Chapter 2
Speech Emotion Recognition: A Review

Abstract This chapter presents the literature related to the databases, features, pattern classifiers used for emotion recognition from speech. Different types of emotional databases such as simulated, elicited and natural are critically reviewed from the research point of view. Review of existing emotion recognition systems developed using excitation source, vocal tract system and prosodic features is briefly presented. Basic pattern classification models used for discriminating the emotions are discussed in brief. Finally, the chapter concludes with motivation and scope of the work presented in this book.

2.1 Introduction

Speakers mainly convey their intentions through non-verbal means such as emotions, in the conversation. In addition to the message conveyed through textual contents, the manner in which the words are spoken conveys essential non-linguistic information. The 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 book is confined to emotions or expressions related to speech. 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. Understanding the text alone is not sufficient to interpret the semantics of a spoken utterance. Therefore, it is important that speech systems should be able to process the non-linguistic information such as emotions, along with the message.

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 [43]. The purpose of the sophisticated speech systems should not be limited to mere message processing, rather it should understand the underlying intentions of the speaker by detecting their expressions in the speech [1, 44, 45]. In the recent past, processing a speech signal for recognizing

S.R. Krothapalli and S.G. Koolagudi, *Emotion Recognition using Speech Features*,
SpringerBriefs in Electrical and Computer Engineering,
DOI 10.1007/978-1-4614-5143-3_2,

the underlying emotions has 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.

This chapter provides a review of the literature on speech emotion recognition, in view of emotion-specific features extracted from different aspects of speech. The features are broadly classified into three categories namely, excitation source, vocal tract system and prosodic features. Review of some important existing emotional speech corpora is given in Sect. 2.2. Section 2.3 discusses the role of excitation source features for developing different speech systems. The research on recognizing emotions from speech using system features is analyzed in Sect. 2.4. Section 2.5 highlights some of the existing works on speech emotion recognition using prosodic features. Review of the classification models used for speech emotion recognition is briefly discussed in Sect. 2.6. The motivation of the book from the available literature is given in Sect. 2.7. The chapter concludes with Sect. 2.8, by providing the scope of the work.

2.2 Emotional Speech Corpora: A Review

For characterizing the emotions, either for synthesis or for recognition, a suitable emotional speech database is a necessary prerequisite [21]. The design and collection of emotional speech corpora mainly depends on the research goals. For example: a single speaker emotional speech corpus would be enough for the purpose of emotional speech synthesis, whereas recognizing emotions needs a database with multiple speakers and various styles of expressing the emotions. The survey presented in this section critically analyzes 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 [46].

- The scope of the emotion database both in terms of number of subjects contributing for recording and number of emotions to be recorded is to be decided properly.
- The decision about the nature of the speech 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 and general context.
- Labeling of soft emotions present in the speech databases is highly subjective.
- Size of the database used for speech emotion recognition plays an important role in deciding the properties such as scalability, generalizability, and reliability of the developed systems. Most of the existing emotional speech databases used for developing emotion systems are too small in size to capture the influence of speakers, gender, and language for characterizing the emotions [46].

In the literature, emotional speech databases are collected mainly using three different methods. (1) Actors are asked to portray the given emotion in the case of simulated databases. These are the most popular databases used in emotional speech research. (2) Elicited databases are not completely natural, but are recorded under simulated natural situations. (3) Naturalistic databases are recorded form the natural situations. The properties of some important emotional speech corpora being used for emotional speech research are briefly discussed in Table 2.1. From Table, it may be observed that there is a huge disparity among the databases, in terms of language, number of emotions, number of subjects, purpose of corpus collection and methods of database collection.

