Analysis of Emotional Speech—A Review

Chapter · March 2016

DOI: 10.1007/978-3-319-31056-5_11

3 authors, including:

Gangamohan Paidi
International Institute of Information Techno…
14 PUBLICATIONS **39** CITATIONS
SEE PROFILE

Sudarsana Reddy Kadiri
International Institute of Information Techno…
15 PUBLICATIONS **37** CITATIONS
SEE PROFILE

Chapter 11
Analysis of Emotional Speech—A Review

P. Gangamohan, Sudarsana Reddy Kadiri and B. Yegnanarayana

Abstract Speech carries information not only about the lexical content, but also about the age, gender, signature and emotional state of the speaker. Speech in different emotional states is accompanied by distinct changes in the production mechanism. In this chapter, we present a review of analysis methods used for emotional speech. In particular, we focus on the issues in data collection, feature representations and development of automatic emotion recognition systems. The significance of the excitation source component of speech production in emotional states is examined in detail. The derived excitation source features are shown to carry the emotion correlates.

11.1 Introduction

Humans have evolved various forms of communication like facial expressions, gestures, body postures, speech, etc. The form of communication depends on the context of interaction, and is often accompanied by various physiological reactions such as changes in the heart rate, skin resistance, temperature, muscle activity and blood pressure. All forms of human communication carry information at two levels, the message and the underlying emotional state.

Emotions are essential part of real life communication among human beings. Various descriptions of the term emotion are studied in [21, 22, 60, 88, 92, 98, 100]. Some of the descriptions are:

(a) "Emotions are underlying states which are evolved and adaptive. Emotion expressions are produced by the communicative value of underlying states" [22].

P. Gangamohan (✉) · S.R. Kadiri · B. Yegnanarayana
International Institute of Information Technology, Hyderabad, India
e-mail: gangamohan.p@students.iiit.ac.in

S.R. Kadiri
e-mail: sudarsanareddy.kadiri@research.iiit.ac.in

B. Yegnanarayana
e-mail: yegna@iiit.ac.in

A. Esposito and L.C. Jain (eds.), *Toward Robotic Socially Believable Behaving Systems - Volume I*, Intelligent Systems Reference Library 105,
DOI 10.1007/978-3-319-31056-5_11

(b) "Emotions are experienced when something unexpected happens at times," [92].
(c) "Emotions are an organism's interface to outside world," and carry three principle functions, *the significance and relevance of particular stimuli*, *preparation of the organism's physiology for appropriate action*, and *communication of the organism's state and intention to other organisms* [98].

Many studies reported the relationship between emotions and expressive states. Darwin [22] defined emotion as an inner state and expression as a communicative manifestation of the underlying emotional state. Studies in the literature [15, 88, 92, 99, 119] have referred to the concept of "basic emotions". Typically, "basic emotions" are perceived cross-cultural, while non-basic emotions are learnt in a culture-specific manner [15]. Although the concept of "basic emotions" has been widely accepted, there has been considerable debate on the composition and number of "basic emotions" [27, 115]. Typically, anger, happiness, fear, sadness and neutral are identified as "basic emotions".

In speech communication, humans have the natural ability to grasp the underlying emotional state, as well as the lexical content. A fundamental research problem is, given a speech signal, can the underlying emotional state of the speaker be identified? Normally, human beings perceive emotions of an unknown speaker through deviations from normal state. For example, in the perception of angry speech, there is increase in voice intensity, raise in pitch and faster speaking rate. From this, one can infer that there is reference (neutral/normal), and deviations from the reference are being perceived. The objective is to derive an emotion-specific feature representation, by focusing on the deviations in the components of the speech production mechanism.

Progress in human-computer voice interaction systems is limited due to difficulty of the machine in recognizing and responding to even basic emotions. Whereas, it happens effortlessly in human-human communication. Systems recognizing speaker's emotions, and also responding expressively, are essential for natural interaction. Research on emotions in speech has applications in spoken dialogue systems, automated response systems, call centers, etc.

This chapter gives a review of emotional speech analysis, along with description of studies addressing specific research issues in emotion analysis. The organization of the chapter is as follows. The data collection issues for emotion related studies are discussed in Sect. 11.2. A review of studies on emotion analysis and emotion recognition is given in Sect. 11.3. Studies addressing some specific issues are presented in Sect. 11.4. Finally, Sect. 11.5 discusses some research challenges in emotion analysis.

11.2 Data Collection of Emotional Speech

Progress in applications related to emotional speech relies heavily on the availability of suitable databases [44]. Emotion databases developed by different research groups can be categorized into simulated, semi-natural and natural databases [26,

73, 112, 126]. Simulated parallel emotion databases are collected from speakers (voice-talents) by prompting them to enact emotions through specified text in a given language. The simulated parallel emotion databases in [17, 28, 74, 113, 138] were collected from speakers by asking them to emote the same text in different emotions. The main disadvantage of such databases is that deliberately enacted emotions are quite at variance from 'spontaneous' emotions, and also at times they are out of context [30, 126]. Semi-natural is a kind of enacted corpus, where the context is given to the speakers. The semi-natural emotion database in German language was developed by asking speakers to enact the scripted scenarios eliciting each emotion [13, 116]. The third kind of emotion database is a natural database, where the recordings do not involve any prompting or eliciting of emotional responses. Sources for such natural situations could be talk shows, interviews, panel discussions and group interactions in TV broadcast. A brief overview of databases is given in Table 11.1.

For developing high quality expressive text-to-speech (TTS) synthesis systems, a large natural database of each target emotion is required [103]. But it is impractical to obtain a large natural emotion database. Emotion conversion systems are adopted as a post-processing block for speech synthesis systems. In this, a large database of neutral speech is used to generate speech by a TTS system, which is then fed to an emotion conversion system [88, 103]. The output speech from these systems is unnatural, and also has the constraint of requiring parallel speech corpus [31].

As the collection of natural databases is mostly carried out from TV talk shows, call centers and interaction with robots, speakers involved in these cases control their emotion/expressive states. There is a trade-off between the controllability and naturalness of the interaction [51]. There is also difficulty in identifying the emotion or expressive state of a dialogue. Therefore, the annotation is described mostly by three basic primitives or dimensions, namely, valence, arousal/activation and dominance/power [51, 125]. Also, there are two other cases of ambiguity in annotating. One of them is occurrence of mixed emotions in an utterance. For example, there is a possibility of combinations like, surprise-happy, frustration-anger and anger-sad, occurring in the dialogue. The second reason for ambiguity is due to unsustainabililty of emotion throughout the dialogue. In natural communication among human beings, emotion may not be sustainable over the entire duration of the dialogue. The emotion is expressed mostly in some segments of the dialogue, with the rest of the dialogue being neutral. It is also difficult to define the boundaries for the occurrence of an emotion in continuous speech [64].

11.3 Emotional Speech Analysis

Feature representation plays a key role in developing any emotion related applications. The features used in many analysis studies can be broadly categorized into prosody, voice quality and spectral features. In this Section, the overview of these features is given along with the literature of emotion recognition studies.

Table 11.1 An overview of emotional speech data collection

	Simulated	Semi-natural	Natural
Description	Collected from trained speakers by asking them to emote same text in different emotions	Developed by asking speakers to enact the scripted scenarios eliciting each emotion	• Recording does not involve any prompting or the obvious elicitation of emotional responses
			• Sources: TV broadcast, call center calls, court rooms, etc.
Examples	• LDC [138]	• IEMOCAP [18]	• VAM [51]
	• EMO-DB [17]	• Belfast [116]	• Call centers [77, 87]
	• DES [28]	• NIMITEK [44]	• AIBO [117]
	• IITKGP-SESC [74]		
Advantages	• Most commonly used and standardized	• Near to natural database	• Completely natural
	• Comparison of results is easy	• Even though contextual information is available, it is still artificial	• Useful for real world emotion systems modeling
	• Number of emotions available are large		
Disadvantages	• It tells how emotions should be portrayed rather than how they are portrayed	• Less number of emotions are available	• Emotions are continuous
	• Context, environment and purpose dependent information is absent	• If speakers are aware that they are being recorded, the emotions become unnatural/artificial	• All emotions are not available
	• Episodic in nature, not true in real-world situations		• Contains multiple and concurrent emotions
			• Difficult to model
			• Copyright and privacy issues arise
Preferred applications	• Emotion conversion systems	• Emotion recognition systems	• Emotion recognition systems
	• Expressive speech synthesis systems	• Dimensional emotion analysis	• Dimensional emotion analysis
		• Categorical emotion analysis	• Categorical emotion analysis

Table 11.2 Summary of prosody and spectral features of emotional speech w.r.t. neutral speech

	Angry	Happy	Sad	Fear
Prosody features				
Average F_0	Higher	Higher	Lower	Much higher
F_0 range	Wider	Wider	–	Wider
F_0 contour (shape)	Irregular fluctuations	Descending and ascending patterns at irregular intervals	Downward inflections	Much irregular up-down fluctuations
Average intensity	Higher	Higher	Lower	Normal
Intensity range	Wider	Wider	Narrower	Normal
Speaking rate	Slightly faster	Normal	Slightly slower	Faster
Spectral features				
F_1 mean	Increase	–	Increase	Increase
F_1 bandwidth	–	Increase	Decrease	Decrease
F_2 mean	Increase	–	Decrease	–
High frequency energy	Increase	Increase	Decrease	Increase

11.3.1 Prosody Features

Efforts have been made to understand the contribution of speech prosody features towards production of emotion. Early studies were based on the fundamental frequency (F_0) of emotional speech [23, 36, 37, 40, 132, 133]. Along with F_0, various aspects of prosody like speaking rate, relative durations and intensity were examined [24, 50, 87]. One of the early studies using speaking rate parameter was by [36], where the speaking rate was described in terms of words per minute. The analysis of pauses in angry, happy, fear and sad speech was also performed in the study. The related features of speaking rate like the average duration of voiced speech, ratio of voiced to unvoiced speech, average duration of pauses and syllables per second, have been used in [7, 23, 24, 50, 58, 87]. A summary of the prosody features corresponding to basic emotions is given in Table 11.2. Some of the features share similar properties across different emotions. As observed from Table 11.2, angry and happy utterances have similar trends in F_0 and speaking rate, with respect to (w.r.t.) neutral speech.

11.3.2 Voice Quality Features

In a broader sense, the term voice quality refers to the characteristic auditory colouring of an individual's speech [67]. In general, speakers have their own voice quality signature. By varying voice qualities, they convey important information like inten-

tions, emotions and attitudes. From the perspective of Laver's approach [75], voice quality is expressed in terms of laryngeal and supralaryngeal settings. Laryngeal settings are described by phonation types, pitch and loudness ranges. Supralaryngeal settings are described by longitudinal, latitudinal, tension modifications of vocal tract, and nasalisation.

