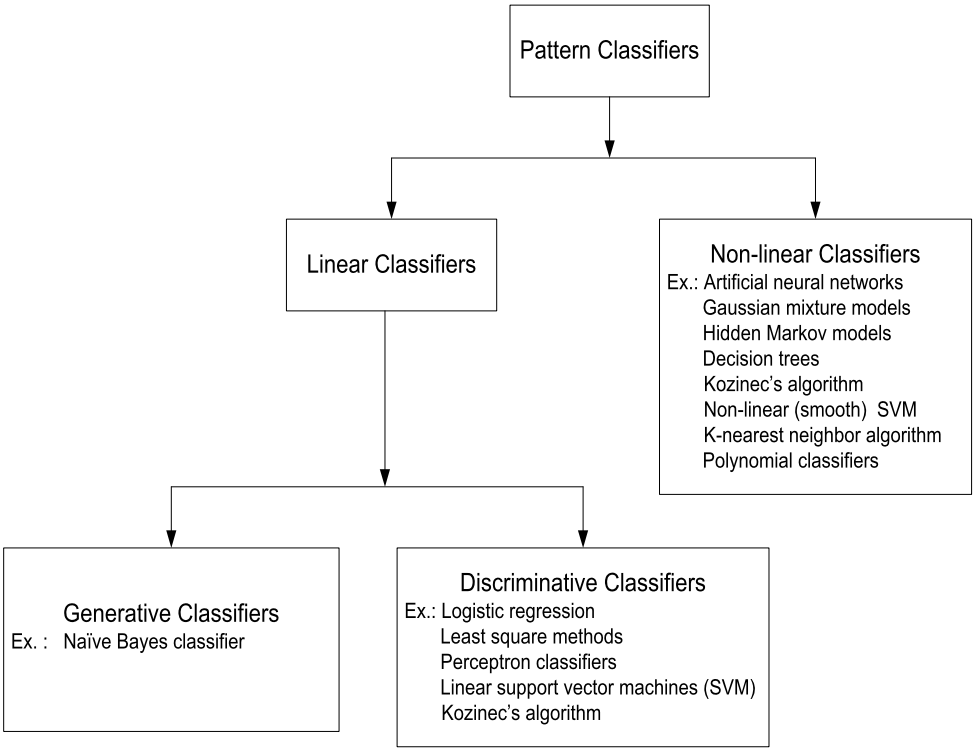

Fig. 1 Types of classifiers used for speech emotion recognition

mentation, proper selection of a kernel function makes the classifier either linear, or non-linear (Gaussian, polynomial, hyperbolic, etc.). In addition, each kernel function may take one or more parameters that would need to be set. Determining an optimal kernel function and parameter set for a given classification problem is not really a solved problem. There are only useful heuristics to reach satisfying performance. While adopting the classifiers to the specific problem, one should be aware of the facts that, non-linear classifiers have a higher risk of over-fitting, since they have more dimensions of freedom. On the other hand a linear classifier has less degree of freedom to fit the data points, and it severely fails in the case of data that is not linearly separable.

Determination of classifier parameters for linear classifiers is done by two broad methods. The first method uses probability density functions (generative classifiers) and the second method works on discriminative properties (discriminative classifiers) of the data points. Some important examples of classifiers using probability density functions are linear discriminant analysis, Fischer's linear discriminant analysis, Naive Bayes classifier, principal component analysis and so on. Important examples of linear classifiers working on discrimination of feature vectors are logistic regression, least square methods, perceptron algorithm, linear support vector machines, Kozinec's algorithm and so on. Discriminative classifiers perform mainly on the principle of non-probabilistic binary classification by adopting supervised learning. Whereas, probabilistic classifiers adopt unsupervised learning algorithms. Common non-linear classification tools used for general pattern recognition are Gaussian mixture models, hidden Markov models, soft (non-linear) SVMs (Support Vector Machines), neural networks, polynomial classifiers, universal approximators, and decision trees. Types of the pattern classifiers mainly used for speech emotion recognition are given in Fig. 1.

