Chapter 2
Robust Emotion Recognition using Pitch Synchronous and Sub-syllabic Spectral Features

Abstract This chapter discusses the use of vocal tract information for recognizing the emotions. Linear prediction cepstral coefficients (LPCC) and mel frequency cepstral coefficients (MFCC) are used as the correlates of vocal tract information. In addition to LPCCs and MFCCs, formant related features are also explored in this work for recognizing emotions from speech. Extraction of the above mentioned spectral features is discussed in brief. Further extraction of these features from sub-syllabic regions such as consonants, vowels and consonant-vowel transition regions is discussed. Extraction of spectral features from pitch synchronous analysis is also discussed. Basic philosophy and use of Gaussian mixture models is discussed in this chapter for classifying the emotions. The emotion recognition performance obtained from different vocal tract features is compared. Proposed spectral features are evaluated on Indian and Berlin emotion databases. Performance of Gaussian mixture models in classifying the emotional utterances using vocal tract features is compared with neural network models.

2.1 Introduction

In Chap. 1, we have introduced the topic of emotion recognition from speech using emotion specific speech features. In this chapter, we intend to discuss the use of vocal tract system features for speech emotion recognition. Features extracted from vocal tract system are generally known as spectral or system features. They are also sometimes referred to as segmental features, as they are normally extracted from the speech segments of 20–30 ms. MFCCs (Mel frequency cepstral coefficients), LPCCs (Linear prediction cepstral coefficients), perceptual linear prediction coefficients (PLPCs) are widely known spectral features used in the literature [1].

Generally, spectral features are found to be robust for various speech tasks. This may be due to the accurate representation of vocal tract system characteristics by spectral features. Figure 2.1 shows the unique spectral characteristics for 8 emotions

K. S. Rao and S. G. Koolagudi, *Robust Emotion Recognition using Spectral and Prosodic Features*, SpringerBriefs in Speech Technology,
DOI: 10.1007/978-1-4614-6360-3_2, © The Author(s) 2013

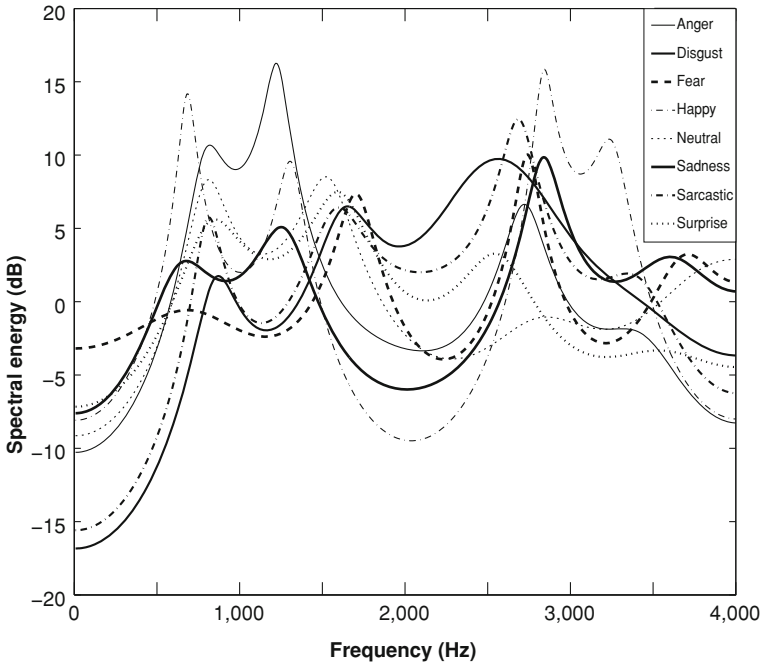

Fig. 2.1 Spectra of a steady region of vowel /A/, taken from the utterance *anni dAnamulalo vidyA dAnamu minnA*, in 8 emotions

of IITKGP-SESC. The spectra shown in Fig. 2.1 represent the steady region of a vowel /A/ from Telugu utterance *anni dAnamulalo vidyA dAnamu minnA*, expressed in eight different emotions. It is observed from the figure that the sharpness of the formant peaks, positions of the formants, formant bandwidths, and spectral tilt have distinctive properties for different emotions. In the literature, spectral features are used for modeling the pitch pattern of the speaker [2]. Variation of the pitch is an important correlate of emotions in speech. It is also known that variation in the pitch leads to changes in other prosodic parameters like duration and energy. Intuitively it may be harder to correlate spectral features with the temporal dynamics in prosody, related to emotional states. However they provide a detailed description of varying speech signal, including variations in prosody [3]. Our intuition here is that spectral features capture not only the information about what is being said (text) but also how it is being said (emotion). Therefore, spectral features are used in this study to recognize speech emotions.

The majority of the works on speech emotion recognition are carried out using spectral features derived from the whole speech signal through a conventional block processing approach. In this chapter, we have proposed the spectral features from sub-syllabic regions of the speech for recognizing the emotions. Generally, a syllable consists of a vowel optionally preceded or/and followed by consonants (this of course,

varies across languages). Vowel is treated as the nucleus of the syllable. In this context, we have used vowel onset points (VOPs) to identify consonant, vowel and consonant-to-vowel transition regions from a CV unit. In this work, these three regions are referred to as sub-syllabic regions. Features extracted from these three regions are used to recognize the emotions from a speech utterance. Most of the speech signal contains vowel regions. Steady portions of the vowel, mostly contain redundant spectral information. In this context, there is no need to process entire speech signal for feature extraction, to perform speech emotion recognition. This chapter will show that the approach of processing only useful portions from speech signal for feature extraction increases the desired performance of the system and reduces the time and computational complexities.

In case of a block processing approach, the speech signal is assumed to be stationary within a considered block of 20 ms, which includes multiple pitch cycles. In reality, the speech signal is not stationary, as it is the outcome of a time varying vocal tract system excited by the time varying excitation source. Even two consecutive cycles of a speech signal are slightly different. Generally, manifestation of emotions through speech is observed to be a gradual process, leading to emotion-specific modulation, during production of speech. Hence, instead of using the average spectrum over multiple pitch cycles, the spectrum derived from each pitch cycle is considered for feature extraction. Variations in instantaneous pitch information may also be captured using pitch synchronous spectral features. Therefore analyzing spectral features of each cycle (pitch period) of a speech signal may provide more emotion distinctive information than the emotion-specific information obtained from a block processing approach. This approach of analyzing a speech signal considering one pitch cycle each time is known as pitch synchronous analysis. Specifically no evidence is found in the literature, for pitch synchronously analyzing the speech signal to recognize emotions. In this chapter, pitch synchronously extracted spectral features are explored for capturing emotion-specific information. The proposed approach is verified using two databases: IITKGP-SESC and Emo-DB. Details of these databases are explained in Sect. 2.2.

The remaining part of the chapter is organized as follows, Sect. 2.2 briefs about the two databases used in this work. Extraction of different types of spectral features mentioned above is discussed in Sect. 2.3. Section 2.4 discusses the methodology of Gaussian mixture models used to develop emotion recognition models. Emotion recognition results are discussed in Sect. 2.5. Chapter concludes with a summary.

2.2 Emotional Speech Corpora

In this work, the Indian Institute of Technology Kharagpur- Simulated Emotional Speech Corpus (IITKGP-SESC) and Berlin emotional speech database (Emo-DB) are used to analyze the performance of proposed spectral features for Emotion Recognition (ER).

2.2.1 *Indian Institute of Technology Kharagpur-Simulated Emotional Speech Corpus: IITKGP-SESC*

This database is particularly designed and developed at the Indian Institute of Technology, Kharagpur, to support the study on speech emotion recognition. The proposed speech database is the first one developed in an Indian language (Telugu), for analyzing the common emotions present in day-to-day conversations. This corpus is sufficiently large to analyze the emotions in view of speaker, gender, text and session variability.

The corpus is recorded using 10 (5 male and 5 female) professional artists from All India Radio (AIR) Vijayawada, India. The artists are sufficiently experienced in expressing the desired emotions from the neutral sentences. All the artists are in the age group of 25–40 years, and have the professional experience of 8–12 years. For analyzing the emotions we have considered 15 semantically neutral Telugu sentences. Each of the artists has to speak the 15 sentences in 8 given emotions in one session. The number of sessions considered for preparing the database is 10. The total number of utterances in the database is 12000 (15 *sentences* $\times$ 8 *emotions* $\times$ 10 *artists* $\times$ 10 *sessions*). Each emotion has 1500 utterances. The numbers of words and syllables in the sentences are varying from 3–6 and 11–18 respectively. The total duration of the database is around 7 h. The eight emotions considered for collecting the proposed speech corpus are: Anger, Disgust, Fear, Happiness, Neutral, Sadness, Sarcasm and Surprise. The speech samples are recorded using SHURE dynamic cardioid microphone C660N. The distance between the microphone and the speaker is maintained approximately around 3–4 inches. The speech signal is sampled at 16 kHz, and each sample is represented as a 16-bit number. The sessions are recorded on alternate days to capture the inevitable variability in the human vocal tract system. In each session, all the artists have given the recordings of 15 sentences in 8 emotions. The recording is done in such a way that each artist has to speak all the sentences at a stretch in a particular emotion. This provides the coherence among the sentences for each emotion category. The entire speech database is recorded using a single microphone and at the same location. The recording was done in a quiet room, without any obstacles in the recording path [4].