The set of emotional speech databases, given in Table 2.1, is dominated by the English language, followed by German and Chinese. Very few databases are collected in languages such as: Russian, Dutch, Slovenian, Swedish, Japanese and Spanish. There is no reported reference of an emotional speech database in any of the Indian languages. Among the emotional speech databases, given in Table 2.1, 24 speech corpora are collected for the purpose of recognition and 8 are collected with the intention of synthesis. Subjective listening tests confirm that the average emotion recognition rate in case of any database has not crossed beyond 80%. For full blown emotion subjective listening tests have shown more than 90% of recognition performance. Most of the automatic emotion recognition systems have achieved the recognition performance close to subjective listening tests. About 70% of databases contain only 4–5 basic emotions. Few emotional speech databases contain seven and eight emotions. Most of the existing databases rarely contain uncommon emotions like: antipathy, approval, attention, prohibition, etc. A majority of the databases contain clearly distinguishable emotions such as anger, sadness, happiness and neutral. Since actor based simulated database collection is a straight forward and comparatively easy process, more than half of the databases mentioned in Table 2.1 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 are no availability of standard, internationally approved database 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 [47]. 'HUMAINE', a group of researchers dedicated to speech emotion recognition, has started the *INTERSPEECH emotion challenge* since 2009, to facilitate- feature, classifier, and open performance comparison for non-prototypical spontaneous emotion recognition. In Indian context, some organizations such as the Linguistic Data Consortium for Indian Languages (LDCIL), Centre 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 confined themselves to collect speech corpora in different Indian languages in the context of speech recognition/synthesis and speaker recognition tasks only.

Table 2.1 Literature survey of speech databases used for emotion processing

Sl. no.	Emotions	Number of speakers	Type of database	Purpose and approach	References
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	[48]
02	Anger, disgust, fear, joy, neutral, sadness and surprise (07)	Eight actors (two per language) (two per language)	Simulated	Synthesis. Speech in four languages (English, Slovenian, Spanish, and French) is recorded	[49]
03	Anger, boredom, joy, and surprise (04)	51 children	Elicited	Recognition. Recorded at the university of Maribor, in German and English	[50]
04	Anger, fear, happiness, neutral, and sadness (05)	40 native speakers	Natural	Recognition. Two broad domains of emotions are proposed based on prosodic features	[51]
05	Different natural emotions	125 TV artists	Natural	Recognition. It is known as Belfast natural database and is used for several emotion processing applications	[52]
06	Anger, boredom, fear, happiness, neutral, and sadness (06)	Single actor	Simulated	Synthesis. F_0, duration and energy are modeled for synthesizing the emotions	[53]
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	[31]

German emotional speech corpora					
08	Depression and neutral (02)	Different native speakers	Elicited	Recognition	[54]
09	Negative and positive (02)	Customers and call attendants	Natural	Recognition. Call center conversations are recorded	[25]
10	Annoyance, shock and stress (03)	29 native speakers	Elicited	Recognition	[24]
11	Hot anger, cold anger, happiness, neutral, and sadness (05), 40 utterances per emotion are recorded.	29 native speakers	Elicited	Recognition. Dimensional analysis of emotions is performed using F0 parameters	[55]
12	Anger, fear, neutral, and sadness (04)	Different native speakers	Simulated	Recognition. Prosodic, spectral and verbal cues are used for emotion recognition	[56]
13	Five stress levels (05)	6 soldiers	Natural	Recognition	[57]
14	Two task load stress conditions and two 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	[58]
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	[59]

(continued)

Table 2.1 (continued)

Sl. no.	Emotions	Number of speakers	Type of database	Purpose and approach	References
Japanese emotional speech corpus					
16	Anger, happiness, neutral, sadness (04), 112 utterances per emotion are recorded.	Single actress	Simulated	Recognition. Speech prosody, vowel articulation and spectral energy distribution are used to analyze four emotions	[60]
17	Anger, Boredom, disgust, fear, joy, neutral, and sadness (07)	Ten actors	Simulated	Synthesis	[61]
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 h of effective emotional expressions of children are recorded	[62]
19	Anger, Boredom, disgust, fear, joy, neutral, and sadness (07)	Ten actors (5M+5F)	Simulated	Recognition. About 800 utterances are recorded using 10 neutral German sentences	[22]
20	Soft, modal, and loud (03)	Single actor	Simulated	Synthesis. Di-phone based approach is used for emotional speech synthesis	[63]
21	Anger, Boredom, disgust, and worry (04)	Six native speakers	Simulated	Recognition. Affective bursts and short emotional non-speech segments are analyzed for discriminating the emotions	[64]
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. Twelve 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	[65]