Use of classifiers mainly depends upon nature of the data. If nature of data is known before, then deciding on type of the classifier would be an easier task. Linear classifiers would classify the features better and faster, if they are clearly, linearly separable. Supervised learning would be helpful, if training data set is properly labeled. Feature vectors those are not linearly separable, would need non-linear classifiers for classification. In most of the real world situations, nature of the data is rarely known. Therefore, researchers use non-linear classifiers always at the cost of complexity and computational time. However systematic approach based on nature of speech features is required while choosing the pattern classifiers for emotion recognition. Diversified nature of features (excitation source, vocal tract system, and prosodic) would help to decide the use of proper classifier. Systematic study in this regard would be useful and appreciable as it saves lot of computational resources. Table 8 provides the list of classifiers used for speech emotion recognition.

Table 8 Literature review on use of different classifiers for speech emotion recognition task

Sl.No.	Classifiers	Features	References
01	Gaussian mixture models (GMM)	Prosodic	Slaney and McRoberts (2003), Schuller et al. (2004), Zhou et al. (2009), Neiberg et al. (2006), and Wang and Guan (2004)
		Spectral	Slaney and McRoberts (2003), Schuller et al. (2004), Zhou et al. (2009), Mubarak et al. (2005), Wang and Guan (2004), and Luengo et al. (2005)
02	Support vector machines (SVM)	Prosodic	Yu et al. (2001b), Schuller et al. (2004), Zhou et al. (2009), Luengo et al. (2005), Wang et al. (2008), Kao and Lee (2006), and Zhang (2008)
		Spectral	Yu et al. (2001b), Schuller et al. (2004), Zhou et al. (2009), and Kao and Lee (2006)
03	Artificial neural networks (ANN)	Prosodic	Petrushin (1999), Schuller et al. (2004), Nakatsu et al. (2000), Nicholson et al. (2000), Tato et al. (2002), Fernandez and Picard (2003), Zhu and Luo (2007), Nakatsu et al. (2000), Petrushin (1999), and Wang and Guan (2004)
		Spectral	Petrushin (1999), Schuller et al. (2004), Nakatsu et al. (2000), Nicholson et al. (1999), Nakatsu et al. (2000), Petrushin (1999), and Wang and Guan (2004)
04	k-Nearest neighbor classifier	Prosodic	Dellaert et al. (1996), Yu et al. (2001b), Pao et al. (2005), Wang and Guan (2004), and Dellert et al. (1996)
		Spectral	Petrushin (2000), Yu et al. (2001b), Lee et al. (2001), and Wang and Guan (2004)
05	Bayes classifier	Prosodic	Dellaert et al. (1996), Fernandez and Picard (2003), Lugger and Yang (2007), and Wang et al. (2008)
		Spectral	Lugger and Yang (2007)
06	Linear discriminant analysis with Gaussian probability distribution	Prosodic	Yildirim et al. (2004), Ververidis et al. (2004), and McGilloway et al. (2000)
		Spectral	Lee et al. (2001), Yildirim et al. (2004), Ververidis et al. (2004), and Lee and Narayanan (2005)
07	Hidden Markov models (HMM)	Prosodic	Fernandez and Picard (2003), Zhou et al. (2001), Nwe et al. (2003), and Bitouk et al. (2010)
		Spectral	Williams and Stevens (1981), Zhou et al. (2001), Nwe et al. (2003), and Kamaruddin and Wahab (2009)

5 Discussion on some important issues related to speech emotion recognition

Some of the important research issues in speech emotion recognition are discussed below in brief.