The quality of the database is also evaluated using subjective listening tests. Here, the quality represents how well the artists simulated the emotions from the neutral text. The subjects are used to assess the naturalness of the emotions embedded in speech utterances. This evaluation is carried out by 25 post graduation and research students of the Indian Institute of Technology, Kharagpur. This subjective listening test is useful for the comparative analysis of emotions in a human versus machine perspective. In this study, 40 sentences (5 sentences from each emotion) randomly selected from male and female speakers are considered for evaluation. Before taking the test, the subjects are given the pilot training by playing 8 sentences (a sentence from each emotion) from each artist's speech data, for understanding (familiarizing) the characteristics of emotion expression. Forty sentences used in this evaluation are randomly ordered, and played to the listeners. For each sentence, the listener has to

Table 2.1 Emotion classification performance based on subjective evaluation

	Male Artist								Female Artist							
	A	D	F	H	N	Sa	S	Sur	A	D	F	H	N	Sa	S	Sur
Anger	73	17	2	3	4	0	0	1	69	19	3	2	5	0	0	2
Disgust	28	56	7	0	4	0	5	0	40	44	5	0	3	2	3	3
Fear	7	6	49	0	8	19	1	10	6	8	37	2	7	25	1	14
Happiness	0	2	6	62	8	9	5	6	0	4	4	66	10	7	3	6
Neutral	0	5	0	6	86	2	0	1	1	8	1	6	83	0	1	0
Sadness	0	3	16	3	13	61	4	0	4	2	25	1	12	52	3	1
Sarcasm	6	5	0	0	0	4	85	0	4	5	0	0	0	3	88	0
Surprise	0	7	5	16	5	5	7	55	6	6	3	17	3	16	1	48

A Anger, *D* Disgust, *F* Fear, *H* Happiness, *N* Neutral, *Sa* Sadness, *S* Sarcasm, *Sur* Surprise

mark the emotion category from the set of 8 given emotions. The overall emotion classification performance for male and female speech data is given in Table 2.1. The observation shows that the average emotion recognition rates of male and female speech utterances are 61 and 66 % respectively.

In this book, emotion recognition studies have been carried out in three ways, accordingly three data sets are derived from IITKGP-SESC. They are (1) Set1: Session independent speech emotion recognition, (2) Set2: Text independent speech emotion recognition, and (3) Set3: Speaker and text independent speech emotion recognition. Set1 is used to analyze the emotion recognition in view of session variability. Here 8 sessions of all speakers' speech data is used for training the emotion models, and the remaining 2 sessions of all speakers' speech data is used for testing. Set2 is used to study emotion recognition in view of text independent speech data. Here, 8 sessions of all speakers' speech data containing the first 10 text prompts are used for training, and while testing the remaining 5 text prompts of the last 2 sessions from all the speakers' speech data are used. Set3 is used to analyze the emotion recognition with respect to speaker and text independent speech data. Here, training is performed with 8 speakers' (4 males and 4 females) speech data, from all 10 sessions. Testing is performed with the remaining 2 speakers' (one male and one female) speech data. To realize the text independent data, during training the speech utterances corresponding to the first 10 text prompts of the database are used, and the remaining 5 text prompts are used while testing. Set3 is designed to have text and speaker independent properties and is more generalized than the other 2 sets. Therefore the majority of the results discussed in this book are derived using the Set3 dataset. Brief results of Set1 and Set2 are discussed at the end of the respective chapters. Table 2.2 shows the details of the three datasets used in this work.

2.2.2 Berlin Emotional Speech Database: Emo-DB

F. Burkhardt et al. have collected the actor-based simulated-emotion Berlin database in the German language [5]. Ten (5 male + 5 female) actors have contributed in preparing the database. The emotions recorded in the database are anger, boredom,

Table 2.2 Details of the datasets derived from IITKGP-SESC, for various studies on speech emotion recognition

Data set	Purpose and description	Training data	Testing data
Set1	Session independent emotion recognition	The utterances of all 15 text prompts, recorded from 10 speakers are used in training. Out of 10 sessions, 8 (1–8) sessions of each speaker are used in training.	The utterances of all 15 text prompts, recorded from 10 speakers are used in testing. Out of 10 sessions, 2 (9 and 10) sessions of each speaker are used in testing.
Set2	Session and text independent emotion recognition	Out of 15 text prompts, the utterances of 10 (1–10) prompts, recorded from 10 speakers are used in training. Out of 10 sessions, 8 (1–8) sessions of each speaker are used in training.	Out of 15 text prompts, the utterances of 5 (11–15) prompts, recorded from 10 speakers are used in testing. Out of 10 sessions, 2 (9 and 10) sessions of each speaker are used in testing.
Set3	Session, text, and speaker independent emotion recognition	Out of 15 text prompts, the utterances of 10 (1–10) prompts, recorded from 8 (4 males and 4 females) speakers are used in training. All 10 sessions of each speaker are used in training.	Out of 15 text prompts, the utterances of 5 (11–15) prompts, recorded from 2 (a male and a female) speakers are used in testing. All 10 sessions of each speaker are used in testing.

disgust, fear, happiness, neutral and sadness. Ten linguistically neutral German sentences are chosen for database construction. The database is recorded using the Sennheiser MKH 40 P48 microphone, with the sampling frequency of 16 kHz. Samples are stored as 16 bit numbers. Eight hundred and forty (840) utterances of Emo-DB are used in this work.

In the case of the Berlin database, 8 speakers' speech data is used for training the models and the remaining 2 speakers' speech data is used for validating the trained models.

2.3 Feature Extraction

In this work, LPCCs, MFCCs and formant features are used for representing the spectral information. Sub-syllabic spectral features are derived from the speech segments of consonants, vowels, and consonant to vowel transitions. Pitch synchronous spectral features are derived from each pitch cycle of the speech signal. Extraction of different spectral features, mentioned above is discussed in the following subsections.

2.3.1 Linear Prediction Cepstral Coefficients (LPCCs)

The cepstral coefficients derived from either linear prediction (LP) analysis or a filter bank approach are almost treated as standard front end features. Speech systems developed based on these features have achieved a very high level of accuracy, for speech recorded in a clean environment. Basically, spectral features represent phonetic information, as they are derived directly from spectra. The features extracted from spectra, using the energy values of linearly arranged filter banks, equally emphasize the contribution of all frequency components of a speech signal. In this context, LPCCs are used to capture emotion-specific information manifested through vocal tract features. In this work, the 10th order LP analysis has been performed, on the speech signal, to obtain 13 LPCCs per speech frame of 20 ms using a frame shift of 10 ms. The human way of emotion recognition depends equally on two factors, namely: its expression by the speaker as well as its perception by a listener. The purpose of using LPCCs is to consider vocal tract characteristics of the speaker, while performing automatic emotion recognition [6].

Cepstrum may be obtained using linear prediction analysis of a speech signal. The basic idea behind linear predictive analysis is that the nth speech sample can be estimated by a linear combination of its previous p samples as shown in the following equation.

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + a_3 s(n-3) + \cdots + a_p s(n-p)$$

where $a_1, a_2, a_3 \cdots$ are assumed to be constants over a speech analysis frame. These are known as predictor coefficients or linear predictive coefficients. These coefficients are used to predict the speech samples. The difference of actual and predicted speech samples is known as an error. It is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k)$$

where $e(n)$ is the error in prediction, $s(n)$ is the original speech signal, $\hat{s}(n)$ is a predicted speech signal, $a_k s$ are the predictor coefficients.

To compute a unique set of predictor coefficients, the sum of squared differences between the actual and predicted speech samples has been minimized (error minimization) as shown in the equation below

$$E_n = \sum_{m} \left[s_n(m) - \sum_{k=1}^{p} a_k s_n(m-k) \right]^2$$

where m is the number of samples in an analysis frame. To solve the above equation for LP coefficients, E_n has to be differentiated with respect to each a_k and the result is equated to zero as shown below

$$\frac{\partial E_n}{\partial a_k} = 0, \qquad \text{for } k = 1, 2, 3, \cdots, p$$

After finding the $a_k s$, one may find cepstral coefficients using the following recursion.

$$C_0 = \log_e p$$

$$C_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} C_k a_{m-k}, \qquad \text{for } 1 < m < p \text{ and}$$

$$C_m = \sum_{k=m-p}^{m-1} \frac{k}{m} C_k a_{m-k}, \qquad \text{for } m > p$$

2.3.2 Mel Frequency Cepstral Coefficients

A human auditory system is assumed to process a speech signal in a nonlinear fashion. It is well known that lower frequency components of a speech signal contain more phoneme specific information. Therefore a nonlinear mel scale filter bank has been used to emphasize lower frequency components over higher ones. In speech processing, the mel frequency cepstrum is a representation of the short term power spectrum of a speech frame using a linear cosine transform of the log power spectrum on a nonlinear mel frequency scale. Conversion from normal frequency f to mel frequency m is given by the equation

$$m = 2595 \log_{10}\left(\frac{f}{700} + 1\right)$$

The steps used for obtaining mel frequency cepstral coefficients (MFCCs) from a speech signal are as follows:

1. Pre-emphasize the speech signal.
2. Divide the speech signal into a sequence of frames with a frame size of 20 ms and a shift of 5 ms. Apply the hamming window over each of the frames.
3. Compute the magnitude spectrum for each windowed frame by applying DFT.
4. Mel spectrum is computed by passing the DFT signal through a mel filter bank.
5. DCT is applied to the log mel frequency coefficients (log mel spectrum) to derive the desired MFCCs.

Twenty filter banks are used to compute 8, 13 and 21 MFCC features from a speech frame of 20 ms, with 5 ms overlap each time. The purpose of using MFCCs is to take the listener's non-linear auditory perceptual system into account, while performing automatic emotion recognition.