Chinese emotional speech corpora					
23	Antipathy, anger, fear, happiness, sadness, and surprise (06)	Two actors	Simulated	Recognition	[66]
24	Anger, disgust, fear, joy, sadness, and surprise (06), 60 Utterances per emotion per speaker are recorded	12 actors	Simulated	Recognition. Log frequency power coefficients are proposed for emotion recognition using HMMs	[67]
25	Anger, happiness, neutral, and sadness (04), 721 short utterances per emotion are recorded	Native TV actors	Simulated	Recognition	[68]
26	Anger, fear, joy, neutral and sadness (05), 288 sentences per emotion are recorded	Nine native speakers	Elicited	Recognition. Phonation, articulation and prosody are used to classify four emotions	[69]
Spanish emotional speech corpora					
27	Desire, disgust, fear, fury (anger), joy, sadness, and surprise (07)	Eight actors (4M+4F)	Simulated	Synthesis. Acoustic modeling of Spanish emotions is studied. Rules are used to identify significant behavior of emotional parameters	[70]
28	Anger, disgust, happiness, and sadness (04), 2,000 phones per emotion are considered	Single actor	Simulated	Synthesis. Pitch, tempo, and stress are used for emotion synthesis	[71]

(continued)

Table 2.1 (continued)

Sl. no.	Emotions	Number of speakers	Type of database	Purpose and approach	References
29	Anger, joy, and sadness (03)	Two native speakers	Simulated	Synthesis Concatenative synthesis approach is used	[72]
Russian emotional speech corpus					
30	Anger, fear, happiness, neutral, sadness, and surprise (06), ten 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)	[73]
Swedish emotional speech corpus					
31	Happiness and neutral (02)	Single native speaker	Simulated	Synthesis. Variations in articulatory parameters are used for uttering Swedish vowels in two emotions	[74]
Italian emotional speech corpus					
32	Anger, disgust, fear, joy, sadness, and surprise (06)	Single native speaker	Simulated	Synthesis	[75]

From the above mentioned survey, it is observed that there is a need of an emotional speech corpus in Indian languages, to support the research on speech emotion recognition and synthesis, in the context of Indian languages. The database may contain seven and eight common emotions. Though it is actor based simulated speech corpus, it should contain underlying emotions, rather than full blown emotions, so that the expression of emotions is more real and natural. The database has to contain sufficient variability in terms of number of speakers, gender, and sessions of recording. The text prompts used for recording are to be neutral in nature, so that linguistic contents do not influence the expressiveness of emotions.

2.3 Excitation Source Features: A Review

Speech features derived from the excitation source signal are known as source features. The 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 the speech signal, and then separating it by inverse filter formulation. The resulting signal is known as a *linear prediction residual*, and it contains mostly the information about the excitation source [76]. In this book, features derived from the LP residual are referred to as the excitation source or simply source features. The sub-segmental analysis of the speech signal is aimed at studying characteristics of the glottal pulse, open and closed phases of glottis, strength of the excitation and so on. The characteristics of glottal activity, specific to the emotions may be estimated using excitation source features. The LP residual signal and glottal volume velocity (GVV) signal are explored in the literature as the correlates of excitation source information [77]. In the literature, very few attempts have been made to explore the excitation source information for any of the speech tasks. The reasons may be

1. Popularity of spectral features
2. The excitation signal (LP residual) obtained from LP analysis is viewed mostly as an error signal [78] or unpredictable components of the predicted speech signal.
3. The LP residual basically contains the higher order relations, and capturing these higher order relations is not well known [79].