- Majority of the research results produced on emotional speech recognition have used databases with limited number of speakers. While developing emotion recognition systems using limited speaker databases; speaker specific information may play considerable role, if speech utterances of the same speakers are used for training and testing the models. On the other hand, developed models may produce poor results, due to lack of generality, if speech utterances of different speakers are used for training and testing the models. Therefore, there is a need of larger emotional speech database with reasonably large number of speakers and text prompts. Emotion recognition studies have to be conducted on large databases in view of speaker, text and session variabilities.
- Most research on emotional speech mainly focuses on characterizing the emotions from classification point of view. Hence, the main task carried out was deriving the emotion specific information from speech, and using it for classifying the emotions. On the other hand, emotion synthesis through speech is also an important task. Here, emotion specific information may be predicted from the text, and then it has to be incorporated during synthesis. For predicting the emotion specific information, appropriate models have to be developed using sufficiently large emotional speech corpus. In emotion synthesis, the major issues are the design of accurate prediction models and preparation of appropriate emotional speech corpus.
- Expression of emotions is an universal phenomenon, which may be independent of speaker, gender and language. Cross lingual emotion recognition study may be another interesting work for further research. The emotion recognition models developed using the utterances of a particular language should yield appreciably good recognition performance for any test utterance of the other language. By using cross lingual emotion analysis, one can group the languages based on their emotional similarity.
- Majority of the work done and results produced in the literature, are on recognizing speech emotions using simulated databases. Real challenge is to recognize speech emotions from natural emotions. The features and techniques discussed in the literature may be applied to the natural speech corpora, to analyze emotion recognition.

Realization of this, needs the collection of good natural emotional speech corpus, covering wide range of emotions, which is another challenge.

- More often, in the literature, emotion classification task is performed using single model (i.e., GMM, AANN, or SVM). Hybrid models can be explored for studying their performance in the case of emotion recognition. The basic idea behind using the hybrid models is that, they derive the evidence from different perspectives, and hence, the combination of evidence may enhance the performance, if the evidence are complementary in nature.
- The trend of emotion recognition is not clearly known in the case of many other languages. It would be helpful to evaluate the established features on different Indian languages for emotion recognition. This helps to comment on whether the methods and features used in literature are language independent? This analysis is also helpful to group languages based on their emotion characteristics, which in turn would improve the performance of language identification systems.
- The study on discrimination of emotions may be extended to the emotion dimensions (arousal, valance and power), that are derived from the psychology of production and perception of emotions. Deriving the appropriate speech features related to the emotion dimensions can be explored for further improving the recognition performance.
- Expression of emotions is a multi-modal activity. Therefore, other modalities like facial expression, bio-signals may be used as the supportive evidence along with the speech signal for developing the robust emotion recognition systems.
- The affect of emotion expression also depends upon the linguistic contents of the speech. Identification of emotion salient words from emotional speech, and the features extracted from these words along with other conventional features may enhance emotion recognition performance.
- In real time applications such as call analysis in the emergency services like ambulance and fire brigade, verification of emotions to analyze genuineness of requests is important. In this context, under the framework of emotion verification appropriate features and models can be explored.
- Most of the today's emotion recognition systems experience high influence of speaker specific information during emotion classification. An efficient technique may be developed to remove speaker specific information from the speech utterances.

6 Summary and conclusions

Processing of emotions from speech helps to assure naturalness in the performance of existing speech systems. Considerable amount of work in this area is done in the recent past. Due to lack of information and standardization lot of research overlap is a common phenomenon. Since 2006, exhaustive review paper is not published on speech emotion recognition, specifically in Indian context. Therefore, we thought that, the survey paper covering recent work in speech emotion recognition may ignite the research community for filling some important research gaps. This paper contains the review of recent works in speech emotion recognition from the points of views of emotional databases, speech features, and classification models. Some important research issues in the area of speech emotion recognition are also discussed in the paper.

References

Alpert, M., Pouget, E. R., & Silva, R. R. (2001). Reflections of depression in acoustic measures of the patient's speech. *Journal of Affective Disorders*, *66*, 59–69.

Ambrus, D. C. (2000). *Collecting and recording of an emotional speech database*. Tech. rep., Faculty of Electrical Engineering, Institute of Electronics, Univ. of Maribor.