2.3.3 Formant Features

Cepstral coefficients are used as standard front end features for developing various speech systems, however, they perform poorly with noisy or real life speech. Therefore the supplementary features along with basic cepstral coefficients are essential to handle real life speech. The higher amplitude regions of a spectrum, such as formants, are relatively less affected by the noise. K. K. Paliwal et al. have extracted spectral sub-band centroids from high amplitude regions of the spectrum and used for noisy speech recognition [7]. With this viewpoint, formant parameters are used in this study as the supplementary features to cepstral features. Also note that the conventional cepstral features utilize only amplitude (energy) information from the speech power spectrum, whereas the proposed formant features utilize frequency information as well.

In general, formant tracks represent the sequences of vocal tract shapes, hence formant analysis using their strength, location and bandwidth may help to extract vocal tract related emotion specific information from a speech signal. Figure 2.2 shows different spectra for 8 emotions of IITKGP-SESC. The spectra are derived from the syllable *tha* from Telugu sentence *thallidhandrulanu gauravincha valenu*. In this case, the language, text, speaker and contextual information is maintained the same. This is speculative from the figure that the variation in the spectra is due to the emotions. Formant frequencies are very crucial in view of speech perception. Hence, a slight change in these parameters causes a perceptual difference, which may lead to manifestation of different emotions. It is evident from Fig. 2.2 that the position and strength of formants are clearly distinct for different emotions. Spectral peaks indicate the intensity of specific frequency components (or frequency band). Their distinctive nature for different emotions is the indication of presence of emotion specific information. The rate of decrease in spectrum amplitude, as a function of frequency is known as spectral roll-off or spectral tilt. This happens mainly because of decreasing strength of harmonics, as the frequency increases. A speaker can induce more strength into higher harmonics by consciously controlling the glottal vibration. Abrupt closing of the glottis increases the energy in the higher frequency components. This leads to the variation in spectral roll-off for different emotions. Figure 2.2 shows distinct spectral roll-offs for each of the emotions.

Though it is assumed that the bandwidth of a formant does not influence phonetic information [6], it represents some speaker specific information. Figure 2.2 depicts the variation in the formant bandwidths in case of different emotions. Even a slight variation in the bandwidth may be due to speaker induced emotion specific information as speaker, text, language and context related information do remain the same. Formant bandwidth is the frequency band measured at around 3 dB downward from the respective formant peak.

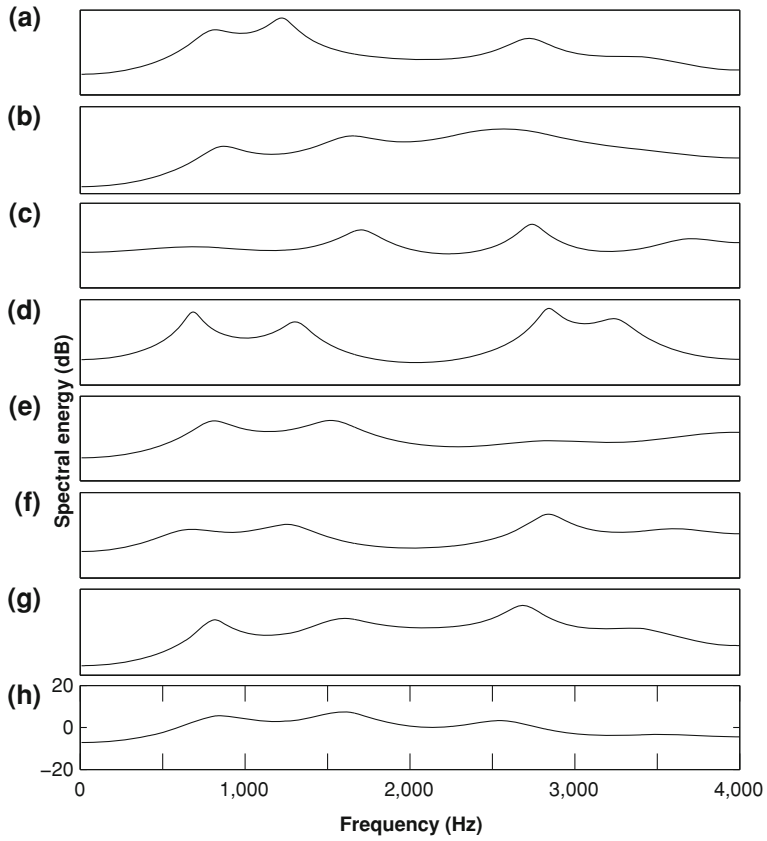

Fig. 2.2 Spectra for the steady region of the syllable /tha/ from the utterance *thallidhandrulanu gauravincha valenu*. **a** Anger, **b** Disgust, **c** Fear, **d** Happy, **e** Neutral, **f** Sadness, **g** Sarcastic, and **h** Surprise

2.3.4 Extraction of Sub-syllabic Spectral Features

In the context of Indian languages, a syllable can be viewed as the combination of consonants and a vowel, in the form $C^m V C^n$, where C is the consonant and V is the vowel such that, $m, n \geq 0$ and $m, n \leq 3$. In the syllable, the vowel is treated as the nucleus and consonants may or may not be present. In the context of Indian languages, most common syllable forms are CV, CCV, CCVC, and CVC. Among these forms, more than 90 % of the syllables are of the type CV. In a CV unit, the speech signal to the left of the VOP (before VOP) is a consonant region and to the right of it is the vowel region [8]. Vowel onset point as a junction point between the consonant and vowel of a CV unit. At this point the characteristics of the consonant segment are terminated and the characteristics of vowel are originated. Hence, it is important to extract features around this crucial point [9]. After determining the VOP, 40 ms to the left of the VOP

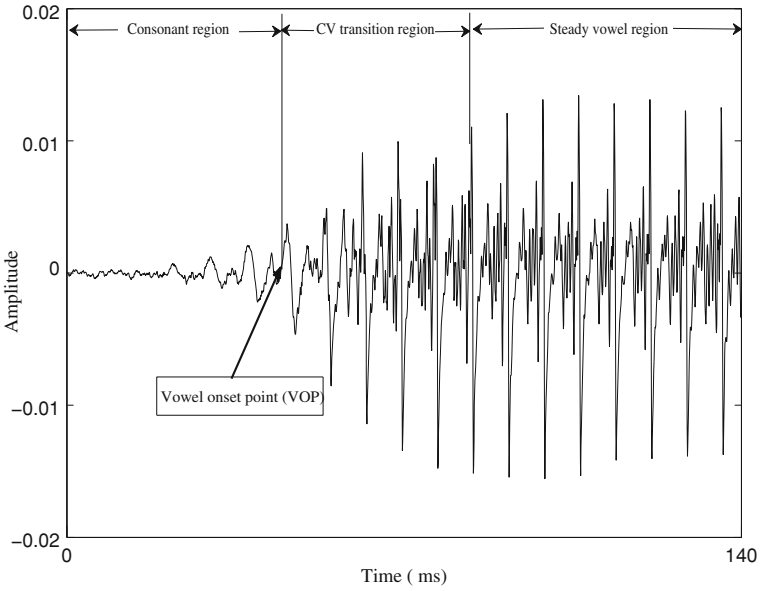

Fig. 2.3 Vowel onset point (consonant, vowel and CV transition regions are marked using vowel onset point) of a CV unit

is marked as the consonant region, and to the right of the VOP is the vowel region. In the vowel region, a small region of about 30 ms, following the VOP, is treated as transition region. After the transition region, 60 ms of speech signal is considered as the steady vowel region. Figure 2.3 shows the identified consonant, vowel, and CV transition regions from a typical CV unit. In this study, consonant, vowel and CV transition regions of a syllable are referred to as sub-syllabic speech regions. The features are extracted from vowel, consonant and transition regions, to analyze speech emotions. Processing complexity may be considerably reduced by avoiding processing of redundant information present in the speech regions such as the steady portion of the vowels. The purpose of this analysis is to study the contribution of different sub-syllabic regions toward emotion-specific information and also to avoid processing redundant speech regions from feature extraction. Therefore in this work, spectral (LPCCs, MFCCs and formant) features are extracted from consonant, vowel and CV transition regions of each syllable separately, using VOP locations as the anchor points.

VOP detection method employed in this work, uses the combination of measures from excitation source, spectral peaks, and modulation spectrum. This method is known as the combined method for the detection of VOP. Excitation source information is represented using the Hilbert envelope (HE) of the linear prediction (LP) residual. A sequence of the sum of the ten largest peaks of the spectra of speech frames represents the shape of the vocal tract. The slowly varying temporal envelope of the speech signal can be represented using a modulation spectrum. Each of these

three features represents supplementary information about the VOP, and hence they are combined for the enhancement in the performance of VOP detection. VOP detection using the combined method is carried out with the following steps: (1) Derive the VOP evidence from the excitation source, spectral peaks, and modulation spectrum. Here, the evidence from excitation source information is obtained from the Hilbert envelope of the linear prediction residual signal. The evidence from the spectral peaks is obtained by summing the ten largest spectral peaks of each speech frame. The evidence due to the modulation spectrum is derived by passing the speech signal through a set of critical band pass filters, and summing the components corresponding to 4–16 Hz. (2) The above evidence is further enhanced by computing the slope with the help of a first order difference (FOD). (3) These enhanced parameters are convolved with the first order Gaussian difference (FOGD) operator for deriving the final VOP evidence. (4) Individual VOP parameters derived from the excitation source, spectral peaks and modulation spectrum are combined to provide the robust VOP evidence plot. (5) The positive peaks in the combined VOP evidence signal are hypothesized as the locations of VOPs [10]. Figure 2.4 shows the intermediate steps of VOP detection using the evidence from excitation source, spectral and modulation spectrum energy plots. More than 90 % of the automatically located VOPs are observed to be accurate with a maximum tolerance of 40 ms [10]. The sentence *Dońt ask me to carry an oily rag like that*, chosen from TIMIT database, is used for illustrating the automatic detection of VOPs in Fig. 2.4. From the figure, it is observed that, the detected VOPs are close to the manually marked VOPs (See Figs. 2.4a, e).