In the literature, many studies performed voice quality analysis by considering the laryngeal activity (mainly phonations). Non modal phonations are often observed in emotional speech. Breathiness is associated with angry and happy speech [71, 88]. Vocal fry voice is observed in sad and relaxed speech [48, 71], which may be because of very low fundamental frequencies in these cases. Harsh voice, which corresponds to irregularity in voicing, was observed in fear speech [71].

Studies [4, 20, 46, 47] have shown that glottal source parameters like closed quotient (CQ), abruptness in closing and normalized amplitude quotient (NAQ) are useful in distinguishing different phonations. Also, these glottal source parameters are analyzed for emotional speech [1, 121, 122, 131]. These features were extracted from the glottal waveform derived using inverse filtering (IF) technique [3, 96]. There are several limitations in IF based approaches such as deriving the accurate transfer function by canceling out the effect of the vocal tract system, and obtaining the closed phase duration of the glottal cycle [38, 45]. Although these glottal source parameters give emotion correlates, the dynamic ranges of these features are observed to be speaker-specific [1, 48].

11.3.3 Spectral Features

The spectrum characterized by formant frequencies and their respective bandwidths is extensively analyzed for emotional speech [13, 100, 133]. In [133], it is observed that the vowels in angry speech are produced with wide open vocal tract, and inferred that the first formant (F_1) has higher mean than that of neutral speech. The predictions of formants (F_1, F_2 and F_3), their bandwidths and high frequency energy for the emotion classes are made in [13], which are given in Table 11.2 along with prosody features.

It is also interesting to note that there are certain changes in the spectral component which are associated with the glottal source excitation [54, 86, 133]. The syllables produced with higher fundamental frequencies in angry speech tend to have weaker F_1 amplitudes [133]. More closed phase of glottis configuration results in relatively higher amplitudes at high frequencies [86].

A fundamental characteristic of the spectrum of a speech signal is that it is sound-specific [72, 83]. The deviations in spectral features are analyzed for utterances having the same lexical content [87, 100, 133]. Also, some studies on spectral features [53, 78, 128] have shown that changes in the magnitude and shift of the formants in emotional states vary across vowels.

11.3.4 Emotions as Points in Continuous Dimensional Space

Viewing emotions as points in continuous dimensional space was first suggested in [101]. Emotions are mainly viewed as combinations of three dimensions/primitives, namely, valence, arousal/activion and dominance/power [101, 102, 104]. Several studies [12, 23, 93, 105] have explored the relation of features like F_0, durations, loudness and spectral parameters, to these dimensions. Some important findings of these studies are that high arousal speech (like angry and happy) is associated with increase in the average F_0, wider F_0 range and decrease in spectral tilt. Low arousal speech (like sad and disgust) is associated with decrease in the average F_0 and narrow F_0 range.

11.3.5 Studies on Emotion Recognition

Feature representation is the most important step for developing an automatic emotion recognition system. The prosody-related features are statistical measures of F_0 and intensity contours, and features related to speaking rate [29, 77, 80, 82, 107, 137]. The spectral features are mel frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs), modulation spectral features, formant frequencies, bandwidths of formant frequencies [55, 61, 77, 107, 134]. The voice quality features are shimmer, jitter and NAQ, which are related to the glottal excitation characteristics [32, 34, 49, 95, 97, 120, 121, 123].

Feature representation is done at different levels of speech such as frame level [56, 59, 68, 80, 81, 91], segment level [25, 61, 84, 90] and utterance level [19, 29, 55, 76, 82, 85, 97, 127, 137]. A common approach adopted for feature representation at segment and utterance levels is by statistical analysis of frame level features and prosody features. Some of the toolkits which are widely used for feature extraction are PRAAT (prosody and voice quality features) [16], APARAT (voice quality features) [2], OpenSMILE (prosody, voice quality and spectral features) [35] and OpenEAR (prosody and spectral features) [33].

There are two major approaches used in the automatic emotion recognition task. In the first approach, the utterances in a corpus are labeled with discrete emotion categories such as anger, happiness, sadness and fear. Feature representations at different levels are used for training models like (GMMs) [29, 61, 69, 81], support vector machines (SVMs) [55, 97, 137], artificial neural networks (ANNs) [6], deep neural networks (DNNs) [5], hidden Markov models (HMMs) [59, 80, 91]. The second approach is a primitive-based classification, where emotions are described by arousal, valence and dominance [32, 39, 50, 76, 82, 123], and the classification is done using a hierarchical binary tree approach. For example, the activation states are detected initially, and then the identification of valence states follows.

Most of the pattern recognition algorithms used for developing emotion recognition systems require large amount of labeled emotion data. Also, emotion recognition systems that use spectral feature representations require a phonetically/phonemically

balanced database. In [32, 82], it is reported that prosody features can effectively discriminate activation states, but there is difficulty in discriminating valence states.

There are a few emotion recognition systems developed using linguistic features [14, 25, 29, 68, 77, 94, 97, 106, 108–111, 118]. Typically there are two approaches, one is a language modeling approach [25, 29, 68, 77, 94] and the other is a bag-of-words approach [14, 97, 106, 108, 111, 118]. In language modeling approach, utterances are transcribed manually, then language models are developed for each emotion. In the testing phase, the decision of emotion is given by the likelihood scores of each model for the test utterance. The bag-of-words approach was primarily developed for document retrieval tasks [63, 106]. In this approach each word in the vocabulary adds a dimension to the linguistic vector. The linguistic vector which gives the word term frequency within the utterance is used for classification of emotions. The limitations of these approaches are that the data required for each emotion should be large, and also these methods are based on ASR framework with its own limitations.

11.3.6 Limitations of the Studies

The prosody and voice quality features are mostly speaker-specific, and the spectral features are mostly sound-specific. In order to use these features for developing emotion recognition systems, a phonetically/phonemically balanced database covering several speakers might be required.

It is also observed that there are many interrelations among the type of database used, features, approaches and evaluation procedures. Some of the emotion recognition studies with various features, databases, pattern recognition algorithms and recognition accuracies are given in Table 11.3.

The following are some important observations from Table 11.3.

- The performance in terms of accuracy of emotion recognition systems built and tested using simulated parallel corpus [50, 61, 82, 120, 137] is high when compared to systems built and tested with semi-natural or natural databases [11, 32, 62, 76, 97]. As per spectral analysis of emotions [78, 87, 100, 128], for speech segments of the same lexical content, there exist deviations in the formant frequencies, bandwidths and spectral tilt, in somewhat emotion-specific way. These spectral deviations might help in discriminating emotions in the case of simulated parallel corpus.
- There are a few emotion recognition studies focused on cross-corpora and cross-lingual aspects [34, 62, 113, 120]. There are two ways of cross-corpora evaluations, a system developed using a database and tested with other database, or a mixture of cross-corpora used for both training and testing systems [113, 114, 120]. The reported recognition results are low in these cases.
- In several studies [6, 50, 61, 62, 69, 80, 82, 127], it is reported that there is confusion between anger and happiness emotions. The anger versus happiness confusion is also observed in the case of temporal modeling of intonation variations using HMMs [80].

Table 11.3 Review of some emotion recognition studies

References	Features	Database and no. of emotions	Pattern recognition algorithms	Accuracy and remarks
Jeon et al. [61]	Spectral and prosody	EMO-DB (simulated parallel), 4 emotions (anger, happiness, sadness and neutral)	GMM	85 % and more anger versus happiness confusion
Yeh and Chi [137]	Spectral and prosody	EMO-DB (simulated parallel), 7 emotions (anger, happiness, fear, disgust, boredom, sadness and neutral)	SVM	83 % and confusion matrix not reported
Sun and Moore II [120]	Spectral, prosody and voice quality	EPST, EMA and EMO-DB (all are simulated parallel), 4 emotions (anger, happiness, sadness and neutral)	SVM	Training and testing with same database, EPST—64 %, EMA—80 % and EMO-DB—83 %. Accuracy < 50 % for cross-corpora
Lugger and Yang [82]	Prosody and voice quality	EMO-DB (simulated parallel), 7 emotions	GMM	66 % and more anger versus happiness confusion
Grimm et al. [50]	Spectral and prosody	EMA (simulated parallel), 4 emotions (anger, happiness, sadness and neutral)	kNN(k-nearest neighbors)	83 % and more anger versus happiness confusion
Espinosa et al. [32]	Spectral, prosody and voice quality	VAM (natural), 3 categories (arousal, valence and dominance)	SVM	68 %
Rozgic et al. [97]	Spectral, prosody and lexical	USC-IEMOCAP (semi-natural), 4 emotions (anger, happiness, sadness and neutral)	SVM	69 %
Lee et al. [76]	Spectral, prosody and voice quality	AIBO (natural), 5 emotions (anger, emphatic, positive, neutral and others)	Bayesian logic regression	49 %

(continued)

Table 11.3 (continued)

References	Features	Database and no. of emotions	Pattern recognition algorithms	Accuracy and remarks
Atassi and Esposito [10]	Prosody and voice quality (emotion selective)	EMO-DB (simulated parallel) 6 emotions (anger, happiness, fear, disgust, boredom and sadness)	GMM	80 % and more anger versus happiness confusion
Ververidis and Kotropoulos [127]	Spectral and prosody	DES (simulated parallel), 5 emotions (anger, surprise, sadness, happiness and neutral)	GMM	More anger versus happiness confusion
Wu et al. [134]	Modulation spectral	VAM (natural), 3 categories (arousal, valence and dominance). EMO-DB (simulated parallel) 7 emotions	SVM	VAM—67 %, EMO-DB—85 %
Lin et al. [80]	Spectral and prosody	MHMC (semi-natural)	HMM	79 % and more anger versus happiness confusion
Atassi et al. [11]	Spectral, prosody and voice quality	COST 2102 Italian (natural) database, 6 emotions (anger, fear, happiness, irony, sadness and surprise)	ANNs	60.7 %

- There are a few studies on emotion recognition which addressed speaker dependent and speaker independent criteria [50, 70, 82, 85]. In these studies, the recognition accuracies of speaker independent systems are reported to be low when compared to the accuracies of speaker dependent systems. This may be because of the speaker-specific variations in the voice quality and prosodic features.

11.4 Studies on Some Specific Research Issues

The following are the specific research issues addressed in some recent studies

- *Analysis-by-synthesis to explore relative contribution of different components of emotional speech*: These experiments are performed using a flexible analysis-synthesis tool (FAST) [41]. This tool is used to modify the components of speech of the source utterance to that of the target, and vice-versa. These analysis-by-synthesis experiments are discussed in Sect. 11.4.1.
- *Identification of emotion-specific regions of speech*: Speaker-specific neutral versus non neutral regions are detected from speech signals using neural networks [65], and the studies are described in Sect. 11.4.2.
- *Significance of excitation source features in emotional speech*: The excitation source features extracted around the glottal closure instants (GCIs) are analyzed [42]. These studies are described in Sect. 11.4.3. A speaker-specific emotion recognition system developed based on these features [66] is described in Sect. 11.4.4.
- *Anger and happiness discrimination*: Emotion recognition studies indicate that there exists confusion among higher activation states like anger and happiness. Features related to the excitation source of speech are examined for discriminating anger and happiness emotions [43], and the studies are described in Sect. 11.4.5.
- *Discrimination between high arousal and falsetto voices*: Speakers tend to increase pitch when they raise their voice intensity in natural communication. Speakers can also shift to a falsetto register, where the pitch can be deliberately increased without raising the voice intensity. The excitation source related features are analyzed for these two cases in Sect. 11.4.6.