It may be difficult to parameterize the LP residual signal, however 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 [80], as the first and 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

Table 2.2 Literature review on use of excitation source information for various speech tasks

Sl. no.	Features	Purpose and approach	References
01	LP residual energy	Vowel and speaker recognition	[36]
02	LP residual	Instants of significant excitation are determined	[37]
03	Higher order relations among LP residual samples	Categorizing audio documents	[38]
04	LP residual	Speech enhancement in multi-speaker environment	[39]
05	LP residual	Characterizing loudness, lombard effect, speaking rate, and laughter segments	[86]
06	Glottal excitation signal	Analyzing the relation between emotional state of the speaker and glottal activity	[41]
07	Glottal excitation signal	To analyze emotion related disorders	[41]
08	Excitation source signal	To discriminate emotions in continuous speech	[42]

such as message, speaker, language, and emotion-specific information. Probably, 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 the LP residual signal is successfully used in [81], for speaker recognition. LP residual energy is used in [36], for vowel and speaker recognition. Cepstral features derived from the LP residual signal are used in [82], for capturing the speaker specific information. The combination of features derived from the LP residual and LP residual cepstrum has been used to minimize the equal error rate in case of speaker recognition [83]. By processing LP residual signal using Hilbert envelope and group delay function, the instants of significant excitation are accurately determined [37]. The higher order relations among the samples of LP residual are also used for categorizing different audio documents like: sports, news, cartoons, music in noisy and clean environments [38]. The instants of significant excitation obtained from the LP residual signal during the production of voiced speech are used to determine the relative delays between the speech segments of different speakers in a multi-speaker environment, and they are further used to enhance the speech of individual speakers [39]. The epoch (instants of glottal closure) properties of LP residual are exploited in [84], 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 detecting the laughter segments from the speech [40, 85, 86]. Table 2.2 shows some of the important achievements in speech research using excitation source information.

From the available literature, it is clear that the excitation source information can be used equally well to develop any speech 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. Hence, there is a scope for conducting the detailed and systematic study on excitation source information for characterizing the emotions.

2.4 Vocal Tract System 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 signals. The Fourier transform of a speech frame gives its short time spectrum. Features like formants, their bandwidths, spectral energy and slope may be observed from the spectrum. The cepstrum of a speech frame is obtained by taking the Fourier transform on the log magnitude spectrum [15]. MFCCs (Mel frequency cepstral coefficients) and 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. 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 [16]. 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 [21].

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. MFCC features are used in [87], to distinguish speech and non-speech (music) information. It has been observed that the lower order MFCC features carry phoneme information, whereas higher order features contain non-speech information. A combination of MFCCs, LPCCs, RASTA PLP coefficients and log frequency power coefficients (LFPCs) is proposed as the feature set, to classify anger, boredom, happiness, neutral and sadness emotions in Mandarin [88, 89]. Log frequency power coefficients (LFPC) are used to represent the emotion-specific information in [10], 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 [10, 90]. The MFCC features extracted from lower frequency components (20–300 Hz) of the 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.

Table 2.3 Literature review on emotion recognition using vocal tract system features

Speech emotion research using vocal tract system features			
Sl. no.	Features	Purpose and approach	References
01	MFCC features	Discrimination of speech and music. Higher order MFCCs contain more music specific information and a lower number of MFCCs contain more speech specific information	[87]
02	MFCCs, LPCCs RASTA PLP coefficients, log frequency power coefficients	Classification of four emotions in the Mandarin language. Anger, happiness, neutral and sadness emotions are considered in this study	[88, 89]
03	Combination of MFCCs and MFCC-low features	Emotion classification using Swedish and English emotional speech databases	[91]
04	MFCC features from consonant, stressed and unstressed vowels (class-level MFCCs)	Emotion classification on English LDC and Emo-DB databases	[92]
05	Spectral features obtained using Fourier and Chirp transformations	Modeling human emotional states under stress	[93]

MFCC-low features are reported to perform better than pitch features in the case of emotion recognition [91]. 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 than prosodic or utterance-level spectral features. The combination of class-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 [92]. In [93] 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 2.3. From the references mentioned in Table 2.3, it is observed that, in most of the cases, spectral features are extracted through a conventional block processing approach, wherein the entire speech signal is processed frame by frame, considering the frame size of 20 ms, and a shift of 10 ms.