The crucial part in CV units (syllable) is a region that represents transition from consonant to vowel. Figure 2.5 indicates the spectra obtained from the CV transition region of 40 ms length, for different emotions. The text, gender, speaker and contextual information is kept the same. From the figure, it may be observed that, except for the first formant, there is a clear distinction in the spectral characteristics for different emotions. For formant positions and energies are varying with respect to the emotions. Therefore, spectral characteristics derived from the CV transition regions of the speech signal are useful for discriminating the emotions. However, there is not much variation in the spectral characteristics of a speech signal in the steady portion of vowel. Consonant portion seem like noise containing frequency components with lower energy. The parameters extracted from these regions may not be much discriminative while capturing emotion-specific vocal tract characteristics. Figure 2.6 shows the spectra obtained from vowel, transition and consonant regions. It is clearly observed from the figure that the spectral behavior of a speech signal is mostly redundant in the case of vowels. Each frame in the transition region displays different spectral characteristics. The consonant region does not even show the formant structure, and energy of the spectral peaks is also very much less. To study this phenomenon, in this work, we have extracted LPCCs, MFCCs and formant features from 40 ms of each of CV transition and constant regions. Similarly, 60 ms of speech is considered from the steady vowel regions to extract the features.

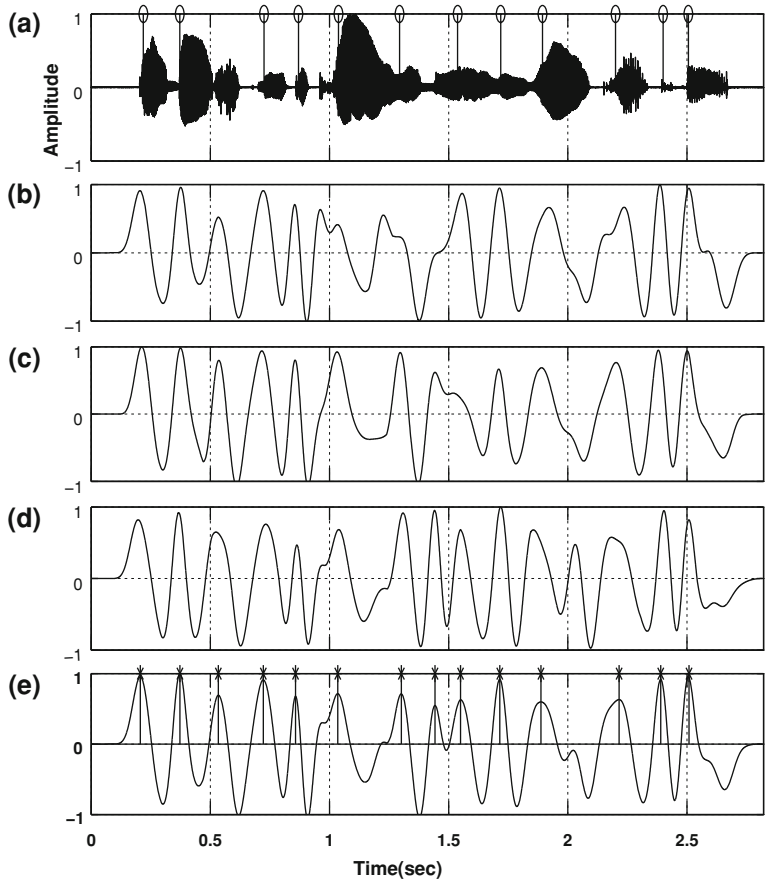

Fig. 2.4 Identified VOPs using the combined approach of excitation source energy, spectral peak energy, and modulation spectrum energy. **a** Speech signal with manually marked VOPs. **b** Evidence plot using excitation source energy. **c** Evidence plot using spectral peak energies. **d** Evidence plot using modulation spectrum energies. **e** Combined evidence plot

2.3.5 Pitch Synchronous Analysis

Analysis of the speech signal with respect to each pitch period is known as pitch synchronous analysis. In a normal block processing approach, physical decomposition of speech signal into the segments of length 20 ms is performed. In case of pitch synchronous analysis, logical decomposition of the speech signal, to cover one or multiple pitch cycles, is considered for feature extraction. This approach helps to derive logically related feature vectors. Variation in the speech signal between the consecutive pitch cycles is captured through pitch synchronous analysis.

Figure 2.7 shows the variation in spectral properties obtained from the consecutive pitch cycles of a voiced speech segment. From the gross observation, the spectra

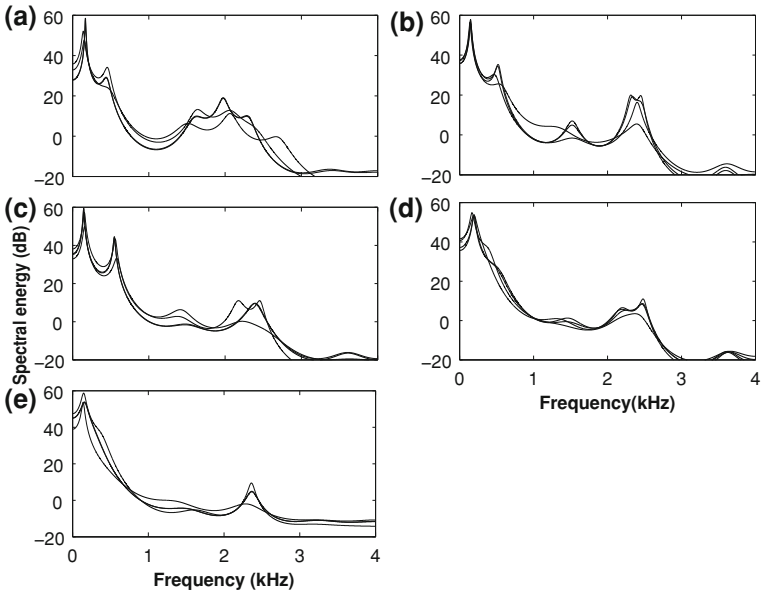

Fig. 2.5 Spectra of CV transition regions for 5 emotions. **a** Anger, **b** Fear, **c** Happiness, **d** Neutral and **e** Sadness

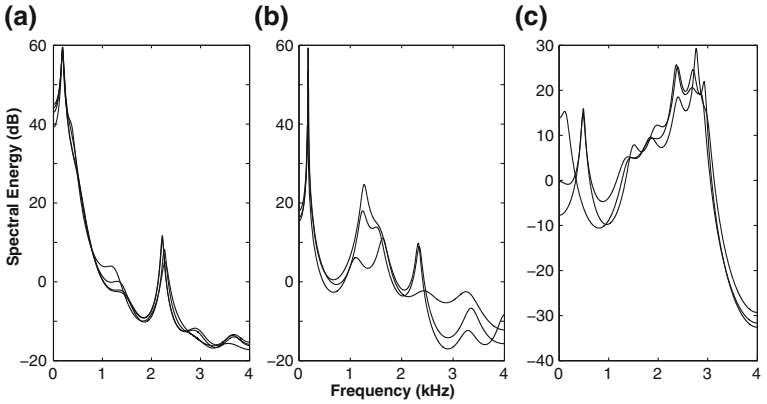

Fig. 2.6 Spectral properties of vowel, CV transition and consonant regions in the continuous speech. **a** Spectra for steady portion of the vowel, **b** Spectra for CV transition region and **c** Spectra for the consonant region

appear to be similar (close to each other), but at the finer level one can observe the variations in formant strengths and bandwidths in a successive pitch cycles. The sequence of finer variations may provide the desired emotion discrimination. In the literature 3–4 consecutive pitch cycles are considered for feature extraction [11, 12], but in this study, every pitch cycle of entire voiced region of the utterances is

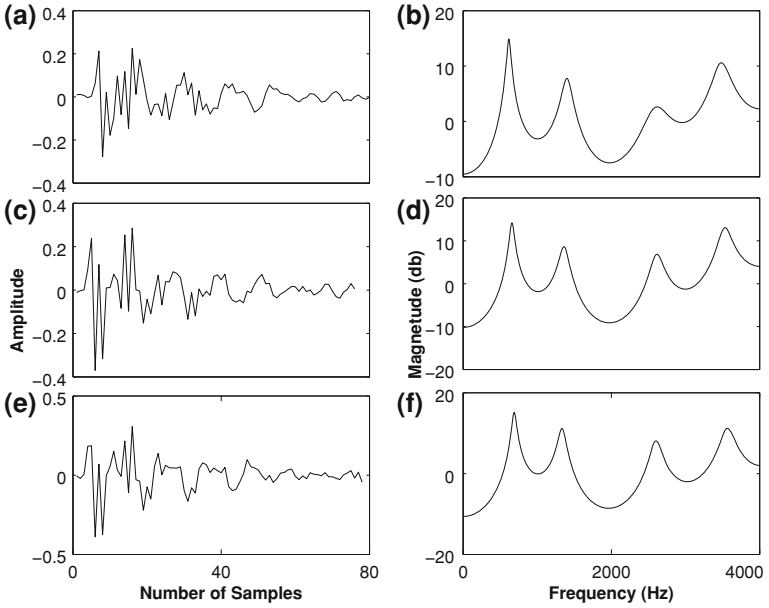

Fig. 2.7 Pitch synchronous analysis. **a**, **c** and **e** Three consecutive pitch cycles of the speech signal. **b**, **d** and **f** Corresponding spectra

independently processed for extracting the spectral features. Some of the important intuitions behind using pitch synchronous analysis of speech signals are as follows: The illogical approach associated with physical framing of the speech signal practiced in block processing, can be eliminated by carrying the analysis of speech signal within a pitch period. The assumption that a speech signal is stationary within the frame of 20 ms is not completely acceptable as both source and system are continuously varying with respect to time. In this work, pitch periods are marked using glottal closure instants (GCIs). A signal between two consecutive GCIs is treated as one pitch cycle. A zero frequency filter based method is used to determine the GCIs [13]. LPCCs, MFCCs and formant features are computed for each pitch cycle of a speech signal.

