Chapter 2
# Emotion Recognition Using Prosodic Features

2.1 Introduction

In computer vision, a feature is a set of measurements. Each measurement contains a piece of information and specifies the property or characteristics of the object. In speech recognition techniques, how the speech signals are produced and perceived by the human is starting point of the research. Human speech communication produces ideas (word sequence) which are made within the speaker brain. These word sequence are delivered by his/her text generator. The general human vocal system is modeled by the speech generator. The speech generator converts the word sequence into speech signal and is transferred to listener through air. At the listener side, the human auditory system receives these acoustic signal and listeners brain starts the processing of signal to understand its content. The speech recognizer modeled by the speech decoder, it decodes the acoustic signal into word sequence. So speech production and speech perception are in inverse processes in the speech recognition application.

When we analyze speech signals, most of them are more or less stable within a short period of time. When we do frame blocking, there may be some overlaps between neighboring frames as shown in Fig. 2.1. Each frame is the basic unit for analysis. The basic approach to the extraction of acoustic features from the speech signal can be summarized as follows:

- Converting the stream of speech signals into set of frames by performing frame blocking. The size of each frame must not be too big and too small. If frame size is too big, we can't extract the time-varying characteristics of the audio signals. And if the frame size is too small, we can't extract the valid acoustic features, then extraction cannot be done. In general, a frame contains several fundamental periods of the given speech signal. Usually the frame size (in terms of sample points) is equal to the powers of 2 (such as 256, 512, 1024 etc.) such that it is suitable for fast Fourier transform.
- For frame blocking we should also consider the duration between the frames. If we want to reduce the difference between neighboring frames, we can allow overlap between them. Usually the overlap is 1/2 to 2/3 of the original frame.
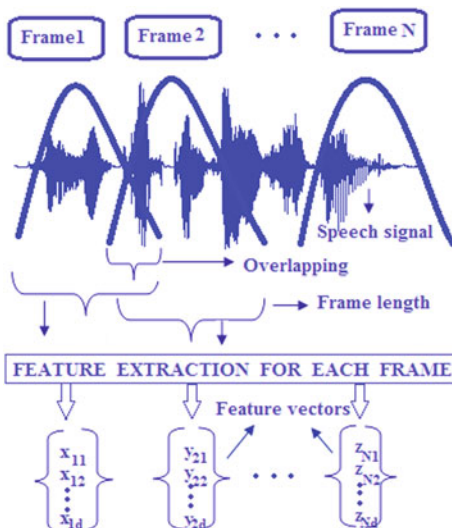
Fig. 2.1 The process of splitting the speech signal into several frames, appling an hamming window for each frame and its corresponding feature vector

The more overlap, the more computation is needed. So we must take care of the distance between two frames.

- After frame blocking, we extract the acoustic features from the audio signals such as zero crossing rate, short time energy, pitch, MFCC etc, can be done by assuming the audio signals within the frame as stationary [87].
- By observing the zero crossing rate and short time energy of the each frame, we have to analyze that particular frame of audio signal is voice speech signal or unvoiced speech signal and keep the voice speech (non-silence frames) for further analysis.

In case of acoustic information, features are mainly classified as temporal and spectral features. Zero Crossing Rate (ZCR), Short Time Energy (STE) are the examples of temporal features. Mel frequency cepstral coefficient (MFCC) and Linear Prediction Cepstral Coefficeints (LPCC) are examples of spectral features.

2.2 Pre-Processing

Speech signals are normally preprocessed before features are extracted to enhance the accuracy and efficiency of the feature extraction process. The modules filtering, framing and windowing are considered as steps under preprocessing.
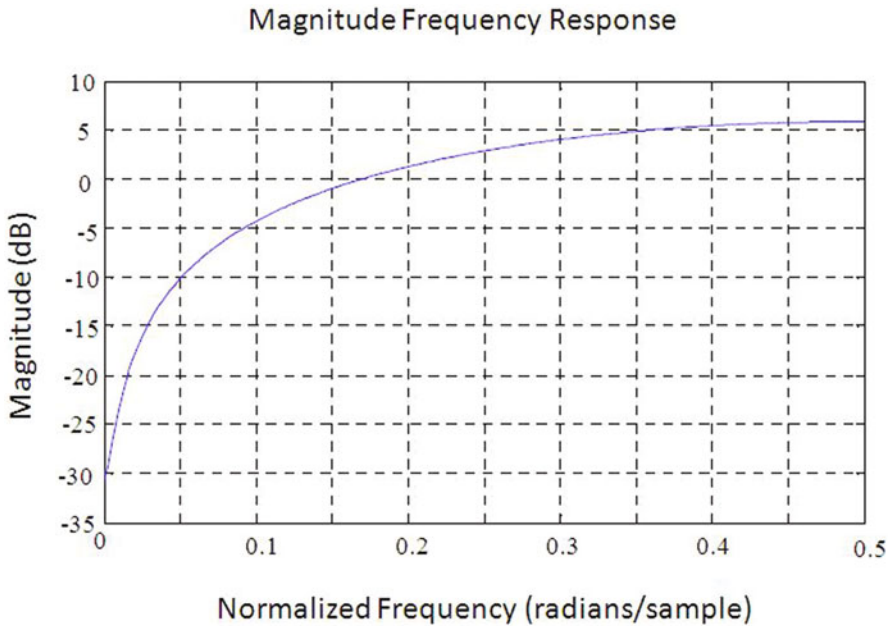
## Magnitude Frequency Response