11.4.1 Analysis-by-Synthesis to Explore Relative Contribution of Different Components of Emotional Speech

The objective of this study is to determine the components of speech that contribute to the perception of emotion. A controlled set of experiments are conducted to modify the speech signal of the same sentence in neutral and emotional states. This is accomplished by using a flexible analysis-synthesis tool (FAST) described in [41].

Analysis of the relative contribution of F_0 and the amplitude of the speech signal is given in [79]. Similarly, experiments investigating the role of F_0 contour and duration

parameter are described in [129]. A time-domain pitch synchronous overlap and add (PSOLA) algorithm is used for modification of these components. The general observation is that the modification of parameters individually in neutral speech does not give the perception of emotion.

In addition to the prosody-related components, components highlighting the characteristics of the excitation source and the vocal tract system are also considered in this section. The following five components of speech are used: Vocal tract shape, excitation source, excitation energy, F_0 contour and relative durations. The time varying vocal tract information is represented by linear prediction coefficients (LPCs) for each frame of 20 ms with a frame shift of 10 ms. The excitation source information is represented by the linear prediction (LP) residual signal within each glottal cycle. The glottal cycle relates to a signal between two adjacent epochs (or GCIs). These epochs are extracted using the zero frequency filtering (ZFF) method [89]. In the zero frequency filtered signal, the negative to positive zero crossings corresponds to the GCIs. The F_0 contour is obtained from the intervals between epochs. The energy of the LP residual within each glottal cycle gives the excitation energy. The set of experiments (denoted as $E1$ to $E31$ in Table 11.4) are performed to modify the components of the source utterance to that of the target utterance.

The IIT-KGP SESC database in Telugu language [74] is used in this study. From this database, utterances of 5 emotions (neutral, anger, happiness, sadness and fear) of the same sentence of two speakers are considered. Two sets of studies are made: Set-1 consists of modification of neutral to emotion, and set-2 consists of modification of emotion to neutral. The modified speech is subjected for evaluation. The subjective evaluation is carried out by 10 (student) listeners from the Speech and Vision Laboratory at IIIT Hyderabad. Each subject was asked to give a similarity score ranging from 1 to 5 for a pair of utterances. The score 5 indicates that the utterances have high similarity. The score 1 indicates that both the utterances are very much different. The original target utterance and synthesized speech of each emotion category are used for determining the similarity.

The experiments with average similarity scores greater than 3.0 for different emotion categories in both the sets are given in Table 11.4. It is interesting to note that the entries are almost same in both the sets, indicating that the parameters for modifying neutral to emotion and vice-versa gives similar perception of target utterance. It is interesting to note that modification of any one component is not adequate for creating or suppressing the characteristics of any emotion category. Experiments which include modification of F_0, duration and LPCs components seem to be having high scores in the case of anger category. Angry speech is produced under extreme displeasure and frustration, it exhibits very high and wider F_0 when compared to other emotions [88]. Also, it is observed that there is wide opening of vocal tract during speech in anger state [133]. Therefore, combination of these parameters might give a better perception of anger. For the perception of sadness, happiness and fear emotions, several combination of the components are possible.

Table 11.4 The analysis-by-synthesis experiments with average similarity scores greater than 3.0 for two sets

Exp.	Components for modification	Set-1 (neutral to emotion)	Set-2 (emotion to neutral)
*E*1	F_0	–	–
*E*2	Duration (dur)	–	–
*E*3	LPCs	–	–
*E*4	Excitation (exc) energy	–	–
*E*5	Excitation (exc) source	–	–
*E*6	F_0 and dur	AN(3.2)	–
*E*7	F_0 and LPCs	HA(3.1)	HA(3.1)
*E*8	F_0 and exc energy	–	–
*E*9	F_0 and exc source	SA(3.1)	
*E*10	Dur and LPCs	–	–
*E*11	Dur and exc energy	–	–
*E*12	Dur and exc source	–	–
*E*13	LPCs and exc energy	–	–
*E*14	LPCs and exc source	–	–
*E*15	Exc energy and exc source	–	–
*E*16	F_0, dur and LPCs	AN(3.7), HA(3.7), SA(3.1), FE(3.7)	AN(3.5), HA(3.2), SA(3.1), FE(3.6)
*E*17	F_0, dur and exc energy	FE(3.6)	FE(3.2)
*E*18	F_0, dur and exc source	SA(3.0), FE(3.5)	SA(3.1)
*E*19	F_0, exc energy and LPCs	HA(3.9)	
*E*20	F_0, exc source and LPCs	HA(3.0), SA(3.7)	SA(3.6)
*E*21	F_0, exc energy and exc source		SA(3.0)
*E*22	Dur, exc energy and LPCs	–	–
*E*23	Dur, exc source and LPCs	–	–
*E*24	Dur, exc energy and exc source	–	–
*E*25	Exc energy, exc source and LPCs	–	–
*E*26	F_0, dur, exc energy and exc source	AN(3.9), HA(3.4), SA(3.4), FE(3.7)	AN(3.9), HA(3.5), SA(3.4), FE(3.6)
*E*27	F_0, dur, exc energy and LPCs	AN(4.3), HA(3.8), SA(3.4), FE(3.9)	AN(4.0), HA(3.6), SA(3.5), FE(3.5)
*E*28	F_0, dur, exc source and LPCs	AN(4.1), HA(3.3), SA(3.6), FE(3.7)	AN(3.7), HA(3.5), SA(3.7), FE(3.6)
*E*29	F_0, exc energy, exc source and LPCs	SA(3.7)	HA(3.3), SA(3.8)
*E*30	Dur, exc energy, exc source and LPCs –	–	
*E*31	F_0, Dur, exc energy, exc source and LPCs	AN(4.4), HA(4.5), SA(4.3), FE(4.2)	AN(4.4), HA(4.1), SA(4.3), FE(4.5)

(*AN*—anger, *HA*—happiness, *SA*—sadness, and *FE*—fear)

Table 11.5 Neutral versus non neutral discrimination for EMO-DB and IIIT-H Telugu emotion databases [65]

	EMO-DB database (%)	IIIT-H Telugu database (%)
Excitation source	91.03	94.01
Vocal tract system	83.25	88.23

11.4.2 Identification of Emotion-Specific Regions of Speech—Neutral Versus Non Neutral Speech

In this section, the issue of (speaker-specific) neutral versus non neutral speech detection is discussed using models of neutral speech as reference. Autoassociative neural network (AANN) models are developed for capturing the excitation source and the vocal tract system components separately. For this purpose, LP residual and LPCs are used as approximations of the excitation source and vocal tract system components, respectively. A 10th order LP analysis with a frame length of approximately twice the instantaneous pitch period of the speech signal is chosen. For extracting the excitation source information, a 4 ms segment of the LP residual around each GCI is chosen. The vocal tract system characteristics are represented by the 15 dimensional weighted LPCC vector derived from the LPCs.

The network structure 33*L* 80*N* *xN* 80*N* 33*L* is chosen for developing the model for the excitation source component of the neutral speech. Here *L* refers to linear units, *N* refers to nonlinear (*tanh*()) output function of units, and *x* refers to the number of units in the compression layer. The structure 15*L* 40*N* *xN* 40*N* 15*L* is used for developing the model for the vocal tract system component of neutral speech. For both the AANN models, a universal background model (UBM) is developed using 15 s of neutral speech from each speaker. Speaker-specific AANN models for neutral speech are developed by training over the UBM using approximately 20 s of neutral speech data from a speaker. In the testing phase, the emotional speech utterance is presented to the neutral speech AANN models, and the mean squared error between the output and input is normalized with the magnitude of the input. Since the AANN models are developed using neutral speech, it is expected that the error should provide discrimination between neutral and emotional speech. It is observed that, error values are high when the test utterance is not neutral, and low when the test utterance is neutral. Using a threshold on the averaged normalized error value, the emotion regions can be detected.

The results of neutral versus non neutral detection using the ANN models for EMO-DB (in German language) and IIIT-H Telugu emotion database [42] are shown in the Table 11.5. From the Table 11.5, it appears that the excitation source information provides better discrimination of neutral versus non neutral states than the vocal tract system information. It is observed that the proposed excitation source and vocal tract system AANN models provide an improvement of approximately 10 and 3 %, respectively, over the recently proposed method [8, 9] (accuracy is 80.4 %)

for EMO-DB. It is also observed that the high arousal emotion states (like anger and happiness) are more discriminative compared to the low arousal emotion states (sadness and boredom). This is in conformity with the studies reported in [61, 76, 114]. It is to be noted that emotion information may not be uniformly distributed across all frames in time. It is also necessary to explore methods to combine the evidence from the excitation source and from the vocal tract system characteristics.

11.4.3 Significance of Excitation Source Features in Emotional Speech

This study demonstrates the significance of the excitation source features. The excitation features are extracted around the epoch locations. The following four features considered for this study: Instantaneous F_0, strength of excitation (*SoE*), energy of excitation (*EoE*) and loudness parameter (η). The instantaneous F_0 and *SoE* features are extracted using the ZFF method [89]. The slope of the zero frequency filtered signal at each epoch location is called *SoE*. The *SoE* parameter is related to the strength of the impulse-like excitation at the epoch [136]. The *EoE* feature is computed using the energy of the samples of the Hilbert envelope (HE) of the LP residual over 2 ms around each epoch. The loudness measure η gives the abruptness of the glottal closure [52]. The η feature is given by the ratio of the standard deviation and mean of the samples of the HE of the LP residual over 2 ms around each epoch location.

The deviations of the excitation source features of an emotional speech w.r.t. neutral speech are analyzed in the following six 2-dimensional (2-D) feature spaces: F_0 versus *SoE*, F_0 versus *EoE*, F_0 versus η, *SoE* versus *EoE*, η versus *SoE* and η versus *EoE*. Sample 2-D scatter plots for a pair of neutral and angry utterances are shown in Fig. 11.1.

The Kullback-Leibler (KL) distance [57] measure of the distributions in the 2-D feature spaces of neutral (reference) and emotion (test) utterances is used for representing the deviations. The KL distance measure is given by

$$D = \frac{1}{2}\left(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1}(\mu_1 - \mu_0) - k - ln\left(\frac{det\,\Sigma_0}{det\,\Sigma_1}\right)\right), \quad (11.1)$$

where D is the KL distance, k is the dimension of the features space, and Σ_0, Σ_1 are the covariance matrices, and μ_0, μ_1 are the mean vectors of the neutral and emotion utterances, respectively.