2.4 Classifiers

GMMs and AANNs are known to capture the general distribution of data points in the feature space and one can be used as an alternative to the other [14]. Two classifiers are used in this study, to mutually compare their emotion classification results.

2.4.1 Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMMs) are among the most statistically mature methods for clustering. A Gaussian mixture model is used as a classification tool in this task. They model the probability density function of observed variables using a multivariate Gaussian mixture density. Given a series of inputs, a GMM refines the weights of each distribution through the expectation-maximization algorithm. Mixture models are a type of density model, which comprise a number of component functions, usually Gaussian in nature. These component functions are combined to provide a multi-modal density. Mixture models are a semi-parametric alternative to non-parametric models and provide greater flexibility and precision in modeling the underlying statistics of sample data. They are able to smooth over gaps resulting from sparse sample data and provide tighter constraints in assigning object membership to cluster regions. Once a model is generated, conditional probabilities can be computed for test patterns (unknown data points). An expectation maximization (EM) algorithm is used for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables.

Expectation Maximization is an iterative method that alternates between performing an expectation (E) step, which computes an expectation of the log likelihood with respect to the current estimate of the distribution for the latent variables, and a maximization (M) step, which computes the parameters that maximize the expected log likelihood found on the E step. These parameters are then used to determine the distribution of the latent variables in the next E step.

The number of Gaussians in the mixture model is also known as the number of components. They indicate the number of clusters in which data points are to be distributed in order to cover local variations. In this work, one GMM is developed to capture the information about one emotion. Depending on the number of training data points, the number of components may be varied in each GMM. The presence of few components in a GMM, and training it with large number of data points may lead to more generalized clusters, failing to capture specific details related to each class. On the other hand over-fitting of the data points may happen, if too many components represent few data points. Obviously the complexity of the model increases, if they contain higher numbers of components. Therefore a tradeoff has to be reached between the complexity and the accuracy of the classification results required. In this work, GMMs are designed with 64 components and iterated 30 times to attain convergence. A diagonal covariance matrix is used to derive the model parameters.

2.4.2 Auto-Associative Neural Networks

AANN models are basically feed-forward neural network (FFNN) models, which try to map an input vector onto itself, and hence the name auto-association or identity

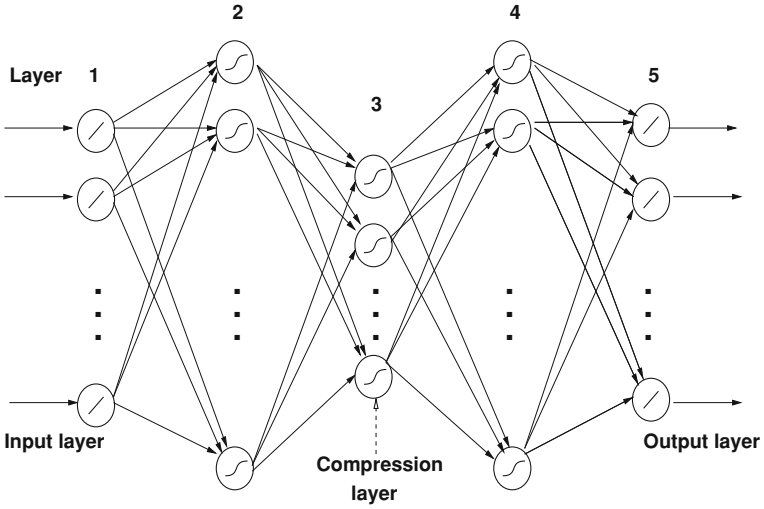

Fig. 2.8 Five layer auto-associative neural network

mapping [15, 16]. It consists of an input layer, an output layer and one or more hidden layers. The number of units in the input and output layers is equal to the dimension of the input feature vectors. The number of nodes in one of the hidden layers is less than the number of units in either the input or output layer. This hidden layer is also known as the dimension compression layer. The activation function of the units in the input and output layers is linear, whereas in case of hidden layers it is either linear or nonlinear. A five-layer AANN model with the structure shown in Fig. 2.8 is used in this study. The structure of network is generally represented by PL-QN-RN-QN-PL, where P, Q and R refers to integer values, L refers to linear units and N to nonlinear units. The integer value indicates the number of units present in that layer. The number of linear elements at the input layer indicates, the size of the feature vectors used for developing the models. The nonlinear units use $tanh(s)$ as the activation function, where s is the net input value of that unit. The structure of the network was determined empirically.

The performance of AANN models can be interpreted in different ways, depending on the application and the input data. If the data is a set of feature vectors in the feature space, then the performance of AANN models can be interpreted as linear or nonlinear principal component analysis (PCA) or capturing the distribution of input data [17–19]. On the other hand, if the AANN is presented directly with signal samples, such as LP residual signal, the network captures the implicit linear/nonlinear relations among the samples [20–22].

Determining the network structure is an optimization problem. At present there are no formal methods for determining the optimal structure of a neural network. The key factors that influence the neural network structure are learning ability of a network and capacity to generalize the acquired knowledge. From the available

literature, it is observed that 5 layer symmetric neural networks, with three hidden layers have been used for different speech tasks. The first and the third hidden layers have more number of nodes than the input or output layer. The middle layer (also known as dimension compression layer) contains fewer units [23, 24]. In this type of network, generally the first and third hidden layers are expected to capture the local information among the feature vectors and the middle hidden layer is meant for capturing global information. Most of the existing studies [23–26] have used the 5 layer AANNs with the structure $N_1L - N_2N - N_3N - N_2N - N_1L$, for their optimal performance. Here N_1, N_2, and N_3 indicate the number of units in the first, second and third layers respectively, of the symmetric 5-layer AANN. Usually N_2 and N_3 are derived experimentally, for achieving the best performance in the given task. From the existing studies, it is observed that N_2 is in the range of 1.3–2 times N_1 and N_3 is in the range of 0.2–0.6 times N_1. For designing the structure of the network, we have used the guidelines from the existing studies and experimented with few structures for finalizing the optimal structure. The performance of the network does not depend critically on the structure of the network [21, 27–29]. The number of units in the two hidden layers is guided by the heuristic arguments given above. All the input and output features are normalized to the range $[-1, +1]$ before presenting to the neural network. The back-propagation learning algorithm is used for adjusting the weights of the network to minimize the mean squared error [24].

2.5 Results and Discussion

In this work, 21 emotion recognition systems (ERS) are developed to study speech emotion recognition using different spectral features. In the beginning, emotion recognition systems are developed individually, using MFCCs, LPCCs, and formant features. Formant features alone have not given appreciably good emotion recognition performance, therefore, in the later stages, they are used in combination with the other features. Totally 5 sets of emotion recognition systems are developed as shown in Fig. 2.9. They are the ERSs developed using the spectral features derived from (a) the entire speech signal, (b) the vowel region, (c) the consonant region, (d) the CV transition region, and (e) pitch synchronous analysis. In each set, emotion recognition systems are developed using LPCCs, MFCCs, LPCCs+formant features, and MFCCs+formant features. In the following paragraphs, the emotion recognition performance of all individual emotion recognition systems, developed using Set3 of IITKGP-SESC, are discussed. Out of 10 speakers' speech data, the utterances of 8 speakers (4 male and 4 female) are used for training the ER models and the utterances of 2 (a male and a female) speakers are used for validating the trained models. Thirteen spectral features are extracted from a frame of 20 ms, with a shift of 5 ms. GMMs with 64 components are used to develop ERSs. The results of emotion recognition performance using session and text independent (Set1 and Set2) speech data are also given at the end of the chapter.