Fig. 2.2 Pre-emphasis filter

2.2.1   Filtering

The filtering technique is applied to reduce the noise, which is occured due to the environmental conditions or any other disturbances while recording the speech sample. To reduce the noise effect, filter operations are performed which optimizes the class separability of features [1]. This is done by using the high pass filter.

The main goal of Pre-emphasis is to boost he amount of energy in the higher frequencies with respect to lower frequencies. Mainly boosting is used to get more information from the higher frequencies available to the acoustic model and to improve the recognition performance [2, 7]. This pre-emphasis is done by using a first-order high pass filter as shown in Fig. 2.2.

2.2.2   Framing

Frame blocking is converting the stream of audio signal into set of frames and analyzed independently. The original vector of sampled values will be framed into overlapping blocks. Each block will contain 256 samples with adjacent frames being separated by 128 samples. This will yield a minimum of 50 % overlap to ensure that all sampled values are accounted for within at least two blocks. Two hundred and
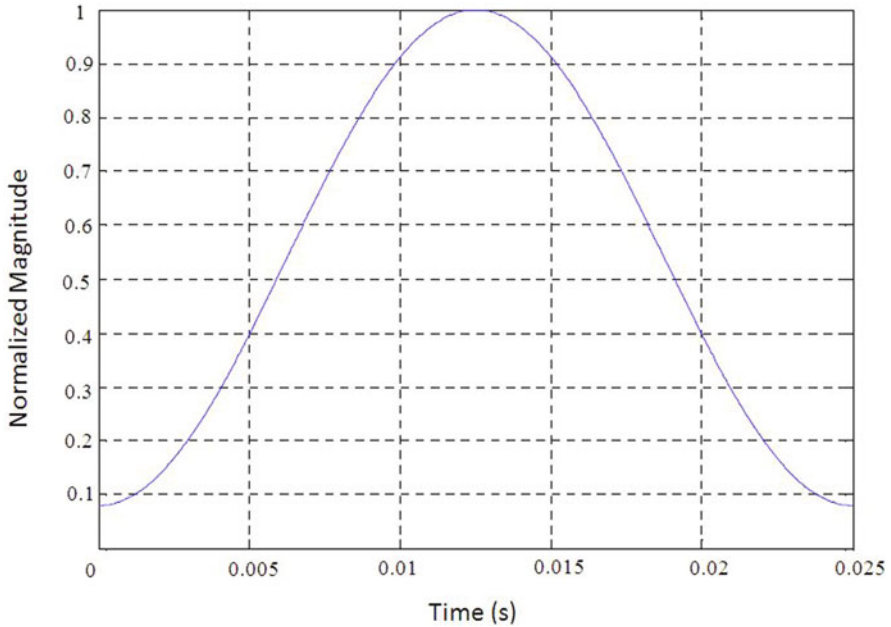
Fig. 2.3 25 ms hamming window fs $= 16$ khz [1]

fifty six was chosen so that each block is 16 ms. In this step the continuous speech signal is made into frames of N samples, with adjacent frames being separated by M (M < N). Typical values used are M $= 100$ and N $= 256$ [1]. Frame blocking is shown in Fig. 2.1.

2.2.3 Windowing

In the window operation, the large input data is divided into small data sets and stored in sequence of frames. While dividing the signal into frames, some of the input data signal may be discontinuous at the edges of the each frame. So a tapered window is applied to each one. The hamming window is used to reduce the spectral leakage in the input data signal.

The generally used window is rectangular window, it is the simplest window. But this window can cause some problems, however, because it abruptly cuts off the signal at its boundaries. These discontinuities create problems when we do Fourier analysis. Therefore it is necessary to keep the continuity of the first and the last points in the frame. For this reason, the Hamming window is used in feature extraction. The 25 ms hamming window is shown in Fig. 2.3.

2.3   Extraction of Prosodic Features

An important module in the design of speech emotion recognition system is the selection of features which best classify the emotions. These features distinguishes the emotions of different classes of speech samples. These are estimated over simple six statistics as shown in Table 2.1.