Three databases, namely, IIIT-H Telugu, IITKGP-SESC and EMO-DB are used in this study. In the case of IIIT-H Telugu (semi-natural) database, 3 test utterances for each emotion (anger, happiness, sadness and neutral) and 2 neutral (reference) utterances for each speaker are used. The KL distance averaged over different test and reference utterances are given for two speakers (1 male and 1 female) in Table 11.6. From Table 11.6, the average KL distance values are low when the reference and test utterances are both neutral. The average KL distance values are higher when the

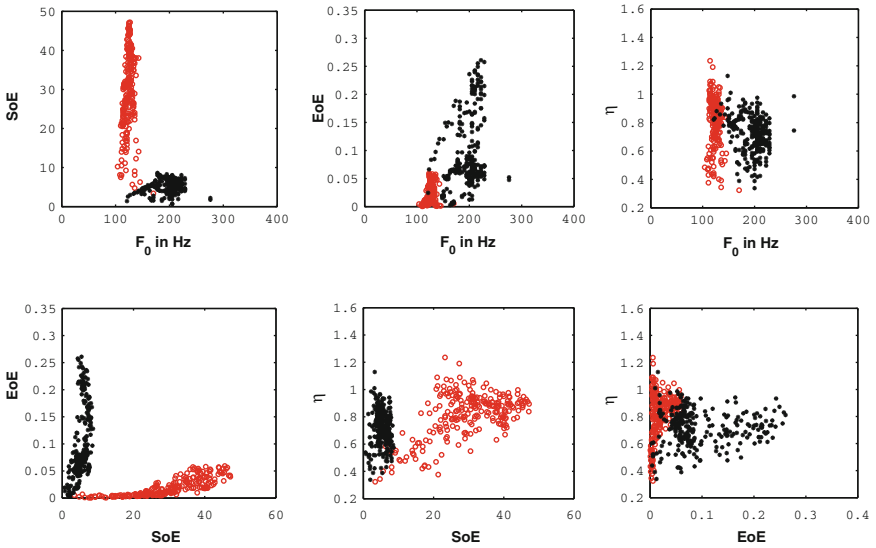

Fig. 11.1 2-D scatter plots of a neutral utterance (marked by 'o') and angry utterance (marked by '$*$') of a speaker [42]

Table 11.6 Average KL distance values between reference (neutral (NU)) utterance and test (angry (AN), happy (HA), sad (SA) and NU) utterances, for IIIT-H Telugu emotion database [42]

	2-D feature spaces					
	F_0 versus *SoE*	F_0 versus *SoE*	F_0 versus η	*SoE* versus *SoE*	η versus *SoE*	η versus *EoE*
Speaker 1						
NU versus NU	0.02	0.02	0.01	0.07	0.02	0.02
NU versus AN	2.10	69.70	1.60	110.00	0.40	93.00
NU versus HA	0.89	28.29	0.70	44.35	0.17	27.56
NU versus SA	0.96	0.91	0.77	0.49	0.27	0.13
Speaker 2						
NU versus NU	0.06	0.09	0.01	0.17	0.06	0.09
NU versus AN	0.10	47.50	0.10	66.20	0.20	46.70
NU versus HA	0.41	22.09	0.25	312.74	0.15	20.90
NU versus SA	0.06	0.29	0.08	0.29	0.07	0.30

test utterance is not neutral, which indicates that the speakers modify the excitation characteristics while producing emotional speech.

In the above case, the reference and test utterances have different lexical content. In order to study the effectiveness of the excitation source features, the reference and test utterances with the same lexical content are considered for the simulated parallel databases (IITKGP-SESC and EMO-DB). The average KL distance values for these databases are given in Table 11.7. The average KL distance values are low when the

Table 11.7 Average KL distance values between reference (neutral (NU)) utterance and test (angry (AN), happy (HA), sad (SA) and NU) utterances, for IITKGP-SESC and EMO-DB databases [42]

	IITKGP-SESC						EMO-DB					
	2-D feature spaces						2-D feature spaces					
	F_0 versus *SoE*	F_0 versus *EoE*	F_0 versus η	*SoE* versus *EoE*	η versus *SoE*	η versus *EoE*	F_0 versus *SoE*	F_0 versus *EoE*	F_0 versus η	*SoE* versus *EoE*	η versus *SoE*	η versus *EoE*
Speaker 1												
NU versus NU	0.07	2.08	0.01	2.14	0.06	2.26	0.11	0.33	0.02	0.40	0.10	0.35
NU versus AN	0.29	86.86	0.04	86.76	0.26	98.58	2.19	4.57	1.01	3.70	0.70	4.01
NU versus HA	0.57	14.84	0.24	15.04	0.31	16.83	0.24	1.34	1.13	2.38	0.85	0.92
NU versus SA	0.28	1.01	0.08	1.14	0.18	1.10	2.33	2.59	0.13	4.05	1.53	2.51
Speaker 2												
NU versus NU	0.07	0.10	0.03	0.15	0.04	0.09	0.67	0.25	0.05	0.82	0.61	0.28
NU versus AN	0.35	38.20	0.13	53.20	0.21	38.60	2.41	18.13	1.47	17.66	0.85	19.68
NU versus HA	0.38	4.90	0.19	5.24	0.08	3.46	3.71	15.21	7.41	11.51	5.48	12.42
NU versus SA	0.21	0.55	0.18	1.68	0.06	1.48	5.72	17.90	13.51	19.10	6.92	40.02

reference and test utterances are both neutral, higher otherwise. This indicates that the excitation source features are independent of the lexical content.

Apart from the above general observations, the average KL distance values are observed to be dependent on speaker, emotion and even culture. Among all the considered emotion categories, these values are high in the case of anger. One of the reasons is due to higher and wider F_0 variations in angry speech [88]. As anger is extremely charged state, the resulting F_0 variations might be because of increased vocal effort. The values in the case of sadness for IIIT-H Telugu and IITKGP-SESC databases are observed to be low. In the case of EMO-DB database, the values are observed to be higher for sad speech.

11.4.4 Emotion Recognition System Based on Excitation Source Features

A system for automatic recognition of emotions is developed based on the excitation source feature deviations of emotional speech w.r.t. neutral speech. The emotions considered for this study are: anger, happiness, neutral and sadness. Three excitation source features, namely, instantaneous F_0, *SoE* and *EoE* are used. Distributions of pairs of F_0, *SoE* and *EoE* are used (see Fig. 11.1) to derive feature representations for the emotion recognition system [42].

The distributions of neutral utterances are first normalized as follows: Let the distributions of F_0, *SoE* and *EoE* for a neutral utterance be denoted by R_{F_0}, R_{SoE} and R_{EoE}, and for an emotion utterance by E_{F_0}, E_{SoE} and E_{EoE}, respectively. Let $R_{m_{F_0}}$, $R_{m_{SoE}}$ and $R_{m_{EoE}}$ represent the mean values, and $R_{\sigma_{F_0}}$, $R_{\sigma_{SoE}}$ and $R_{\sigma_{EoE}}$ represent the standard deviations of the distributions of R_{F_0}, R_{SoE} and R_{EoE}, respectively. The distributions are normalized w.r.t. mean and standard deviation. The normalized distribution for R_{F_0} is given by

$$N_{R_{F_0}} = \frac{R_{F_0} - R_{m_{F_0}}}{R_{\sigma_{F_0}}}. \tag{11.2}$$

Similarly, the normalized distributions $N_{R_{SoE}}$ and $N_{R_{EoE}}$ are obtained for R_{SoE} and R_{EoE}, respectively.

The distributions of the features of an emotion utterance are normalized w.r.t. the neutral utterance as follows. The normalized distribution of E_{F_0} is given by

$$N_{E_{F_0}} = \frac{E_{F_0} - R_{m_{F_0}}}{R_{\sigma_{F_0}}}. \tag{11.3}$$

Similarly, the normalized distributions $N_{E_{SoE}}$ and $N_{E_{EoE}}$ are obtained for E_{SoE} and E_{EoE}, respectively. The normalization is carried out in a speaker-specific way using the speaker's neutral utterance.

Three 2-D feature distributions, $D1$: ($N_{E_{F_0}}$ versus $N_{E_{SoE}}$), $D2$: ($N_{E_{EoE}}$ versus $N_{E_{F_0}}$) and $D3$: ($N_{E_{EoE}}$ versus $N_{E_{SoE}}$) are modeled by a Gaussian distribution, represented by mean vector and covariance matrix. These distributions capture the emotion-specific feature deviations. In the training phase, for each speaker, for each emotion utterance, the 2-D feature distributions (templates) are extracted. In the testing phase, a speaker's neutral speech is required to obtain the normalized distributions for the test utterances. The KL distance scores are computed among the corresponding 2-D distributions of test utterance and stored templates. The emotion category with maximum matched templates is declared for the test utterance.

A 2-stage binary hierarchical classification is implemented. In Stage 1, anger and happiness emotions are grouped into one class, and sadness and neutral emotions are grouped into another class. In Stage 2, the comparisons are made between neutral versus sadness categories, and anger versus happiness categories.

The confusion matrices for IIIT-H Telugu emotion database after Stage 1 and Stage 2 are shown in Tables 11.8 and 11.9, respectively. From the values listed in Table 11.8, the binary classification at stage 1 gives 96 % accuracy. This is in conformity with the studies [76], where generally the acoustic features effectively discriminate between high arousal and low arousal emotions. From Table 11.9, it is observed that the confusion between anger and happiness states is high. The recognition for neutral, sadness, anger and happiness emotions are 91.2, 97, 71.43 and 52 %, respectively, giving a total recognition accuracy of 79.23 % for the 4 class problem.

The proposed method is also applied on the EMO-DB database, and the results are shown in Table 11.10. The emotion recognition at Stage 2 of EMO-DB is 75 %. The performance of the EMO-DB is low because the confusions between angry and happy utterances are observed high.

The results of the proposed method indicate that the features corresponding to the excitation source seem to carry emotion-specific information. The performance of the system can be improved by increasing the number of reference (trained) templates and speakers.

11.4.5 Discrimination of Anger and Happiness

The production characteristics of angry and happy speech are examined to determine features that can discriminate these two emotions. In particular, the closed quotient (C_q) of the glottal vibration, *SoE* and the ratio of high to low frequency band energies are used.