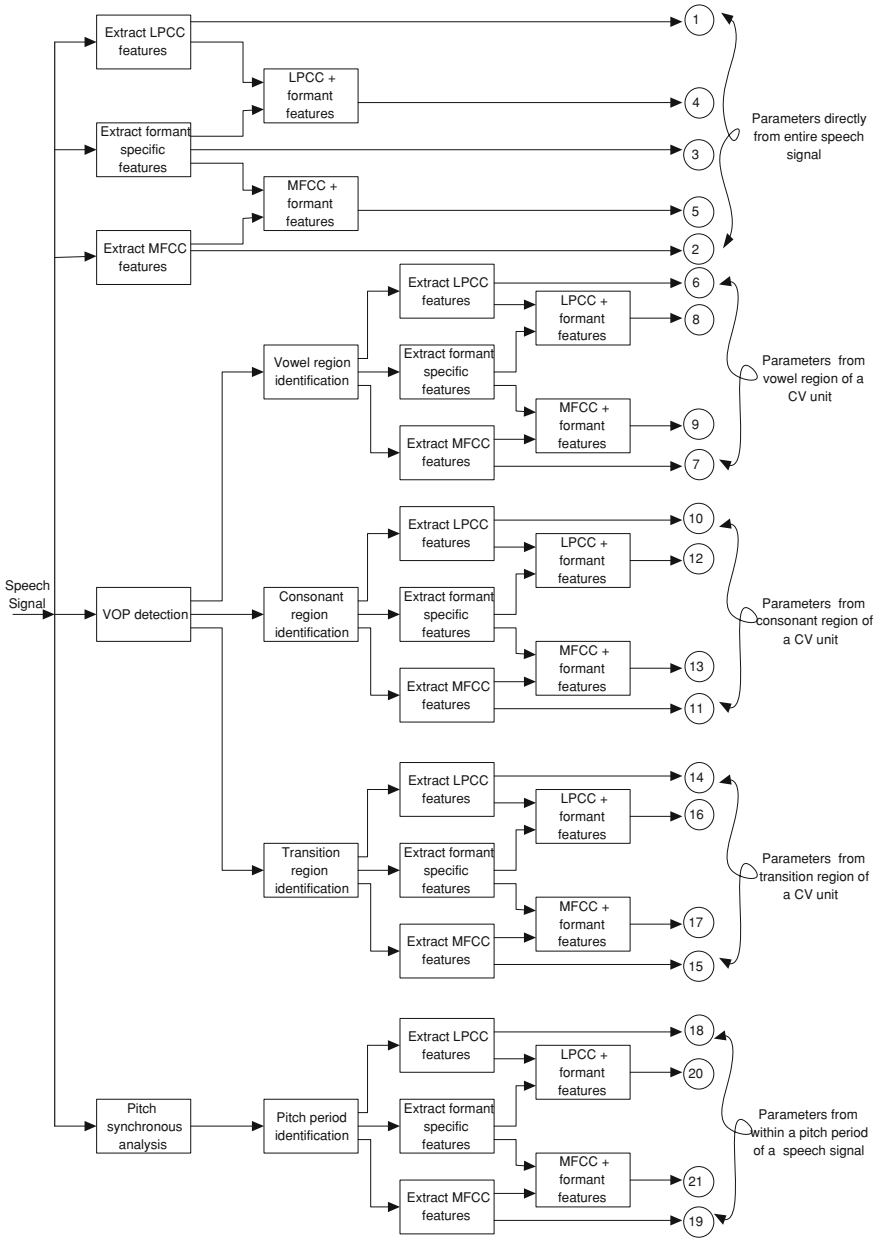

Fig. 2.9 Block diagram indicating various emotion recognition systems developed using the proposed spectral features, extracted from the entire speech utterance, sub-syllabic segments (vowel, consonant, and CV transition), and pitch synchronous analysis

Table 2.3 Emotion classification performance using the spectral features obtained from entire speech signal, adopting conventional block processing approach (ERSs 1–5)

Emotions	LPCCs (ERS1)	MFCCs (ERS2)	Formant features (ERS3)	LPCCs+ formant features (ERS4)	MFCCs+ formant features (ERS5)
	Emotion recognition performance in %				
Anger	53	57	33	60	63
Disgust	63	53	33	60	60
Fear	67	60	37	63	63
Happy	77	70	47	73	70
Neutral	80	77	66	83	77
Sadness	70	63	57	73	73
Sarcasm	73	70	64	77	60
Surprise	60	57	40	63	53
Average	68	63.38	47	69	68

ERS1 is developed using 13 LPCC features obtained from the entire speech signal using the normal block processing approach. ERS2 is developed using 13 MFCC features extracted frame wise from entire speech signal. ERS3 is developed using formant related features. These 13 formant related features (4 frequencies, 4 energy values, 4 bandwidth values and a slope), extracted per frame of 20 ms, are used to represent formant information. Their concatenation forms the 13 dimensional feature vector. The average emotion recognition performance using formant features is about 47 %. ERS4 and ERS5 are developed using the combination of 13 formant features along with 13 LPCCs and 13 MFCCs respectively. The dimension of the resulting feature vectors is 26. Table 2.3 shows the emotion recognition performance of ERS1–ERS5.

ERS6, ERS7, ERS8 and ERS9 are developed using the spectral features extracted from the vowel regions of the utterances [30]. LPCCs, MFCCs and formant features are extracted from 60 ms of speech signal, chosen from the steady portion of the vowel region of each syllable. Table 2.4 shows the performance of ERSs 6–9.

Similar to the emotion recognition systems developed for vowel regions (ERS6–ERS9), four systems are developed for consonant regions (ERS10–ERS13) and four systems are developed for CV transition regions (ERS14–ERS17) [30]. Tables 2.5 and 2.6 show the emotion recognition performance of ERSs developed using LPCC features, MFCC features, LPCCs+formant features and MFCCs+formant features, extracted from consonant and CV transition regions of the syllables, respectively.

Pitch synchronously extracted spectral features are used to capture the finer level spectral dynamics specific to the speech emotions. Therefore, spectral and formant features, extracted from each pitch period are used to develop the ERSs 18–21. Table 2.7 shows the emotion recognition performance using the spectral features extracted through pitch synchronous analysis.

Table 2.4 Emotion classification performance using the spectral features obtained from the vowel regions (ERSs 6–9)

Emotions	LPCCs (ERS6)	MFCCs (ERS7)	LPCCs+ formant features (ERS8)	MFCCs+ formant features (ERS9)
	Emotion recognition performance in %			
Anger	53	40	53	47
Disgust	50	43	53	40
Fear	50	50	57	43
Happy	47	53	50	60
Neutral	63	60	63	67
Sadness	60	57	67	50
Sarcastic	63	50	63	53
Surprise	53	47	57	47
Average	54.88	50	58	50.88

Table 2.5 Emotion classification performance using the spectral features obtained from the consonant regions (ERSs 10–13)

Emotions	LPCCs (ERS10)	MFCCs (ERS11)	LPCCs+ formant features (ERS12)	MFCCs+ formant features (ERS13)
	Emotion recognition performance in %			
Anger	37	33	40	37
Disgust	40	37	50	43
Fear	43	40	47	33
Happy	43	43	50	40
Neutral	47	50	57	53
Sadness	50	47	53	47
Sarcastic	50	53	53	47
Surprise	33	30	43	43
Average	42.88	41.63	49.13	42.88

The contribution of different features and different sub-syllabic regions toward specific emotions can be analyzed by observing individual emotion recognition performance of all ERSs. In general, combination of LPCCs and formants (using block processing and pitch synchronous analysis) are more discriminative with respect to all emotions. Specifically, the features from CV transition regions performed better in case of slow emotions like sadness and neutral. Happy is generally recognized well by most of the features. The results of the above studies show that emotion recognition performance using LPCCs is better than the results of MFCCs. LPCCs in the cases of the conventional block processing and pitch synchronous analysis

Table 2.6 Emotion classification performance using the spectral features obtained from the consonant-vowel transition regions (ERSs 14–17)

Emotions	LPCCs (ERS14)	MFCCs (ERS15)	LPCCs+ formant features (ERS16)	MFCCs+ formant features (ERS17)
	Emotion recognition performance in %			
Anger	47	57	53	57
Disgust	57	53	67	67
Fear	63	53	64	60
Happy	70	50	67	67
Neutral	77	67	77	80
Sadness	63	63	80	73
Sarcastic	67	67	73	70
Surprise	60	57	63	63
Average	63.13	58.38	68	67.13

Table 2.7 Emotion classification performance using pitch synchronously extracted spectral features (ERSs 18–21)

Emotions	LPCCs (ERS18)	MFCCs (ERS19)	LPCCs+ formant features (ERS20)	MFCCs+ formant features (ERS21)
	Emotion recognition performance in %			
Anger	57	50	60	60
Disgust	63	60	67	63
Fear	60	67	63	63
Happy	70	70	73	67
Neutral	80	77	77	77
Sadness	77	73	80	80
Sarcastic	73	63	70	77
Surprise	67	70	73	63
Average	68.38	66.25	70.38	68.75

have achieved highest emotion recognition of around 69 %. The reason for this may be that LPCCs mainly represent the speech production characteristics, by analyzing all frequency components in a uniform manner. The emotion specific information may be present across all the frequencies in a uniform way. The proposed formant features alone are not suitable to develop emotion recognition systems as their individual performance is poor. However, the combination of formant features with other spectral features has been proved to improve the recognition performance.

Emotion recognition performance using the spectral features from the consonant region is very poor. This may be due to poor representation of consonant regions by spectral features. The systems developed using only vowel regions have shown the

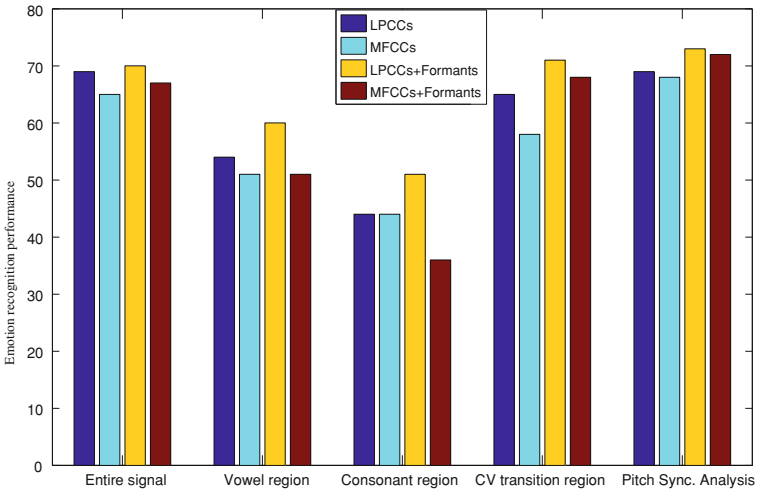

Fig. 2.10 Comparison of emotion recognition performance with respect to the entire speech signal, sub-syllabic speech segments, and pitch synchronous analysis, using proposed spectral features

average emotion recognition performance of around 58 %. The low performance of the spectral features in the vowel regions is due to the presence of redundant information. The special observation of the results presented in Tables 2.4, 2.5, 2.6 and 2.7 is that the CV transition regions contain more emotion specific information than vowel and consonant regions. The models developed using CV transition regions have achieved the performance as equal as that of a conventional block processing approach. It indicates that extracting the features from CV transition regions alone would be sufficient to recognize the underlying emotions. Processing only CV transition regions for emotion classification has two important advantages: (1) it is highly effective in the view of computational complexity and (2) achievement of almost similar performance due to selective processing of crucial information.