2.5 Prosodic Features: A Review

Human beings impose duration, intonation, and intensity patterns on the sequence of sound units, while producing speech. The incorporation of these prosody constraints (duration, intonation, and intensity) makes human speech natural. Lack of prosody knowledge can easily be perceived from the speech. 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 such as: intonation and energy, that are normally used by human beings to perform various speech tasks including emotion recognition [94, 95]. In the literature, mainly, pitch, energy, duration and their derivatives are used as the acoustic correlates of prosodic features [96, 97]. 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 [95]. 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 perception level, speech sound waves enter the 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 the literature, prosodic features such as energy, duration, pitch and their derivatives are treated as good correlates of emotions [25, 34, 67, 98]. 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 [99, 100]. Some studies [100, 101] have also tried to measure the 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 [100]. The importance of prosodic contour trends in the context of different emotions is discussed in [102, 103]. 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 [104]. 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 [67]. 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 a K-nearest neighbor classifier [27]. 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 [1]. The complex relations between pitch, duration and energy parameters are exploited in [72] for detecting the speech emotions. Table 2.4 shows some of the other important and recent works on speech emotion recognition using prosodic features.

From the literature, it is observed that most of the speech emotion recognition studies are carried out using utterance level static (global) prosodic features [23, 34, 67, 72, 98, 112]. Very few attempts have explored the dynamic behavior of prosodic patterns (local) for analyzing speech emotions [104, 113]. Elementary prosodic analysis of speech utterances is carried out in [114], 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.

2.6 Classification Models

In the literature, several pattern classifiers are explored for developing speech systems like speech recognition, speaker recognition, emotion classification, speaker verification and so on. However, the justification for choosing a particular classifier to the specific speech task is not provided in many instances. Most of the times appropriately some classifiers are chosen. 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

Table 2.4 Literature review on emotion recognition using prosodic features

Speech emotion research using prosodic features			
Sl. no.	Features	Purpose and approach	References
01	Initially 86 prosodic features are used, later best 6 features are chosen from the list	Identification of emotions in the Basque language. Around 92% emotion recognition performance is achieved using GMMs	[105]
02	35 dimensional prosodic feature vectors including pitch, energy, and duration are used	Classification of seven emotions of the Berlin emotional speech corpus. Around 51% emotion recognition results are obtained for speaker independent cases using neural networks	[106]
03	Pitch and power based features are extracted from frame, syllable, and word levels	Recognizing emotions in Mandarin. Combination of features from frame, syllable and word level yielded 90% emotion recognition performance	[107]
04	Duration, energy, and pitch based features	Recognizing emotions in the Mandarin language. Sequential forward selection (SFS) is used to select best features from the pool of prosodic features. Emotion classification studies are conducted on a multi-speaker multi-lingual database. Modular neural networks are used as classifiers	[108]
05	Eight static prosodic features and voice quality features	Classification of six emotions (anger, anxiety, boredom, happiness, neutral, and sadness) from the Berlin emotional speech corpus. Speaker independent emotion classification is performed using Bayesian classifiers	[109]
06	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 algorithms	[110]
07	Prosody and voice quality based features	Classification of four emotions namely anger, joy, neutral, and sadness from the Mandarin language. Around 76% emotion recognition performance is reported using support vector machines (SVMs)	[111]

as applied to speech emotion recognition [115]. 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 a feature vector. If the input feature vector to the classifier is a real vector $\mathbf{x}$, then the output score is given by $y = f(\mathbf{w} \cdot \mathbf{x}) = f\left(\sum_j w_j x_j\right)$, where $\mathbf{w}$ 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 $\mathbf{w}$ 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.

A non-linear weighted combination of object characteristics is used to develop non-linear classifiers. During implementation, 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 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 and the second method works on discriminative properties 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. 2.1.