Table 11.8 Confusion matrix after Stage 1 for IIIT-H Telugu emotion database [66]

	Neutral/sad	Angry/happy
Neutral/sad	68/68	0/68
Angry/happy	5/62	57/62

Table 11.9 Confusion matrix after Stage 2 for IIIT-H Telugu emotion database [66]

	Neutral	Sad	Angry	Happy
Neutral	31/34	3/34	0/34	0/34
Sad	1/34	33/34	0/34	0/34
Angry	0/35	0/35	25/35	10/35
Happy	1/27	4/27	8/27	14/27

Table 11.10 Confusion matrix after Stage 2 for EMO-DB database [66]

	Neutral	Sad	Angry	Happy
Neutral	74/79	4/79	0/79	1/79
Sad	27/62	33/62	0/62	2/62
Angry	2/127	0/127	114/127	11/127
Happy	7/71	3/71	27/71	34/71

The C_q of a glottal pulse is the ratio of the closed phase duration to the duration of the total glottal pulse, and is denoted by γ in percentage. The open and closed phases of a glottal cycle are illustrated in Fig. 11.2 through the EGG and the derivative of EGG (dEGG) signals. The amplitude of the EGG (current flow) signal is larger during the close phase region due to low impedance and lower during the open phase region due to high impedance across vocal folds. The locations of the GCI are associated with positive peaks, and the locations of the glottal opening instant are associated with negative peaks in the dEGG signal. The average (A_γ) values of γ for neutral, happiness and anger emotions, for five speakers of IIIT-H Telugu and EMO-DB databases are given in Tables 11.11 and 11.12, respectively. From the values of A_γ, it can be observed that the C_q has increasing trend in happiness and anger emotions for a given speaker, with anger possessing relatively higher value.

The ratio of high (800–5000 Hz) to low (0–400 Hz) frequency band energy is denoted as β. The β values are computed from the short time Fourier transform. In [86], it was reported that, increase in C_q increases the value of β. The average (A_β) values of β and the average (A_{SoE}) values of SoE for neutral, happiness and anger emotions for 5 speakers of IIIT-H Telugu and EMO-DB databases are shown in Tables 11.11 and 11.12, respectively.

Tables 11.11 and 11.12 show that the values of A_β and A_{SoE} can be related to the value of C_q. Increase in C_q (γ in percentage) is observed with increase in β and decrease in SoE. There is increasing trend in the values of A_β and decreasing trend in the values of A_{SoE} for happiness and anger emotions w.r.t. neutral. The values of A_β are relatively higher, and the values of A_{SoE} are relatively lower in case of anger when compared to the case of happiness. Although these features carry discriminative property, the dynamic ranges of their values are speaker-specific.

A speaker-specific anger versus happiness classification is implemented. A sample 2-D feature distributions (β versus SoE) of happy and angry utterances is given in

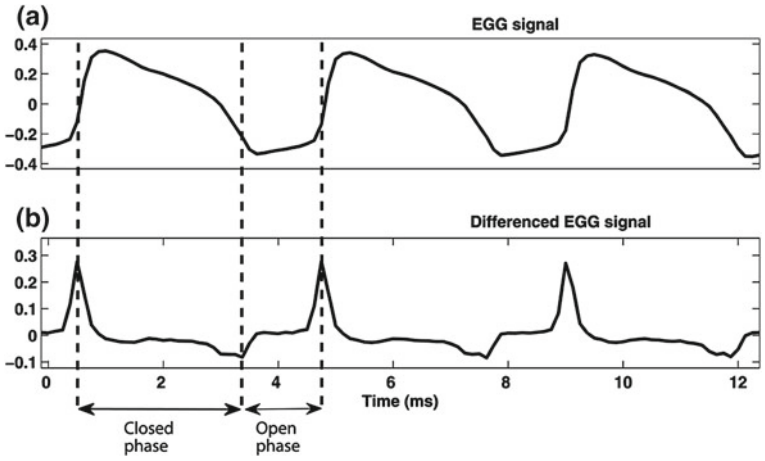

Fig. 11.2 Illustration of open and closed phase durations. **a** EGG signal. **b** dEGG signal [86]

Table 11.11 Average (A_γ, A_β and A_{SoE}) values of γ, β and SoE of neutral, happy and angry utterances for 5 speakers of IIIT-H Telugu database [43]

	Neutral			Happy			Angry		
	A_γ	A_β	A_{SoE}	A_γ	A_β	A_{SoE}	A_γ	A_β	A_{SoE}
Speaker 1	42.21	8.63	43.53	48.38	12.89	21.63	53.12	13.32	13.47
Speaker 2	39.86	9.86	28.93	42.19	13.26	13.91	44.31	14.26	7.76
Speaker 3	41.46	7.34	37.21	46.32	13.90	19.12	51.74	13.95	11.49
Speaker 4	44.55	7.40	18.97	44.19	13.62	7.76	46.14	13.81	5.67
Speaker 5	41.87	8.78	16.45	47.43	13.56	4.98	47.23	13.96	4.04
Mean value	41.99	8.40	29.02	45.70	13.45	13.48	48.51	13.86	8.49

Table 11.12 Average (A_γ, A_β and A_{SoE}) values of γ, β and SoE of neutral, happy and angry utterances for 5 speakers of EMO-DB database [43]

	Neutral			Happy			Angry		
	A_γ	A_β	A_{SoE}	A_γ	A_β	A_{SoE}	A_γ	A_β	A_{SoE}
Speaker 1	42.97	7.23	23.42	48.91	10.44	12.43	50.63	12.45	9.43
Speaker 2	44.14	6.32	19.34	47.89	11.02	11.34	51.01	11.53	8.32
Speaker 3	39.26	8.21	17.44	46.21	9.87	12.23	48.54	13.47	10.21
Speaker 4	44.04	5.56	11.21	48.24	8.32	3.96	49.66	9.61	3.32
Speaker 5	44.83	8.30	12.56	48.56	11.29	6.23	49.15	13.81	4.92
Mean value	43.05	7.12	16.79	47.96	10.19	9.24	49.80	12.17	7.24

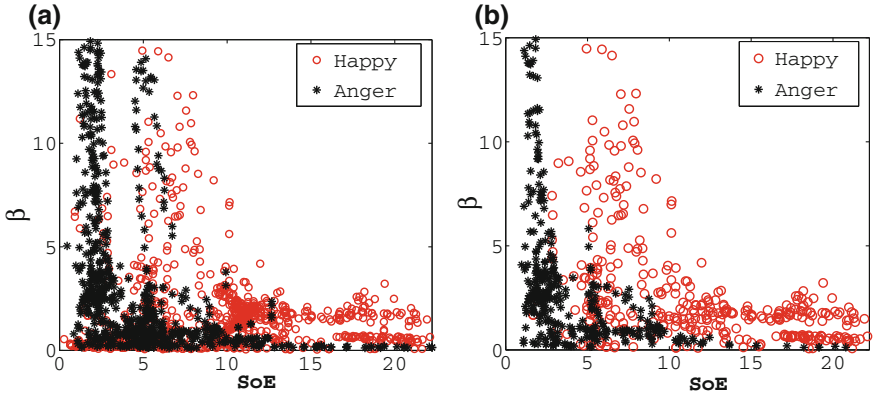

Fig. 11.3 Feature distributions (β versus *SoE*) of happy and angry emotion utterances. **a** Entire utterance. **b** High F_0 regions of the utterance [43]

Fig. 11.3a. As these emotions are produced suddenly in response to a particular stimuli, they are not sustainable. There may be neutral regions in these utterances [88]. To reduce the effect of neutral regions, the regions corresponding to high F_0 (above the average F_0) are considered. This can be justified as the speech in happiness and anger states is produced with increased pitch levels. The 2-D feature distributions of high F_0 regions of the happy and angry utterances are given in Fig. 11.3b. From Fig. 11.3a, b, it is evident that there is better discrimination between happiness and anger in the high F_0 regions.

Given a set of happy and angry utterances, the 2-D feature distributions (β versus *SoE*) of all utterances are computed. The feature distribution with low mean *SoE* and high mean β is considered as reference for anger emotion, and the feature distribution with high mean *SoE* and low mean β is considered as reference for happiness emotion. The remaining feature distributions are classified by computing the KL distance with the reference feature distributions. The lower the KL distance, the closer are the test distributions towards the reference. The classification accuracy of approximately 85 % is observed for both the databases.

In general, the positive emotions possesses more rhythmic behavior than the negative emotions [88]. To exploit this characteristic, the variances of *SoE* values are examined. The variance of *SoE* corresponding to high F_0 regions is compared with the variance of the entire utterance. A parameter called relative *SoE* variance is defined. This is given by

$$S_v = \frac{V_h - V_t}{(\frac{\mu_h + \mu_t}{2})}, \tag{11.4}$$

where V_h and V_t are the *SoE* variances in the selected high F_0 regions and for the entire utterance, respectively, and μ_h and μ_t are the corresponding means. The average value of S_v of happy and angry utterances of IIIT-Telugu and EMO-DB databases is given in Table 11.13. It is observed that the S_v value for happy utterances is mostly

Table 11.13 Average value of S_v of happy and angry utterances of IIIT-H Telugu and German EMO-DB databases [43]

	IIIT-H Telugu		German EMO-DB	
	Happy	Angry	Happy	Angry
Average S_v	0.42	−0.51	−0.21	−0.58

positive in the case of IIIT-H Telugu database. The S_v value for angry utterances is observed to be mostly negative in both the databases. Sample histograms of happy and angry utterances are given in Fig. 11.4. Following this observation, a classification approach is proposed. The percentage of accuracy of 75 and 68 % are observed for IIIT-H Telugu and EMO-DB databases, respectively.

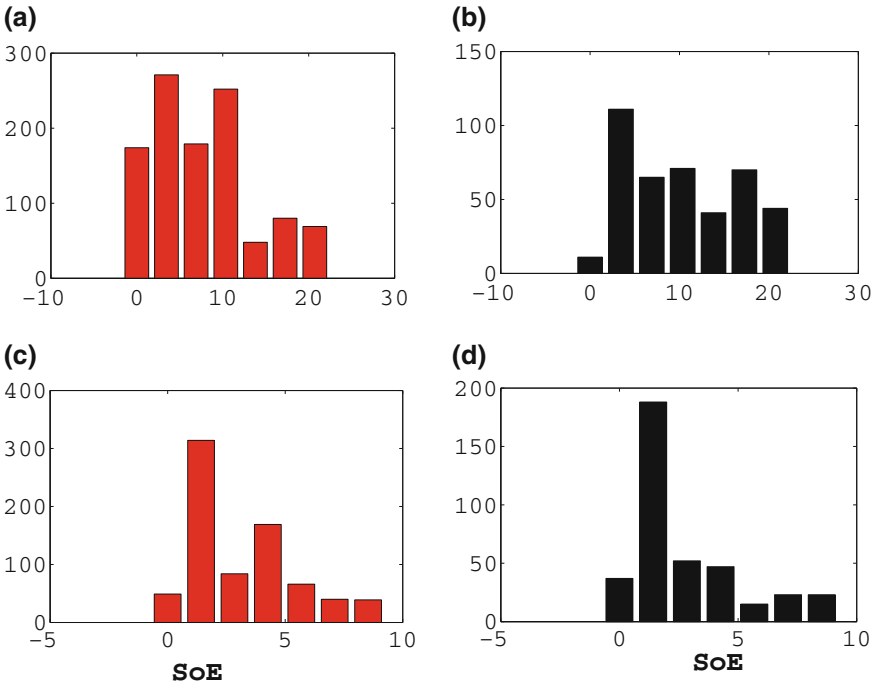

Fig. 11.4 Histograms of *SoE* values of happy and angry utterances. **a** Entire happy utterance. **b** High F_0 regions of the happy utterance. **c** Entire angry utterance. **d** High F_0 regions of the angry utterance [43]

11.4.6 Discrimination Between High Arousal and Falsetto Voices

High arousal speech is produced by a speaker in situations like emotionally charged states, communication over a long distance and in noisy environments. High arousal speech is often produced with increased levels of voice intensity. In natural communication, speakers tend to increase their pitch in high arousal speech. Increase in pitch can also be intentional as in falsetto register. This study is to discriminate high arousal and falsetto voices, as either of these cases there is increase in the average F_0 w.r.t. neutral.