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

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

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

Bajpai, A., & Yegnanarayana, B. (2004). Exploring features for audio clip classification using LP residual and AANN models. In *The international conference on intelligent sensing and information processing 2004 (ICISIP 2004)*, Chennai, India, Jan. 2004 (pp. 305–310).

Bajpai, A., & Yegnanarayana, B. (2008). Combining evidence from sub-segmental and segmental features for audio clip classification. In *IEEE region 10 conference TENCON*, India, Nov. 2008 (pp. 1–5). IIIT, Hyderabad.

Banziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, *46*, 252–267.

Bapineedu, G., Avinash, B., Gangashetty, S. V., & Yegnanarayana, B. (2009). Analysis of lombard speech using excitation source information. In *INTERSPEECH-09*, Brighton, UK, 6–10 September 2009 (pp. 1091–1094).

Batliner, A., Buckow, J., Niemann, H., Noth, E., & Warnke, V. (2000). *Verbmobile Foundations of speech to speech translation*. Berlin: Springer.

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

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

Benesty, J., Sondhi, M. M., & Huang, Y. (Eds.) (2008). *Springer handbook on speech processing*. Berlin: Springer.

Bitouk, D., Verma, R., & Nenkova, A. (2010, in press). Class-level spectral features for emotion recognition. *Speech Communication*.

Bozkurt, E., Erzin, E., Erdem, C. 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, Sept. 6–10, 2009 (pp. 324–327).

Burkhardt, F., & Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant synthesis. In *ITRW on speech and emotion*, Newcastle, Northern Ireland, UK, Sept. 2000 (pp. 151–156).

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A database of German emotional speech. In *Interspeech*.

Cahn, J. E. (1990). The generation of affect in synthesized speech. In *JAVIOS*, Jul. 1990 (pp. 1–19).

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*(1–4), 173–185.

Charles, F., Pizzi, D., Cavazza, M., Vogt, T., & Andr, E. (2009). Emoemma: Emotional speech input for interactive story telling. In Decker, Sichman, Sierra, & Castelfranchi (Eds.), *8th int. conf. on autonomous agents and multiagent systems (AAMAS 2009)*, Budapest, Hungary, May 2009 (pp. 1381–1382).

Chauhan, A., Koolagudi, S. G., Kafley, S., & Rao, K. S. (2010). Emotion recognition using lp residual. In *IEEE TechSym 2010*, West Bengal, India, April 2010. IIT Kharagpur: IEEE.

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

Cowie, R., & Douglas-Cowie, E. (1996). Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Fourth international conference on spoken language processing ICSLP 96*, Philadelphia, PA, USA, October 1996 (pp. 1989–1992).

Cummings, K. E., & Clements, M. A. (1995). Analysis of the glottal excitation of emotionally styled and stressed speech. *The Journal of the Acoustical Society of America*, *98*, 88–98.

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

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

Edgington, M. (1997). Investigating the limitations of concatenative synthesis. In *European conference on speech communication and technology (Eurospeech 97)*, Rhodes/Athens, Greece, 1997 (pp. 593–596).

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

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*, *47*(7), 829–837.

Gobl, C., & Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*, 189–212.

Gonzalez, G. M. (1999). *Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in Chicanos/Latinos*. Tech. report-39, Univ. Michigan.

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *IEEE international conference multimedia and expo*, Hanover, Apr. 2008 (pp. 865–868).

Hansen, J., & Cairns, D. (1995). Icarus: source generator based real-time recognition of speech in noisy stressful and lombard effect environments. *Speech Communication*, *16*(4), 391–422.

Hoque, M. E., Yeasin, M., & Louwerse, M. M. (2006). Robust recognition of emotion from speech. In *Intelligent virtual agents. Lecture notes in computer science* (pp. 42–53). Berlin: Springer.

Hua, L. Z., Yu, H., & Hua, W. R. (2005). *A novel source analysis method by matching spectral characters of LF model with STRAIGHT spectrum*. Berlin: Springer.

I, A. I., & Scordilis, M. S. (2001). Spoken emotion recognition using glottal symmetry. *EURASIP Journal on Advances in Signal Processing, 1*(11).