Use of classifiers mainly depends upon the nature of data. If the nature of data is known before, then deciding on the type of classifier would be an easier task. Linear

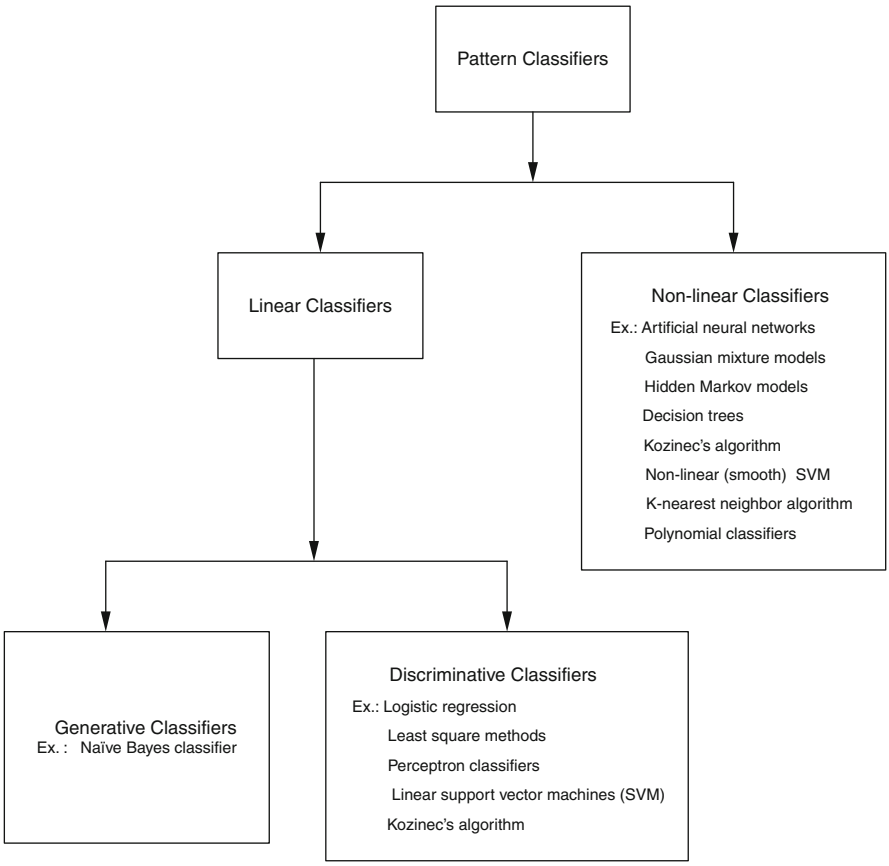

Fig. 2.1 Types of classifiers used for speech emotion recognition

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 not linearly separable would need non-linear classifiers for classification. In most of the real world situations, the nature of the data is rarely known. Therefore, researchers use non-linear classifiers always at the cost of complexity and computational time. Table 2.5 provides the list of classifiers used for speech emotion recognition. From Table 2.5, it may be observed that the majority of the speech emotion recognition tasks have employed non-linear classifiers. Artificial neural networks (ANN), Gaussian mixture models (GMM), and support vector machines (SVM) have been widely used in emotion recognition research. ANNs are known to capture non-linear relations among the feature vectors. GMMs are expected to capture the distribution of input feature space and probabilistically take the decision related to the class of the unknown feature vector. SVMs are discriminative classifiers. Their performance basically depends upon the number of feature vectors with discriminative properties, known as support vectors, rather

Table 2.5 Literature on use of different classifiers for speech emotion recognition task

Sl. no.	Classifiers	Features	References
01	Gaussian mixture models (GMM)	Prosodic	[26, 59, 91, 115, 116]
		Spectral	[26, 59, 87, 105, 115, 116]
02	Support vector machines (SVM)	Prosodic	[26, 105, 107, 110, 111, 116, 117]
		Spectral	[26, 107, 116, 117]
03	Artificial neural networks (ANN)	Prosodic	[26, 118–122], [28, 108, 115, 123]
		Spectral	[26, 28, 115, 118, 119, 123, 124]
04	k-Nearest neighbor classifier	Prosodic	[27, 88, 98, 115, 117]
		Spectral	[115, 117, 125, 126]
05	Bayes classifier	Prosodic	[98, 109, 110, 122]
		Spectral	[109]
06	Linear discriminant analysis with Gaussian probability distribution	Prosodic	[60, 104, 112]
		Spectral	[25, 60, 112, 126]
07	Hidden Markov models (HMM)	Prosodic	[67, 92, 122, 127]
		Spectral	[10, 67, 90, 127]