The vocal folds configuration is significantly different in modal and falsetto registers [124, 130]. It is observed that the C_q of the vocal fold vibration is lower for segments of falsetto voice when compared to that of modal voice.

The average (A_γ) values of γ and the average (A_{F_0}) values of F_0 of neutral, happy, angry, shout and falsetto utterances for a set of five speakers from IIIT-H Telugu database are given in Table 11.14. On observation of A_{F_0} values, it is evident that falsetto modes are produced with distinct pitch variations. It is also clear that the F_0 range of low falsetto voice matches well with that of high arousal. The values of A_γ increases for happy, angry and shout speech w.r.t. neutral speech. For all the cases of falsetto, there is significant decrease in the A_γ values. It is interesting to note that the A_γ values are decreasing with the ascending degree of falsetto. The inconsistencies for different falsetto cases may be because of the lack of control in the voice intensity while producing speech at higher pitch levels by these speakers, who are not trained to produce these voices.

The glottal vibration characteristics also have an effect on the vocal tract system response. The effective length of the vocal tract changes in the open and closed regions of the vocal folds. Therefore, useful information of the excitation source characteristics can be obtained from speech analysis with high spectro-temporal resolution. A recently proposed zero time windowing (ZTW) method [135] is useful for this analysis. A heavily tapering window in the temporal domain, given by

$$h[n] = \frac{1}{8sin(\frac{\omega n}{N})^4}, \tag{11.5}$$

is used at each sampling instant. The Hilbert envelope of the numerator of group delay function (HNGD) of the windowed segment is computed at every sampling instant.

The β feature, which is the ratio of energies in the high (800–5000 Hz) and low (0–400 Hz) frequency bands of the HNGD spectra is computed. The β contour for a speech segment is shown in Fig. 11.5. The sharp peaks in the β contour occur mainly at GCIs. The β contour has lower values during the open phase.

The highest and the lowest values of the β contour within a glottal cycle are denoted as β_c and β_o, respectively. The average (A_{β_c} and A_{β_o}) values of β_c and β_o, respectively, of neutral, happy, angry, shout and falsetto utterances are given in

Table 11.14 Average (A_γ and A_{F_0} in Hz) values of γ and F_0 of neutral, happy, angry, shout and falsetto utterances for 5 speakers

	Neutral		Happy		Angry		Shout		Low falsetto		Mid falsetto		High falsetto	
	A_γ	A_{F_0}	A_γ	A_{F_0}	A_γ	A_{F_0}	A_γ	A_{F_0}	A_γ	A_{F_0}	A_γ	A_{F_0}	A_γ	A_{F_0}
Speaker 1	42.21	135	48.38	161	53.12	197	56.21	257	33.42	213	31.23	362	35.13	428
Speaker 2	39.86	248	42.19	326	44.31	311	48.92	331	37.87	328	39.13	413	42.14	527
Speaker 3	41.46	142	46.32	198	51.74	241	50.46	238	32.19	321	33.12	376	37.98	371
Speaker 4	44.55	221	44.19	239	46.14	318	53.16	326	37.32	367	36.21	432	41.19	463
Speaker 5	41.87	162	47.43	245	47.23	287	46.37	328	34.42	340	34.56	444	39.17	522
Mean value	41.99	182	45.70	234	48.51	271	51.02	296	35.04	314	34.85	405	39.12	462

Table 11.15 Average (A_{β_c} and A_{β_o}) values of β_c and β_o of neutral, happy, angry, shout and falsetto utterances for 5 speakers

	Neutral		Happy		Angry		Shout		Low falsetto		Mid falsetto		High falsetto	
	A_{β_c}	A_{β_o}	A_{β_c}	A_{β_o}	A_{β_c}	A_{β_o}	A_{β_c}	A_{β_o}	A_{β_c}	A_{β_o}	A_{β_c}	A_{β_o}	A_{β_c}	A_{β_o}
Speaker 1	0.21	0.03	0.37	0.05	0.41	0.11	0.43	0.17	0.22	0.07	0.33	0.16	0.39	0.21
Speaker 2	0.51	0.10	0.55	0.08	0.82	0.24	1.11	0.29	0.48	0.27	1.36	0.64	1.9	0.96
Speaker 3	0.27	0.04	0.33	0.04	0.39	0.09	0.45	0.19	0.25	0.08	0.26	0.1	0.48	0.31
Speaker 4	0.49	0.09	0.63	0.04	1.21	0.31	1.40	0.36	0.47	0.27	1.36	0.63	1.51	0.69
Speaker 5	0.23	0.03	0.23	0.02	0.29	0.05	0.39	0.14	0.25	0.07	0.41	0.21	0.53	0.26
Mean value	0.34	0.06	0.42	0.05	0.62	0.16	0.76	0.23	0.33	0.15	0.74	0.35	0.96	0.49

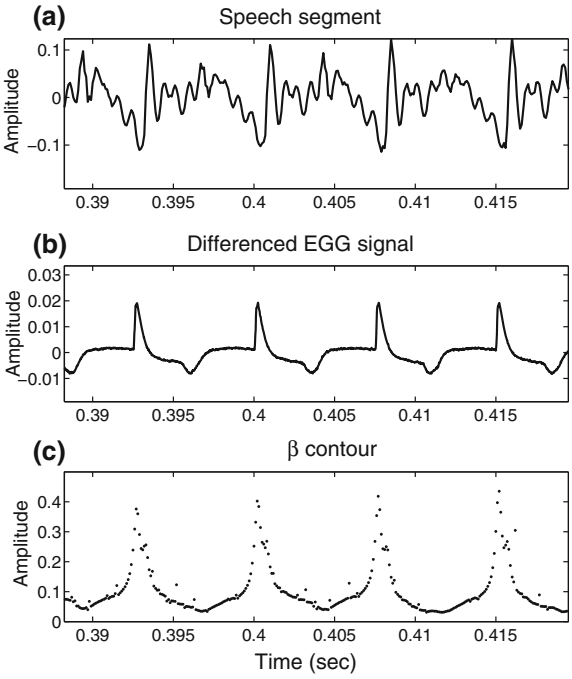

Fig. 11.5 **a** Segment of a speech signal. **b** Differenced EGG signal. **c** β contour

Table 11.15. From Table 11.15, it is evident that there is significant increase in A_{β_c} for angry and shout speech. A similar, but not significant trend can be observed in the case of happy speech. From these results, we can infer that increase in C_q increases A_{β_c}. In the case of low falsetto voice, A_{β_c} is in similar range as for the neutral. But, the results are not consistent for mid and high falsetto voices.

11.5 Research Challenges in Emotional Speech Analysis

The major issues in dealing with emotional speech analysis are the description of emotion, data collection and feature representation. The production and perception of emotions in speech by human beings is a complex phenomenon, which is not well understood. A basic and widely accepted statement is that emotions are underlying states of a speaker, and emotion expressions are the communicative value of the underlying states [22]. In natural communication, speakers may sometimes try to hide the underlying emotional state. It is very difficult to collect spontaneous emotion data, and also it is difficult to annotate the collected data. This statement is supported by the fact that many analysis studies in the literature use full-blown emotional data. Full-blown emotional speech corresponds to intense expressions of underlying emotional states [104].

The features used for analysis are broadly categorized as prosody, voice quality and spectral features. These features carry emotion correlates, but they are observed to be speaker and sound-specific. To derive an emotion-specific feature representation, the speech production knowledge of emotions is required. In general, speech in different emotion states is produced with distinct changes in speech production mechanism. However, it is difficult to describe these emotion-specific deviations in the production mechanism. This makes emotional speech analysis a difficult task.

The state-of-the-art approaches for emotion recognition are adopted from the developments in applications like speech recognition and speaker recognition. In these approaches, the utterances are labeled with the discrete emotion or primitive categories. Pattern recognition models like GMMs, SVMs, ANNs and HMMs are trained on the features extracted. As most of the feature representations are speaker and sound-specific, these pattern recognition models require a phonetically/phonemically balanced data, covering several speakers. In practical sense, it is difficult to collect such a database. Given the limitations of existing features and data collection issues, the ideal scenario is to identify underlying emotion in speech by a feature representation.

The experiments conducted in our research show that the excitation source features carry significant amount of emotion related information, and these features are also observed to be speaker-specific. It is indeed a challenge to determine the excitation source features that are only emotion-specific.

Although it is difficult to define production characteristics that are specific to an emotion, there are clues when emotions are viewed in the 3 dimensions/primitives (arousal, valence and dominance). There is increase in F_0 in the case of high arousal speech, and decrease in F_0 in the case of low arousal speech. But there may be cases where speech can be produced with deliberate increase/decrease in F_0. Discrimination of speech with natural increase/decrease in F_0 and deliberate increase/decrease in F_0 is an important issue in emotion studies. In the case of valence dimension, in [88], it was reported that rhythm and valence were consistently related. The positive feelings possess regular rhythm than negative feelings. From our studies (reported in Sects. 11.4.6 and 11.4.5), it can be said that feature representation of primitives of emotion might help in representing emotional speech effectively. From the current studies, it appears that emotion recognition by a machine appears to be an elusive goal.