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

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, July 2009 (pp. 121–126).

Iriondo, I., Guaus, R., Rodrguez, A., Lzaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., & Longhi, L. (2000). Validation of an acoustical modeling of emotional expression in Spanish using speech synthesis techniques. In *ITRW on speech and emotion*, New Castle, Northern Ireland, UK, Sept. 2000.

Kamaruddin, N., & Wahab, A. (2009). Features extraction for speech emotion. *Journal of Computational Methods in Science and Engineering*, *9*(9), 1–12.

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, Sept. 2006 (pp. 1814–1817).

Kodukula, S. R. M. (2009). *Significance of excitation source information for speech analysis*. PhD thesis, Dept. of Computer Science, IIT, Madras.

Koolagudi, S. G., & Rao, K. S. (2010). Real life emotion classification using VOP and pitch based spectral features. In *INDICON*, (Kolkata, INDIA), Jadavpur University. New York: IEEE Press.

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

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)*, IISc, Bangalore, India, July 2010 (pp. 1–5). New York: IEEE Press.

Kumar, K. S., Reddy, M. S. H., Murty, K. S. R., & Yegnanarayana, B. (2009). Analysis of laugh signals for detecting in continuous speech. In *INTERSPEECH-09*, Brighton, UK, 6–10 September 2009 (pp. 1591–1594).

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

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

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

Liu, J. H. L., & Palm, G. (1997). On the use of features from prediction residual signal in speaker recognition. In *European conf. speech processing and technology (EUROSPEECH)* (pp. 313–316).

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

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

Honolulu, Hawaii, USA, May 2007 (pp. IV17–IV20). New York: IEEE Press.

Makarova, V., & Petrushin, V. A. (2002). RUSLANA: A database of Russian emotional utterances. In *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 *ISCA workshop on speech and emotion*, Belfast.

McMahon, E., Cowie, R., Kasderidis, S., Taylor, J., & Kollias, S. (2003). What chance that a DC could recognize hazardous mental states from sensor inputs? In *Tales of the disappearing computer*, Santorini, Greece.

Montro, J. M., Gutterrez-Arriola, J., Colas, J., Enriquez, E., & Pardo, J. M. (1999). Analysis and modeling of emotional speech in Spanish. In *Proc. int. conf. on phonetic sciences* (pp. 957–960).

Mubarak, O. M., Ambikairajah, E., & Epps, J. (2005). Analysis of an MFCC-based audio indexing system for efficient coding of multimedia sources. In *8th international symposium on signal processing and its applications*, Sydney, Australia, Aug. 2005.

Murray, I. R., & Arnott, J. L. (1995). Implementation and testing of a system for producing emotion by rule in synthetic speech. *Speech Communication*, *16*, 369–390.

Murray, I. R., Arnott, J. L., & Rohwer, E. A. (1996). Emotional stress in synthetic speech: Progress and future directions. *Speech Communication*, *20*, 85–91.

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.

Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. In *INTERSPEECH 2006 - ICSLP*, Pittsburgh, Pennsylvania, 17–19 September 2006 (pp. 809–812).

Nicholson, J., Takahashi, K., & Nakatsu, R. (1999). Emotion recognition in speech using neural networks. In *6th international conference on neural information processing (ICONIP-99)*, Perth, WA, Australia, Aug. 1999 (pp. 495–501).

Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural Computing & Applications*, *11*, 290–296.

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

Nwe, T. L., Foo, S. W., & Silva, L. C. D. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, *41*, 603–623.

O'Shaughnessy, D. (1987). *Speech communication human and machine*. Reading: Addison-Wesley.

Oudeyer, P. Y. (2003). The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, *59*, 157–183.

Pao, T. L., Chen, Y. T., Yeh, J. H., & Liao, W. Y. (2005). Combining acoustic features for improved emotion recognition in Mandarin speech. In J. Tao, T. Tan, & R. Picard (Eds.), *ACII. LNCS* (pp. 279–285). Berlin: Springer.