Compared to other spectral features, system features extracted from individual pitch cycles have performed well. The reason for this is that spectral characteristics are computed from each pitch cycle, with the intention of capturing finer spectral variations among the successive pitch cycles. It may be also noted that finer spectral variations are more emotion-specific than the frame-wise spectral information. The observations discussed above may be visualized from the bar graphs shown in Figs. 2.10 and 2.11. Figure 2.10 shows the emotion recognition performance of various spectral features extracted from the entire speech signal, sub-syllabic regions and pitch synchronous analysis. Figure 2.11 gives the emotion recognition performance by LPCC+formant features extracted from different speech regions and pitch synchronous analysis.

Different numbers of LPCCs/MFCCs are also explored for analyzing the emotion recognition performance. Table 2.8 indicates the average emotion recognition performance of 8 emotions using 8, 13 and 21 spectral features. It may be observed

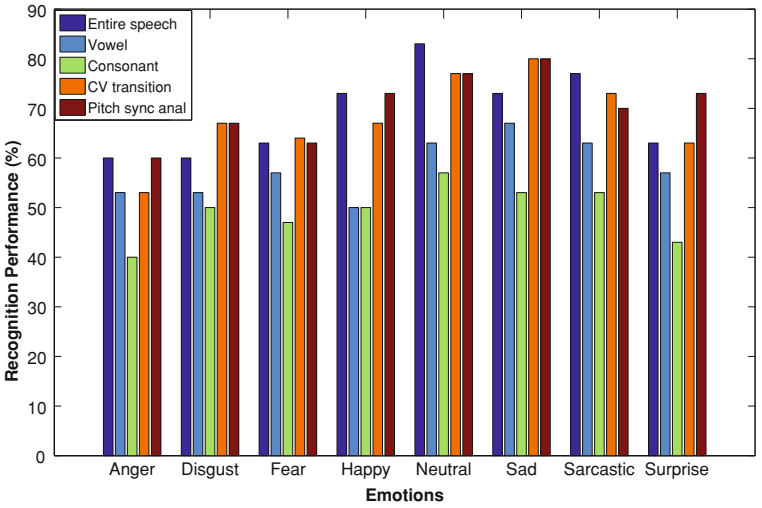

Fig. 2.11 Comparison of emotion recognition performance of the LPCCs+formant features, with respect to different emotions, obtained from entire speech signal, sub-syllabic regions, and pitch synchronous analysis

Table 2.8 Average emotion classification performance on IITKGP-SESC using different lengths of cepstral vectors

Methods & regions used for feature extraction	Features	No. of Ceps. Coeff.		
		8	13	21
		Recognition performance (in %)		
Block processing (Entire speech signal)	LPCCs	59	68	69
	MFCCs	55	63	65
	Formant features		47	
	LPCCs+Formants	58	69	70
	MFCCs+Formants	57	65	67
Vowel region	LPCCs	48	55	54
	MFCCs	42	50	51
	LPCCs+Formants	50	58	60
	MFCCs+Formants	43	51	51
Consonant region	LPCCs	35	43	44
	MFCCs	31	42	44
	LPCCs+Formants	38	49	51
	MFCCs+Formants	31	43	36

continued

Table 2.8 continued

Methods & regions used for feature extraction	Features	No. of Ceps. Coeff. 8	13	21
		Recognition performance (in %)		
CV-transition region	LPCCs	54	63	65
	MFCCs	50	58	58
	LPCCs+Formants	58	68	71
	MFCCs+Formants	55	67	68
Pitch synchronous analysis	LPCCs	56	68	69
	MFCCs	58	66	68
	LPCCs+Formants	61	70	73
	MFCCs+Formants	60	69	72

from Table 2.8 that most of the times, the systems developed using higher numbers of spectral features have performed slightly better than their counterparts developed using smaller numbers of spectral features. It may be due to the reason that higher order spectral features contain more specific information about paralinguistic aspects of the speech, such as speaker, rhythm, melody, timbre, emotion and so on [31].

Auto-associative neural networks are known for capturing the non-linear relations among the feature vectors [32]. AANNs are also capable of capturing the distribution properties as GMMs do [14]. So the best performance by GMMs, using 21 spectral features, is compared with the results of relevant AANNs. AANNs are used as the emotion classifiers to cross validate empirically the results obtained by GMMs. Table 2.9 shows the comparison of emotion recognition results of both GMMs and AANNs. AANN structures used for developing emotion models are also given in the last column of the table. From the results, it is observed that emotion recognition performance using GMMs is better than that of AANN models. This indicates that emotion specific information from the spectral features is better captured by GMMs than by the AANNs. The basic purpose of any emotion recognition system is to recognize the real life emotions with greater accuracy. Recognition of real emotions using the proposed sub-syllabic and pitch synchronous spectral features is discussed in Chap. 6. The proposed spectral features are also tested on internationally known Berlin emotion speech corpus (Emo-DB). The results obtained are on par with the results of our Indian database (IITKGP-SESC). Table 2.10 shows the comparison of emotion recognition results obtained using IITKGP-SESC and Emo-DB.

The results given in Table 2.10 are obtained using 21 spectral and 13 formant features. The emotion recognition systems are developed using GMMs. From the table, it is evident that the trends of emotion recognition performance using different spectral features on IITKGP-SESC and Emo-DB are almost the same.

So far we have carried out emotion recognition using the Set3 data set of IITKGP-SESC, which represents speaker and text independent emotion recognition. Along

Table 2.9 Average emotion classification performance of GMM and AANN models using spectral features on IITKGP-SESC

Features	Methods & regions for feature extr.	GMMs	AANNs	AANN structure
LPCCs	Block	69	63	21-45-10-45-21
MFCCs	processing	65	59	21-45-10-45-21
Formant features	(Entire	47	41	13-28-7-28-13
LPCCs+Formants	speech	70	61	34-60-15-60-34
MFCCs+Formants	signal)	67	60	34-60-15-60-34
LPCCs	Vowel	54	48	21-45-10-45-21
MFCCs	region	51	43	21-45-10-45-21
LPCCs+Formants		60	**57**	34-60-15-60-34
MFCCs+Formants		51	45	34-60-15-60-34
LPCCs	Consonant	44	39	21-45-10-45-21
MFCCs	region	44	37	21-45-10-45-21
LPCCs+Formants		51	46	34-60-15-60-34
MFCCs+Formants		36	33	34-60-15-60-34
LPCCs	CV-	65	57	21-45-10-45-21
MFCCs	transition	58	51	21-45-10-45-21
LPCCs+Formants	region	71	64	34-60-15-60-34
MFCCs+Formants		68	66	34-60-15-60-34
LPCCs	Pitch	69	64	21-45-10-45-21
MFCCs	synchronous	68	61	21-45-10-45-21
LPCCs+Formants	analysis	73	67	34-60-15-60-34
MFCCs+Formants		72	68	34-60-15-60-34

Table 2.10 Average emotion classification performance using the proposed spectral features on IITKGP-SESC and Emo-DB

Features	Methods & regions for feature extr.	IIT KGP-SESC	Emo-DB
LPCCs	Block	69	64
MFCCs	processing	65	63
Formant features	(Entire	47	41
LPCCs+Formants	speech	70	68
MFCCs+Formants	Signal)	67	63
LPCCs	Vowel	54	51
MFCCs	region	51	50
LPCCs+Formants		60	57
MFCCs+Formants		51	49

continued

Table 2.10 continued

Features	Methods & regions for feature extr.	IIT KGP-SESC	Emo-DB
LPCCs	Consonant	44	40
MFCCs	region	44	42
LPCCs+Formants		51	51
MFCCs+Formants		36	34
LPCCs	CV-	65	66
MFCCs	transition	58	57
LPCCs+Formants	region	71	69
MFCCs+Formants		68	64
LPCCs	Pitch	69	67
MFCCs	synchronous	68	65
LPCCs+Formants	analysis	73	72
MFCCs+Formants		72	70

Table 2.11 Average emotion classification performance using proposed spectral features on Set1, Set2, and Set3 datasets of IITKGP-SESC

Features	Methods & regions for feature extr.	IITKGP-SESC (rec.%)		
		Set1	Set2	Set3
LPCCs	Block	74	71	69
MFCCs	processing	67	65	65
Formant features	(Entire	53	49	47
LPCCs+Formants	speech	75	72	70
MFCCs+Formants	Signal)	71	68	67
LPCCs	Vowel	57	55	54
MFCCs	region	56	54	51
LPCCs+Formants		65	62	60
MFCCs+Formants		53	52	51
LPCCs	Consonant	47	45	44
MFCCs	region	47	44	44
LPCCs+Formants		55	53	51
MFCCs+Formants		40	38	36
LPCCs	CV-	68	66	65
MFCCs	transition	63	60	58
LPCCs+Formants	region	74	72	71
MFCCs+Formants		72	71	68
LPCCs	Pitch	75	73	69
MFCCs	synchronous	72	69	68
LPCCs+Formants	analysis	77	75	73
MFCCs+Formants		74	73	72