References

1. Airas M, Alku P (2004) Emotions in short vowel segments: effects of the glottal flow as reflected by the normalized amplitude quotient. In: Affective dialogue systems. Springer, pp 13–24
2. Airas M, Pulakka H, Bäckström T, Alku P (2005) A toolkit for voice inverse filtering and parametrization. In: INTERSPEECH. Lisbon, Portugal, pp 2145–2148
3. Alku P (2011) Glottal inverse filtering analysis of human voice production a review of estimation and parameterization methods of the glottal excitation and their applications. Sadhana 36(5):623–650

4. Alku P, Vilkman E (1996) A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers. Folia Phoniatrica et Logopaedica 48:240–254
5. Amer MR, Siddiquie B, Richey C, Divakaran A (2014) Emotion recognition in speech using deep networks. In: ICASSP. Florence, Italy, pp 3752–3756
6. Amir N, Kerret O, Karlinski D (2001) Classifying emotions in speech: a comparison of methods. In: INTERSPEECH. Aalborg, Denmark, pp 127–130
7. Ang j, Dhillon R, Krupski A, Shriberg E, Stolcke A (2002) Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: INTERSPEECH. Denver, Colorado, USA
8. Arias JP, Busso C, Yoma NB (2013) Energy and F0 contour modeling with functional data analysis for emotional speech detection. In: INTERSPEECH. Lyon, France, pp 2871–2875
9. Arias JP, Busso C, Yoma NB (2014) Shape-based modeling of the fundamental frequency contour for emotion detection in speech. Comput Speech Lang 28(1):278–294
10. Atassi H, Esposito A (2008) A speaker independent approach to the classification of emotional vocal expressions. In: IEEE international conference on tools with artificial intelligence (ICTAI'08), vol 2. Dayton, Ohio, USA, pp 147–152
11. Atassi H, Riviello M, Smékal Z, Hussain A, Esposito A (2010) Emotional vocal expressions recognition using the COST 2102 Italian database of emotional speech. In: Esposito A, Campbell N, Vogel C, Hussain A, Nijholt A (eds) Development of multimodal interfaces: active listening and synchrony. Lecture notes in computer science, vol 5967. Springer, Berlin, pp 255–267
12. Bachorowski J (1999) Vocal expression and perception of emotion. Curr Dir Psychol Sci 8(2):53–57
13. Banse R, Scherer KR (1996) Acoustic profiles in vocal emotion expression. J Personal Soc Psychol 70(3):614–636
14. Batliner A, Schuller B, Seppi D, Steidl S, Devillers L, Vidrascu L, Vogt T, Aharonson V, Amir N (2011) The automatic recognition of emotions in speech. In: Petta P, Pelachaud C, Cowie R (eds) Emotion-oriented systems. Springer, pp 71–99
15. Bezooijen RAMG, Otto SA, Heenan TA (1983) Recognition of vocal expressions of emotion: a three-nation study to identify universal characteristics. J Cross-Cult Psychol 14:387–406
16. Boersma P, Heuven VV (2001) Speak and unSpeak with PRAAT. Glot Int 5(9/10):341–347
17. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of German emotional speech. In: INTERSPEECH. Lisbon, Portugal, pp 1517–1520
18. Busso C, Bulut M, Lee C, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan S (2008) IEMOCAP: interactive emotional dyadic motion capture database. Lang Res Eval 42(4):335–359
19. Chastagnol C, Devillers L (2011) Analysis of anger across several agent-customer interactions in French call centers. In: ICASSP. Prague, Czech Republic, pp 4960–4963
20. Childers DG, Lee CK (1991) Vocal quality factors: analysis, synthesis, and perception. J Acoust Soc Am 90(5):2394–2410
21. Cowie R, Cornelius RR (2003) Describing the emotional states that are expressed in speech. Speech Commun 40(1–2):5–32
22. Darwin C (1872) The expression of emotion in man and animals. reprinted by University of Chicago Press, Murray, London, UK (1975)
23. Davitz JR (1964) Personality, perceptual, and cognitive correlates of emotional sensitivity. In: Davitz JR (ed) The communication of emotional meaning. McGraw-Hill, New York
24. Dellaert F, Polzin T, Waibel A (1996) Recognizing emotion in speech. In: international conference on spoken language processing (ICSLP). Philadelphia, USA, pp 1970–1973
25. Devillers L, Vidrascu L (2006) Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In: INTERSPEECH. Pittsburgh, PA, USA, pp 801–804
26. Douglas-Cowie E, Campbell N, Cowie R, Roach P (2003) Emotional speech: towards a new generation of databases. Speech Commun 40(1–2):33–60

27. Ekman P (1992) An argument for basic emotions. Cognit Emot 6:169–200
28. Engberg IS, Hansen AV, Andersen O, Dalsgaard P (1997) Design, recording and verification of a Danish emotional speech database. In: EUROSPEECH. Rhodes, Greece, pp 1695–1698
29. Erden M, Arslan LM (2011) Automatic detection of anger in human-human call center dialogs. In: INTERSPEECH. Florence, Italy, pp 81–84
30. Erickson D, Yoshida K, Menezes C, Fujino A, Mochida T, Shibuya Y (2006) Exploratory study of some acoustic and articulatory characteristics of sad speech. Phonetica 63:1–5
31. Erro D, Navas E, Hernáez I, Saratxaga I (2010) Emotion conversion based on prosodic unit selection. IEEE Trans Audio Speech Lang Process 18(5):974–983
32. Espinosa HP, Garcia JO, Pineda LV (2010) Features selection for primitives estimation on emotional speech. In: ICASSP. Florence, Italy, pp 5138–5141
33. Eyben F, Wollmer M, Schuller B (2009) OpenEarIntroducing the Munich open-source emotion and affect recognition toolkit. In: International conference on affective computing and intelligent interaction and workshops (ACII). Amsterdam, Netherlands, pp 1–6
34. Eyben F, Batliner A, Schuller B, Seppi D, Steidl S (2010) Cross-corpus classification of realistic emotions—some pilot experiments. In: International workshop on EMOTION (satellite of LREC): corpora for research on emotion and affect. Valletta, Malta, pp 77–82
35. Eyben F, Wöllmer M, Schuller B (2010) OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In: International conference on multimedia. Firenze, Italy, pp 1459–1462
36. Fairbanks G, Hoaglin LW (1941) An experimental study of the durational characteristics of the voice during the expression of emotion. Speech Monogr 8:85–91
37. Fairbanks G, Pronovost W (1939) An experimental study of the pitch characteristics of the voice during the expression of emotion. Speech Monogr 6:87–104
38. Fant G, Lin Q, Gobl C (1985) Notes on glottal flow interaction. Speech Transm Lab Q Progress Status Rep, KTH 26:21–25
39. Fernandez R, Picard R (2011) Recognizing affect from speech prosody using hierarchical graphical models. Speech Commun 53(9–10):1088–1103
40. Fonagy I, Magdics K (1963) Emotional patterns in intonation and music. Kommunikationforsch 16:293–326
41. Gangamohan P, Mittal VK, Yegnanarayana B (2012) A flexible analysis and synthesis tool (FAST) for studying the characteristic features of emotion in speech. In: IEEE international conference on consumer communications and networking conference. Las Vegas, USA pp 266–270
42. Gangamohan P, Sudarsana RK, Yegnanarayana B (2013) Analysis of emotional speech at subsegmental level. In: INTERSPEECH. Lyon, France, pp 1916–1920
43. Gangamohan P, Sudarsana RK, Suryakanth VG, Yegnanarayana B (2014) Excitation source features for discrimination of anger and happy emotions. In: INTERSPEECH. Singapore, pp 1253–1257
44. Gnjatovic M, Rösner D (2010) Inducing genuine emotions in simulated speech-based human-machine interaction: the nimitek corpus. IEEE Trans Affect Comput 1(2):132–144
45. Gobl C (1988) Voice source dynamics in connected speech. Speech Trans Lab Q Progress Status Rep, KTH 1:123–159
46. Gobl C (1989) A preliminary study of acoustic voice quality correlates. Speech Trans Lab Q Progress Status Rep, KTH 4:9–21
47. Gobl C, Chasaide AN (1992) Acoustic characteristics of voice quality. Speech Commun 11(4):481–490
48. Gobl C, Chasaide AN (2003) The role of voice quality in communicating emotion, mood and attitude. Speech Commun 40(1–2):189–212
49. Grichkovtsova I, Morel M, Lacheret A (2012) The role of voice quality and prosodic contour in affective speech perception. Speech Commun 54(3):414–429
50. Grimm M, Kroschel K, Mower E, Narayanan S (2007) Primitives-based evaluation and estimation of emotions in speech. Speech Commun 49(10–11):787–800

51. Grimm M, Kroschel K, Narayanan S (2008) The Vera am Mittag German audio-visual emotional speech database. In: International conference on multimedia and expo. Hannover, Germany, pp 865–868
52. Guruprasad S, Yegnanarayana B (2009) Perceived loudness of speech based on the characteristics of glottal excitation source. J Acoust Soc Am 126(4):2061–2071
53. Hansen JH, Womack BD (1996) Feature analysis and neural network-based classification of speech under stress. IEEE Trans Speech Audio Process 4(4):307–313
54. Hanson HM (1997) Glottal characteristics of female speakers: acoustic correlates. J Acoust Soc Am 101(1):466–481
55. Hassan A, Damper RI (2010) Multi-class and hierarchical SVMs for emotion recognition. In: INTERSPEECH. Chiba, Japan, pp 2354–2357
56. He L, Lech M, Allen N (2010) On the importance of glottal flow spectral energy for the recognition of emotions in speech. In: INTERSPEECH. Chiba, Japan, pp 2346–2349
57. Hershey JR, Olsen PA (2007) Approximating the Kullback Leibler divergence between Gaussian mixture models. In: ICASSP, vol 4. Montreal, Quebec, Canada, pp 317–320
58. Huber R, Batliner A, Buckow J, Nöth E, Warnke V, Niemann H (2000) Recognition of emotion in a realistic dialogue scenario. In: Proceedings of international conference on spoken language processing. Beijing, China, pp 665–668
59. Hübner D, Vlasenko B, Grosser T, Wendemuth A (2010) Determining optimal features for emotion recognition from speech by applying an evolutionary algorithm. In: INTERSPEECH. Chiba, Japan, pp 2358–2361
60. Izard CE (1977) Human emotions. Plenum Press, New York
61. Jeon JH, Xia R, Liu Y (2011) Sentence level emotion recognition based on decisions from subsentence segments. In: ICASSP. Lyon, France, pp 4940–4943
62. Jeon JH, Le D, Xia R, Liu Y (2013) A preliminary study of cross-lingual emotion recognition from speech: automatic classification versus human perception. In: INTERSPEECH. Prague, Czech Republic, pp 2837–2840
63. Joachims T (1998) Text categorization with support vector machines: Learning with many relevant features. In: European conference on machine learning. London, UK, pp 137–142
64. Kadiri SR, Gangamohan P, Mittal VK, Yegnanarayana B (2014) Naturalistic audio-visual emotion database. In: International conference on natural language processing. Goa, India, pp 127–134
65. Kadiri SR, Gangamohan P, Yegnanarayana B (2014) Discriminating neutral and emotional speech using neural networks. In: Interenational conference on natural language processing. Goa, India, pp 119–126
66. Kadiri SR, Gangamohan P, Gangashetty SV, Yegnanarayana B (2015) Analysis of excitation source features of speech for emotion recognition. In: INTERSPEECH. Dresden, Germany, pp 1032–1036
67. Keller E (2005) The analysis of voice quality in speech processing. In: Gèrard C, Anna E, Marcos F, Maria M (eds) Lecture notes in computer science. Springer, pp 54–73
68. Kim W, Hansen JHL (2010) Angry emotion detection from real-life conversational speech by leveraging content structure. In: ICASSP. Dallas, Texas, USA, pp 5166–5169
69. Kim J, Lee S, Narayanan S (2010) An exploratory study of manifolds of emotional speech. In: ICASSP. Dallas, Texas, USA, pp 5142–5145
70. Kim J, Park J, Oh Y (2011) On-line speaker adaptation based emotion recognition using incremental emotional information. In: ICASSP. Prague, Czech Republic, pp 4948–4951
71. Klasmeyer G, Sendlmeier WF (2000) Voice and emotional states. In: Voice quality measurement. Springer, Berlin, Germany, pp 339–358
72. Klatt DH (1980) Software for a cascade/parallel formant synthesizer. J Acoust Soc Am 67(3):971–995
73. Koolagudi SG, Sreenivasa Rao K (2012) Emotion recognition from speech: a review. Int J Speech Technol 15(2):99–117
74. Koolagudi SG, Maity S, Vuppala AK, Chakrabarti S, Sreenivasa Rao K (2009) IITKGP-SESC: speech database for emotion analysis. In: Communications in computer and information science, pp 485–492