Pao, T. L., Chen, Y. T., Yeh, J. H., Cheng, Y. M., & Chien, C. S. (2007). *LNCS: Vol. 4738. Feature combination for better differentiating anger from neutral in Mandarin emotional speech*. Berlin: Springer.

Pereira, C. (2000). Dimensions of emotional meaning in speech. In *Proc. ISCA workshop on speech and emotion*, Belfast, Northern Ireland, 2000 (pp. 25–28).

Petrushin, V. (1999). *Emotion in speech: recognition and application to call centres. Artificial neural networks in engineering (ANNIE).*

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).*

Petrushin, V. A. (2000). Emotion recognition in speech signal: Experimental study, development and application. In *Proc. int. conf. spoken language processing*, Beijing, China.

Polzin, T., & Waibel, A. (2000). Emotion sensitive human computer interfaces. In *ISCA workshop on speech and emotion*, Belfast, 2000 (pp. 201–206).

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

Rahurkar, M., & Hansen, J. H. L. (2002). Frequency band analysis for stress detection using a Teager energy operator based feature. In *Proc. int. conf. on spoken language processing (ICSLP'02)* (pp. 2021–2024).

Rao, K. S., & Yegnanarayana, B. (2006). Prosody modification using instants of significant excitation. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*, 972–980.

Rao, K. S., Prasanna, S. R. M., & Sagar, T. V. (2007a). Emotion recognition using multilevel prosodic information. In *Workshop on image and signal processing (WISP-2007)*, Guwahati, India, Dec. 2007. Guwahati: IIT Guwahati.

Rao, K. S., Prasanna, S. R. M., & Yegnanarayana, B. (2007b). 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., Reddy, R., Maity, S., & Koolagudi, S. G. (2010). Characterization of emotions using the dynamics of prosodic features. In *International conference on speech prosody*, Chicago, USA, May 2010.

Sagar, T. V. (2007). *Characterisation and synthesis of emotions in speech using prosodic features*. Master's thesis, Dept. of Electronics and communications Engineering, Indian Institute of Technology Guwahati.

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

Scherer, K. R., Grandjean, D., Johnstone, L. T., & Klasmeyer, T. B. G. (2002). Acoustic correlates of task load and stress. In *International conference on spoken language processing ICSLP 02*, Colorado, 2002 (pp. 2017–2020).

Schroder, M. (2001). Emotional speech synthesis: A review. In *Seventh European conference on speech communication and technology, Eurospeech* Aalborg, Denmark, Sept. 2001.

Schroder, M. (2003). Experimental study of affect bursts. *Speech Communication, 40*(1–2). Special issue on speech and emotion.

Schroder, M., & Cowie, R. (2006). Issues in emotion-oriented computing toward a shared understanding. In *Workshop on emotion and computing (HUMAINE).*

Schroder, M., & Grice, M. (2003). Expressing vocal effort in concatenative synthesis. In *International conference on phonetic sciences ICPhS 03*, Barcelona.

Schroder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., & Gielen, S. (2001). Acoustic correlates of emotion dimensions in view of speech synthesis. In *7th European conference on speech communication and technology*, Aalborg, Denmark, Sept. 2001.

Schuller, B., Rigoll, G., & Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Proc. IEEE int. conf. acoust., speech, signal processing* (pp. 577–580). New York: IEEE Press.

Seshadri, G. P., & Yegnanarayana, B. (2009). Perceived loudness of speech based on the characteristics of glottal excitation source. *The Journal of the Acoustical Society of America*, *126*, 2061–2071.

Sigmund, M. (2007). Spectral analysis of speech under stress. *International Journal of Computer Science and Network Security*, *7*, 170–172.

Slaney, M., & McRoberts, G. (2003). BabyEars: a recognition system for affective vocalizations. *Speech Communication*, *39*, 367–384.

Tato, R., Santos, R., & Pardo, R. K. J. (2002). Emotional space improves emotion recognition. In *7th international conference on spoken language processing*, Denver, Colorado, USA, Sept. 16–20, 2002.