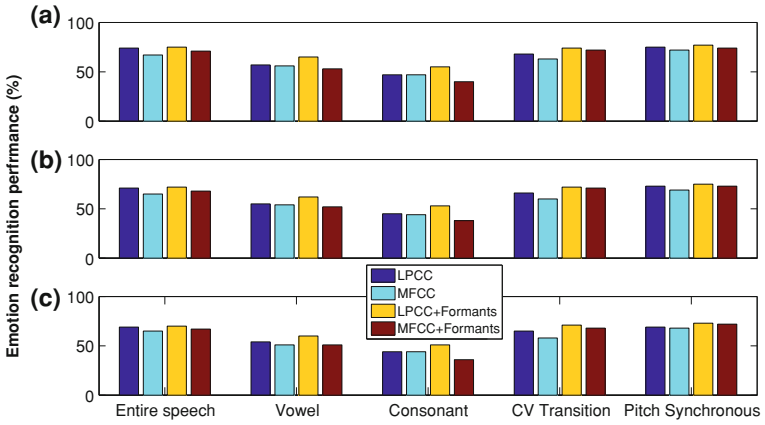

Fig. 2.12 Comparison of emotion recognition performance of different proposed spectral features with respect to the entire speech, sub-syllabic regions, and pitch synchronous analysis on Set1, Set2, and Set3 of IITKGP-SESC. **a** Emotion recognition performance using Set1, **b** Emotion recognition performance using Set2, and **c** Emotion recognition performance using Set3

with Set3, emotion recognition results of Set1 and Set2 are given in Table 2.11. Recognition performance of Set2 data set is 2 % higher than the results of Set3 data set. This is mainly due to the influence of speaker specific information during emotion classification. In the case of Set1, the recognition performance is about 4–5 % more than the results of Set3 and about 2–3 % more than the results of Set2. This improvement in emotion recognition is due to text and speaker specific information. The comparison of emotion recognition performance using proposed spectral features on Set1, Set2, and Set3 of IITKGP-SESC is given in the bar graph shown in Fig. 2.12.

2.6 Summary

In this chapter, spectral features derived from sub-syllabic regions and pitch synchronous analysis are proposed for recognizing the emotions from speech. IITKGP-SESC and Emo-DB are used to carry out the emotion classification using the proposed spectral features. LPCCs, MFCCs and formant features are used as features to represent vocal tract information. Spectral features derived from sub-syllabic regions are independently analyzed for classifying the emotions. It may be concluded from the results that the entire speech signal may not be necessary to recognize underlying emotions. Hence, the redundant information present in the steady region of the vowel may be exempted from feature extraction. The spectral features extracted from CV transition regions have achieved the emotion recognition performance, almost comparable with the performance obtained using the entire speech signal. Pitch synchronously

extracted spectral features outperformed the other spectral features while recognizing the emotions. The combination of LPCCs and formant features also has demonstrated better emotion recognition performance. The studies conducted in this chapter indicate that spectral features contain more discriminating properties with respect to different emotions than excitation source features. In the literature, MFCCs are claimed to be robust features for majority of the speech tasks such as: speech recognition and synthesis. Surprisingly, in this study, LPCCs are found to be outperforming MFCCs while classifying the emotions. Formant features combined with basic spectral features have always improved the emotion recognition performance of the systems by a consistent margin of around 2–3 %. Two classification models, namely GMMs and AANNs, are used for developing emotion recognition systems. The majority of the results reported in this chapter are obtained using GMMs as they performed slightly better than AANNs.

References

1. D. Ververidis, C. Kotropoulos, Emotional speech recognition: Resources, features, and methods. SPC **48**, 1162–1181 (2006)
2. D. Neiberg, K. Elenius, K. Laskowski, Emotion recognition in spontaneous speech using GMMs, in *INTERSPEECH 2006—ICSLP*, (Pittsburgh, Pennsylvania), pp. 809–812, 17–19 Sept 2006
3. D. Bitouk, R. Verma, A. Nenkova, Class-level spectral features for emotion recognition, Speech Commun. (2010) (in Press)
4. S.G. Koolagudi, S. Maity, V.A. Kumar, S. Chakrabarti, K.S. Rao, IITKGP-SESC: Speech Database for Emotion Analysis. Communications in Computer and Information Science, JIIT University, Noida, India, Springer. ISSN: 1865–0929 ed., 17–19 Aug 2009
5. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of german emotional speech, in *Interspeech*, Lissabon, 2005
6. L.R. Rabiner, B.H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, New Jersy, 1993)
7. J. Chen, Y.A. Huang, Q. Li, K.K. Paliwal, Recognition of noisy speech using dynamic spectral subband centroids. IEEE Signal Process. Lett. **11**, 258–261 (2004)
8. S.V. Gangashetty, C.C. Sekhar, B. Yegnanarayana, Detection of vowel on set points in continuous speech using auto-associative neural network models, in *INTERSPEECH*, IEEE, 2004
9. K.S. Rao, B. Yegnanarayana, Duration modification using glottal closure instants and vowel onset points. Speech Commun. **51**, 1263–1269 (2009)
10. S.R.M. Prasanna, B.V.S. Reddy, P. Krishnamoorthy, Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. IEEE Trans. Audio Speech Lang. Process. **17**, 556–565 (2009)
11. Y. Zeng, H. Wu, R. Gao, Pitch synchronous analysis method and fisher criterion based speaker identification, in *Third International Conference on Natural Computation*, vol. 2 (IEEE Computer Society, Washington DC, USA, 2007), pp. 691–695. ISBN: 0-7695-2875-9
12. H. Muta, T. Baer, K. Wagatsuma, T. Muraoka, H. Fukuda, Pitch synchronous analysis of hoarseness in running speech. J. Acoust. Soc. Am. **84**, 1292–1301 (1988)
13. K. Murty, B. Yegnanarayana, Epoch extraction from speech signals. IEEE Trans. Audio Speech Lang. Process. **16**, 1602–1613 (2008)
14. B. Yegnanarayana, S.P. Kishore, AANN an alternative to GMM for pattern recognition. Neural Networks **15**, 459–469 (2002)

15. B. Yegnanarayana, *Artificial Neural Networks* (Prentice-Hall, New Delhi, India, 1999)
16. S. Haykin, *Neural Networks: A Comprehensive Foundation* (Pearson Education Aisa, Inc., New Delhi, India, 1999)
17. K.I. Diamantaras, S.Y. Kung, *Principal Component Neural Networks: Theory and Applications* (Wiley, New York, 1996)
18. M.S. Ikbal, H. Misra, B. Yegnanarayana, Analysis of autoassociative mapping neural networks, in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, (USA, 1999), pp. 854–858
19. S.P. Kishore, B. Yegnanarayana, Online text-independent speaker verification system using autoassociative neural network models, in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, vol. 2 (Washington, DC, USA, 2001), pp. 1548–1553
20. A.V.N.S. Anjani, Autoassociate neural network models for processing degraded speech, Master's Thesis, MS Thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India, 2000
21. K.S. Reddy, Source and system features for speaker recognition, Master's Thesis, MS Thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India, 2004
22. C.S. Gupta, Significance of source features for speaker recognition, Master's Thesis, MS Thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai 600 036, India, 2003
23. S. Desai, A. W. Black, B.Yegnanarayana, K. Prahallad, Spectral mapping using artificial neural networks for voice conversion. IEEE Trans. Audio Speech Lang. Process. **18**, 954–964 (2010)
24. K.S. Rao, B. Yegnanarayana, Intonation modeling for indian languages. Comput. Speech Lang. **23**, 240–256 (2009)
25. C.K. Mohan, B. Yegnanarayana, Classification of sport videos using edge-based features and autoassociative neural network models. Signal Image Video Process. **4**, 61–73 (2008). doi:10.1007/s11760-008-0097-9
26. L. Mary, B. Yegnanarayana, Autoassociative neural network models for language identification, in *International Conference on Intelligent Sensing and Information Processing*, IEEE, pp. 317–320, 24 Aug 2004. doi:10.1109/ICISIP.2004.1287674
27. B. Yegnanarayana, K.S. Reddy, S.P. Kishore, Source and system features for speaker recognition using aann models, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Salt Lake City, UT), May 2001
28. C.S. Gupta, S.R.M. Prasanna, B. Yegnanarayana, Autoassociative neural network models for online speaker verification using source features from vowels, in *International Joint Conference on Neural Networks*, (Honululu, Hawii, USA), May 2002
29. B. Yegnanarayana, K.S. Reddy, S.P. Kishore, Source and system features for speaker recognition using AANN models, in *Proceedings of the IEEE International Conference Acoustics, Speech, Signal Processing*, (Salt Lake City, Utah, USA), pp. 409–412, May 2001
30. S.G. Koolagudi, K.S. Rao, Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. Int. J. Speech Technol. **15**, 495–511 (2012). doi:10.1007/s10772-012-9150-8
31. O.M. Mubarak, E. Ambikairajah, J. Epps, Analysis of an mfcc-based audio indexing system for efficient coding of multimedia sources, in *The 8th International Symposium on Signal Processing and its Applications*, (Sydney, Australia), 28–31 Aug 2005
32. R.O. Duda, P.E. Hart, D.G. Stork, *Pattern Classification*, 2nd edn. (A Wiley-interscience Publications, Singapore, 2004)

http://www.springer.com/978-1-4614-6359-7

Robust Emotion Recognition using Spectral and Prosodic Features

Rao, K.S.; Koolagudi, S.G.

2013, XII, 118 p. 37 illus., 15 illus. in color., Softcover
ISBN: 978-1-4614-6359-7