75. Laver John DM (1968) Voice quality and indexical information. Int J Lang Commun Disord 3(1):43–54
76. Lee C, Mower E, Busso C, Lee S, Narayanan S (2011) Emotion recognition using a hierarchical binary decision tree approach. Speech Commun 53(9–10):1162–1171
77. Lee CM, Narayanan S (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13(2):293–303
78. Lee CM, Yildirim S, Bulut M, Kazemzadeh A, Busso C, Deng Z, Lee S, Narayanan S (2004) Emotion recognition based on phoneme classes. In: INTERSPEECH. JejuIsland, Korea, pp 205–211
79. Lieberman P, Michaels SB (1962) Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. J Acoust Soc Am 34(7):922–927
80. Lin J, Wu C, Wei W (2013) Emotion recognition of conversational affective speech using temporal course modeling. In: INTERSPEECH. Lyon, France, pp 1336–1340
81. Luengo I, Navas E, Hernáez I, Sánchez J (2005) Automatic emotion recognition using prosodic parameters. In: INTERSPEECH. Lisbon, Portugal, pp 493–496
82. Lugger M, Yang B (2007) The relevance of voice quality features in speaker independent emotion recognition. In: ICASSP, vol 4. Honolulu, Hawaii, USA, pp 17–20
83. Makhoul J (1975) Linear prediction: a tutorial review. Proc IEEE 63:561–580
84. Mansoorizadeh M, Charkari NM (2007) Speech emotion recognition: comparison of speech segmentation approaches. In: Proceedings of IKT, Mashad, Iran
85. 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 tutorial and research workshop (ITRW) on speech and emotion. Newcastle, Northern Ireland, UK
86. Mittal VK, Yegnanarayana B (2013) Effect of glottal dynamics in the production of shouted speech. J Acoust Soc Am 133(5):3050–3061
87. Morrison D, Wang R, De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres. Speech Commun 49(2):98–112
88. Murray IR, Arnott JL (1993) Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J Acoust Soc Am 93(2):1097–1108
89. Murty KSR, Yegnanarayana B (2008) Epoch extraction from speech signals. IEEE Trans Audio Speech Lang Process 16(8):1602–1613
90. Nogueiras A, Moreno A, Bonafonte A, Mariño JB (2001) Speech emotion recognition using hidden Markov models. In: EUROSPEECH. Aalborg, Denmark, pp 2679–2682
91. Nwe TL, Foo SW, De Silva LC (2003) Speech emotion recognition using hidden Markov models. Speech Commun 41(4):603–623
92. Oatley K (1989) The importance of being emotional. New Sci 123:33–36
93. Pereira C (2000) Dimensions of emotional meaning in speech. In: ISCA tutorial and research workshop (ITRW) on speech and emotion. Northern Ireland, UK
94. Polzehl T, Sundaram S, Ketabdar H, Wagner M, Metze F (2009) Emotion classification in children's speech using fusion of acoustic and linguistic features. In: INTERSPEECH. Brighton, UK, pp 340–343
95. Prasanna SRM, Govind D (2010) Analysis of excitation source information in emotional speech. In: INTERSPEECH. Chiba, Japan, pp 781–784
96. Rothenberg M (1973) A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. J Acoust Soc Am 53(6):1632–1645
97. Rozgic V, Ananthakrishnan S, Saleem S, Kumar R, Vembu AN, Prasad R (2012) Emotion recognition using acoustic and lexical features. In: INTERSPEECH. Portland, USA
98. Scherer KR (1981) Speech and emotional states. In: Darby JK (ed) Speech evaluation in psychiatry. Grune and Stratton, New York
99. Scherer KR (1984) On the nature and function of emotion: a component process approach. In: Scherer KR, Ekman P (eds) Approaches to emotion. Lawrence Elbraum, Hillsdale
100. Scherer KR (2003) Vocal communication of emotion: a review of research paradigms. Speech Commun 40(1–2):227–256

101. Scholsberg H (1941) A scale for the judgment of facial expressions. J Exp Psychol 29(6):497–510
102. Schlosberg H (1954) Three dimensions of emotion. J Psychol Rev 61(2):81–88
103. Schröder M (2001) Emotional speech synthesis-a review. In: INTERSPEECH. Aalborg,Denmark, pp 561–564
104. Schröder M (2004) Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis. PhD thesis, Saarland University
105. Schröder M, Cowie R, Douglas-Cowie E, Westerdijk M, Gielen SC (2001) Acoustic correlates of emotion dimensions in view of speech synthesis. In: INTERSPEECH. Aalborg, Denmark, pp 87–90
106. Schuller B (2011) Recognizing affect from linguistic information in 3D continuous space. IEEE Trans Affect Comput 2(4):192–205
107. Schuller B, Rigoll G (2006) Timing levels in segment-based speech emotion recognition. In: INTERSPEECH. Pittsburgh, Pennsylvania, pp 17–21
108. 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: ICASSP vol 1. Montreal, Quebec, Canada, pp 577–580
109. Schuller B, Müller R, Lang M, Rigoll G (2005) Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: INTERSPEECH. Lisbon, Portugal, pp 805–808
110. Schuller B, Villar RJ, Rigoll G, Lang MK (2005) Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In: ICASSP. Philadelphia, Pennsylvania, USA, pp 325–328
111. Schuller B, Batliner A, Steidl S, Seppi D (2009) Emotion recognition from speech: putting ASR in the loop. In: ICASSP. Taipei, Taiwan, pp 4585–4588
112. Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun 53(9–10):1062–1087
113. Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A, Rigoll G (2010) Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans Affect Comput 1(2):119–131
114. Shami M, Verhelst W (2007) An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. Speech Commun 49(3):201–212
115. Shaver P, Schwartz J, kirson D, O'Connor C (1987) Emotion, knowledge: further exploration of a prototype approach. J Personal Soc Psychol 52:1061–1086
116. Sneddon I, McRorie M, McKeown G, Hanratty J (2012) The Belfast induced natural emotion database. IEEE Trans Affect Comput 3(1):32–41
117. Steidl S (2009) Automatic classification of emotion related user states in spontaneous children's speech. PhD thesis, Universität Erlangen-Nürnberg, Germany
118. Steidl S, Batliner A, Seppi D, Schuller B (2010) On the impact of children's emotional speech on acoustic and language models. EURASIP J Audio, Speech, and Music Processing
119. Stein N, Oatley K (1992) Basic emotions: theory and measurement. Cognit Emot 6:161–168
120. Sun R, Moore II E (2012) A preliminary study on cross-databases emotion recognition using the glottal features in speech. In: INTERSPEECH. Portland, USA, pp 1628–1631
121. Sun R, Moore II E, Torres JF (2009) Investigating glottal parameters for differentiating emotional categories with similar prosodics. In: ICASSP. Taipei, Taiwan, pp 4509–4512
122. Sundberg J, Patel S, Bjorkner E, Scherer KR (2011) Interdependencies among voice source parameters in emotional speech. IEEE Trans Affect Comput 2(3):162–174
123. Tahon M, Degottex G, Devillers L (2012) Usual voice quality features and glottal features for emotional valence detection. In: Speech Prosody. Shanghai, China, pp 693–696
124. Titze IR (1994) Principles of voice production. Prentice-Hall, Englewood Cliffs
125. Truong Khiet P, van Leeuwen David A, de Jong Franciska M G (2012) Speech-based recognition of self-reported and observed emotion in a dimensional space. Speech Commun 54(9):1049–1063

126. Ververidis D, Kotropoulos C (2003) A review of emotional speech databases. In: Proceedings of panhellenic conference on informatics (PCI). Thessaloniki, Greece, pp 560–574
127. Ververidis D, Kotropoulos C (2005) Emotional speech classification using Gaussian mixture models. In: International symposium on circuits and systems. Kobe, Japan, pp 2871–2874
128. Vlasenko B, Prylipko D, Philippou-Hübner D, Wendemuth A (2011) Vowels formants analysis allows straightforward detection of high arousal acted and spontaneous emotions. In: INTERSPEECH. Florence, Italy, pp 1577–1580
129. Vroomen J, Collier R, Mozziconacci S (1993) Duration and intonation in emotional speech. In: EUROSPEECH, vol 1. Berlin, Germany, pp 577–580
130. Švec Jan G, Schutte Harm K, Miller Donald G (1999) On pitch jumps between chest and falsetto registers in voice: data from living and excised human larynges. J Acoust Soc Am 106(3):1523–1531
131. Waaramaa T, Laukkanen AM, Airas M, Alku P (2010) Perception of emotional valences and activity levels from vowel segments of continuous speech. J Voice 24(1):30–38
132. Williams CE, Stevens KN (1969) On determining the emotional state of pilots during flight: an exploratory study. Aerosp Med 40:1369–1372
133. Williams CE, Stevens KN (1972) Emotions and speech: some acoustical correlates. J Acoust Soc Am 52(2):1238–1250
134. Wu S, Falk TH, Chan W (2011) Automatic speech emotion recognition using modulation spectral features. Speech Commun 53(5):768–785
135. Yegnanarayana B, Dhananjaya N (2013) Spectro-temporal analysis of speech signals using zero-time windowing and group delay function. Speech Commun 55(6):782–795
136. Yegnanarayana B, Murty KSR (2009) Event-based instantaneous fundamental frequency estimation from speech signals. IEEE Trans Audio Speech Lang Process 17(4):614–624
137. Yeh L, Chi T (2010) Spectro-temporal modulations for robust speech emotion recognition. In: INTERSPEECH. Chiba, Japan, pp 789–792
138. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58