Thevenaz, P., & Hugli, H. (1995). Usefulness of LPC residue in textindependent speaker verification. *Speech Communication*, *17*, 145–157.

Tischer, B. (1995). *Acoustic correlates of perceived emotional stress*.

Ververidis, D., & Kotropoulos, C. (2006). A state of the art review on emotional speech databases. In *Eleventh Australasian international conference on speech science and technology*, Auckland, New Zealand, Dec. 2006.

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 *ICASSP* (pp. I593–I596). New York: IEEE Press.

Vidrascu, L., & Devillers, L. (2005). Real-life emotion representation and detection in call centers data. In J. Tao, T. Tan, & R. Picard (Eds.), *LNCS: Vol. 3784*. *ACII* (pp. 739–746). Berlin: Springer.

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, Y., & Guan, L. (2004). An investigation of speech-based human emotion recognition. In *IEEE 6th workshop on multimedia signal processing* (pp. 15–18). New York: IEEE Press.

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

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.

Williams, C., & Stevens, K. (1972). Emotions and speech: some acoustical correlates. *The Journal of the Acoustical Society of America*, *52*(4), 1238–1250.

Williams, C. E., & Stevens, K. N. (1981). Vocal correlates of emotional states. *Speech Evaluation in Psychiatry*, 189–220.

Wu, C. H., Chuang, Z. J., & Lin, Y. C. (2006). Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, *5*, 165–182.

Wu, S., Falk, T. H., & Chan, W. Y. (2009). Automatic recognition of speech emotion using long-term spectro-temporal features. In *16th international conference on digital signal processing*, Santorini-Hellas, 5–7 July 2009 (pp. 1–6). New York: IEEE Press.

Yegnanarayana, B., Murthy, P. S., Avendano, C., & Hermansky, H. (1998). Enhancement of reverberant speech using lp residual. In *IEEE international conference on acoustics, speech and signal processing*, Seattle, WA, USA, May 1998 (Vol. 1, pp. 405–408).

Yegnanarayana, B., Prasanna, S. R. M., & Rao, K. S. (2002). Speech enhancement using excitation source information. In *Proc. IEEE int. conf. acoust., speech, signal processing*, Orlando, Florida, USA, May 2002 (Vol. 1, pp. 541–544).

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.

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 *Int. conf. on spoken language processing (ICSLP 2004)*, Jeju Island, Korea, Oct. 2004.

Yu, F., Chang, E., Xu, Y. Q., & Shum, H. Y. (2001a). Emotion detection from speech to enrich multimedia content. In *Proc. IEEE Pacific Rim conference on multimedia*, Beijing (pp. 550–557).

Yu, F., Chang, E., Xu, Y. Q., & Shum, H. Y. (2001b). Emotion detection from speech to enrich multimedia content. In *Second IEEE Pacific-Rim conference on multimedia*, Beijing, China, Oct. 2001.

Yuan, J., Shen, L., & Chen, F. (2002). The acoustic realization of anger, fear, joy and sadness in Chinese. In *International conference on spoken language processing (ICSLP 02)*, Denver, Colorado, USA, Sept. 2002 (pp. 2025–2028).

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

Zhou, G., Hansen, J. H. L., & Kaiser, J. F. (2001). Nonlinear feature based classification of speech under stress. *IEEE Transactions on Audio, Speech, and Language Processing*, *9*, 201–216.

Zhou, Y., Sun, Y., Yang, L., & Yan, Y. (2009). Applying articulatory features to speech emotion recognition. In *International conference on research challenges in computer science, ICRCCS*, 28–29 Dec. 2009 (pp. 73–76).

Zhou, Y., Sun, Y., Zhang, J., & Yan, Y. (2009). Speech emotion recognition using both spectral and prosodic features. In *International conference on information engineering and computer science, ICIECS*, Wuhan, Dec. 19–20, 2009 (pp. 1–4). New York: IEEE Press.

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