The prosodic features extracted from the speech signal are Zero Crossing Rate, Short Time Energy and Pitch. Their first and second order differentiation provides new useful information hence the information provided by their derivatives also considered [37]. These features were estimated for each frame together with their first and second derivatives, providing six features per frame and applying statistics as shown in Table 2.1 giving a total of 36 prosodic features.

2.3.1   Zero Crossing Rate

In the context of discrete-time signals, a zero crossing is said to occur if the successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of a signal. Zero-crossing rate is measure of number of times the amplitude of the speech signals passes through a value of zero in a given time interval/frame. The calculation of this is as shown in Eq. 2.1

$$ZCR = \frac{1}{N} \sum_{i=0}^{N-1} |sgn(x(i)) - sgn(x_{-1}(i-1))| \qquad (2.1)$$

where the value of sgn are

$$
\begin{array}{cc}
1 & x(i) > 0 \\
0 & x(i) = 0 \\
-1 & x(i) < 0
\end{array}
$$

where $x_{-1}(N)$ is a temporary array created to store the previous frame values Eq. 2.1 shows the mathematical formula to calculate feature values using zero crossing rate.

Table 2.1 The statistics and their corresponding symbols [35]

Statistics	Symbol
Mean	E
Variance	V
Mininum	Min
Range	R
skewness	Sk
kurtosis	K

The function of sgn in the equation is to assign the normalized value $[-1, 0, 1]$ based on the range of input variable value. Since high frequencies imply high zero crossing rates, and low frequencies imply low zero-crossing rates, there is a strong correlation between zero crossing rate and energy distribution with frequency [74]. A reasonable generalization is that if the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced.

2.3.2 Short Time Energy

The amplitude of the speech signal varies with time. Generally, the amplitude of unvoiced speech segments is much lower than the amplitude of voiced segments. The energy of the speech signal provides a representation that reflects these amplitude variations . Short-time energy can define in Eq. 2.2

$$STE = \frac{1}{N} \sum_{i=0}^{N-1} |X(n)|^2 \tag{2.2}$$

where N describes the total number of samples in a frame or a window. X(n) is a speech signal in a frame. A reasonable generalization is that if the Short time energy is high, the speech signal is voiced, while if the Short time energy is low, the speech signal is unvoiced. Based on zero crossing rate and short time energy, voiced sounds are identified. We can extract the following features from the identified voice speech signal.

2.3.3 Pitch

Pitch is the fundamental frequency of audio signals, which is equal to the reciprocal of the fundamental period [75]. This is mainly explained in terms of highness or lowness of a sound. Pitch in reality can be defined as the repeat rate of a complex signal, i.e., the rate at which peaks in the autocorrelation function occur. The three main difficulties in pitch extraction arise due to the following factors:

- Vocal cord vibration does not necessarily have complete periodicity, especially at the beginning and end of the voiced sounds.
- From speech wave, vocal cord source signal can be extracted but its extraction is difficult if it has to be extracted separately from the vocal tract effects.
- The fundamental frequency possesses very large dynamic range.

Above three viewpoints are very much used for the research on pitch extraction recently. The main point is being able to extract quasi-periodic signal's periodicity in a reliable manner. Another is how to correct the pitch extraction error owing to the disturbance of periodicity. The other is how to remove the vocal tract (formant)

**Table 2.2** Classification of major Pitch extraction methods and their major principle features [20]

Classification	Pitch extraction method	Principal features
1. Waveform processing	Parallel processing method	Uses majority rule for pitch periods extracted by many kinds of simple waveform peak detectors
	Data reduction method	Removes superfluous waveform data based on various logical processing and leaves only pitch pulses
	Zero crossing count method	Utilizes iterative pattern in waveform zero crossing rate
2. Correlation processing	Autocorrelation method	Employs autocorrelation of waveform. Applies center and peak clipping for spectrum flattering and computation simplification
	Modified correlation method	Utilizes autocorrelation function for residualsignal of LPC analysis. Computation is simplified by LPF and polarization
	SIFT (simplified Inverse filter tracking) algorithm	Applied LPC analysis for spectrum flattering after down-sampling of speech wave. Time resolution is recovered by interpolation
	AMDF method	Uses average magnitude differential function (AMDF) for speech or residual signal for periodicity detection
3. Spectrum processing	Cepstrum method	Separates spectral envelope and fine structure by inverse Fourier transform of log power spectrum
	Period histogram method	Utilizes histogram for harmonic components in spectral domain. Pitch is decided as the common divisor for harmonic components

effects. Double-pitch and half-pitch are the two main classifications of major errors in pitch extraction. So as the name suggests the double-pitch errors are those which occur while extracting the frequencies which are twice as large as actual value and similarly the half-pitch errors occur due to extraction of half-value of the original fundamental frequency. The employed extraction method plays a major role in deciding the tendency toward which the error is most apt to occur. Precise information regarding all the major pitch extraction methods are given in the Table 2.2. Most of the works uses Auto correlation method to detect pitch.