than the total number of input feature vectors. ANNs are mostly used to capture emotion-specific information present in the feature vectors in the form of non-linear higher order relations. Therefore ANN models are suitable for developing the emotion recognition systems using excitation source information. GMMs are used as classifiers, when the number of feature vectors is large enough to capture the proper distribution. For example in the case of spectral features GMMs are suitable for emotion recognition. Systems to be developed using frame-wise spectral features may perform better with GMMs. SVMs are used to develop emotion recognition models using prosodic features, where the number of feature vectors is less, and they have sufficient discriminative characteristics.

2.7 Motivation for the Present Work

The objective of this work is to develop a suitable simulated emotional speech database to promote the research on processing emotions from speech in the context of Indian languages and exploring various emotion-specific features of speech for characterizing the speech emotions. Excitation source, vocal tract system and prosodic information represent three different aspects of speech. Therefore emotion-specific features can be explored in these three broad categories. From the existing work presented in Sect. 2.3, it is observed that no systematic study is carried out on speech emotion recognition using excitation source features. However, excitation source features have been used successfully for some speech tasks [36–38, 42, 86]. Therefore, in this work we are exploring different excitation source features for recognizing speech emotions.

From the literature related to emotion recognition using system features, it is observed that spectral features are mostly extracted through a conventional block processing approach. In this approach feature vectors are extracted from the entire speech signal using overlapped frames. High amplitude regions of spectrum are robust in the case of speech with background disturbances. Therefore, formant features extracted from speech are explored for emotion recognition in combination with other spectral features.

Considering the literature provided in Sect. 2.5, one can observe that most of the existing works on prosodic features mainly exploited their gross statistics at the utterance level such as maximum, minimum, mean, and standard deviation of the feature set for recognizing the emotions. However the variations in prosodic features with respect to time are not explored. In reality, the dynamics of prosodic parameters (i.e., local or fine variations in prosodic parameters with respect to time) are crucial in analyzing the manifestation of the emotions at the suprasegmental level. With this motivation, dynamics of prosodic features are explored along with static features for characterizing the emotions.

2.8 Summary of the Literature and Scope for the Present Work

From the literature, it is observed that there is no proper emotional speech corpus in any of the Indian languages for carrying out the research on emotional speech processing. A real life emotional speech database is also not available in the context of Indian languages. It is also observed from the literature that excitation source information is not thoroughly investigated for the purpose of emotion recognition task. Most of the researchers have used frame-wise spectral features extracted from entire utterance for speech emotion classification. Most of the existing emotion recognition systems are developed using only gross prosodic features extracted from the entire utterances. Dynamics of prosodic patterns and emotion-specific prosody are not explored in view of recognizing the emotions. From this summary of the available work, the scope of this book may be outlined as follows.

- Design and collection of a simulated Telugu emotional speech database from the artists.
- The excitation source features such as LP residual signal, LP residual phase, epoch parameters such as strength of epochs, instantaneous frequency, sharpness of epochs, slope of strength of epochs, glottal volume velocity (GVV) waveform, GVV parameters, dynamics of epoch parameters at syllable and utterance levels, may be used as features for speech emotion recognition.
- Exploring basic spectral features such as LPCCs and MFCCs, extracted from entire speech utterances through block processing approach for recognizing the emotions.

- Static (global) and dynamic (local) features of prosodic contours may be explored for emotion classification.
- Different linear and non-linear pattern classifiers may be explored to study their suitability for emotion classification tasks. AANNs, GMMs, and SVMs are identified as suitable models for developing emotion recognition systems, for classifying the emotions using excitation source, spectral, and prosodic features respectively.

http://www.springer.com/978-1-4614-5143-3

Emotion Recognition using Speech Features
Rao, K.S.; Koolagudi, S.G.
2013, XII, 124 p. 30 illus., 6 illus. in color.,
ISBN: 978-1-4614-5143-3