Pitch Detection using Autocorrelation Method:

Autocorrelation function is used to estimate pitch, directly from the waveform. xcorr function is used to estimate the statistical cross-correlation sequence of random process and is given by

$$R[m] = E(x[n + m]x[n]) = E(x[n]x[n - m]) \qquad (2.3)$$
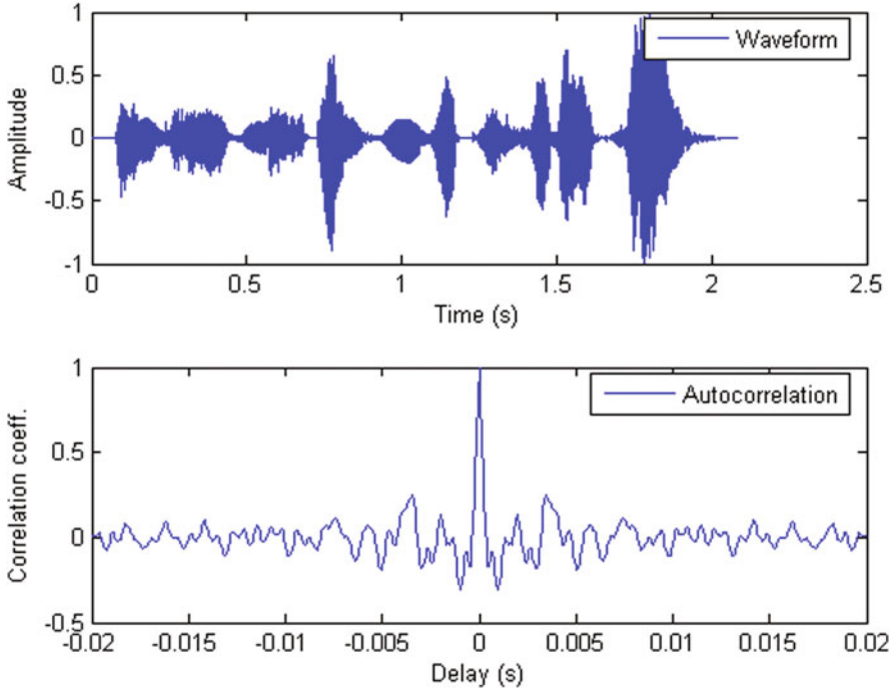
Fig. 2.4 Waveform and autocorrelation function in pitch estimation

When x and y are not the same length then the shorter vector is zero padded to the length of the longer vector. If x and y are length N vectors (N > 1), c = xcorr(x,y) returns the cross-correlation sequence in a length 2*N−1 vector [45, 75]. The autocorrelation function is given by

$$R(m) = \frac{1}{N} \sum_{n=0}^{N-m-1} (x[n+m]x * [n])m \geq 0 \tag{2.4}$$

Where x(n) is a speech signal, n is time for discrete signal, m is a lag number. If x(n) is similar with x(n + m), then R[m] has a large value. Whenever x(n) has a period of P, then the value of R[m] has peaks at m = lP where l is an integer [25, 31]. Here we need to estimate R[m] from N samples. The empirical autocorrelation function is given by $R(m) = \frac{1}{N} \sum_{n=0}^{N-m-1} (w[n+m]x[n+m]w[n]x * [n])$, where w[n] is a window function of length N. The peaks of autocorrelation function and waveform of a speech signal as shown in Fig. 2.4.

We can estimate the fundamental frequency by using autocorrelation function, peaks at delay intervals corresponding to the normal pitch range in speech, say 2 ms (= 500 Hz) and 20 ms (= 50 Hz). Figure 2.4 shows the pitch tracking using autocorrelation method. Pitch is very low for male voice and very high for female voice [69].

2.4   **Importance of Prosodic Features**

The features extracted from the speech signal plays a major role in identifying emotion and are capable of detecting an exact emotion of an unknown speaker who is not visible during training period. But the detection of feature set which best classify the emotion is a complex task [60].

Some physiological and psychological changes occurs due to each and every emotion. For instance happiness occurs when we stood stood first in competitive examination. This causes some changes in characteristic of speech also like amplitude, frequency and speech rate etc. These are nothing but prosodic features [38]. Because of this reason, emotion identification systems uses prosodic features for many years. Many works in the literature deals with these prosodic features in identifying the emotion [9, 24, 62]. A detailed summary of these works is given in [18]. Most of the emotional information is obtained by using these prosodic features and are estimated over short term speech segments by using simple statistics like minimum, maximum, mean, variance, skewness and kurtosis etc., As shown in Fig. 1.2 prosodic features discriminate high arousal emotions to low arousal emotions more accurately.

Springer