Chapter 2
Background and Literature Review

Abstract This chapter provides the systematic review of the existing approaches for vowel onset point detection, speech systems in mobile environment, Consonant-Vowel (CV) recognition in Indian languages, and time scale modification (TSM). In addition to providing the review of above-mentioned topics, authors have discussed about the short comings present in the existing approaches and derived the motivation and scope of the present work.

This chapter discusses about the state of the art related to the contents of this book. Authors have provided detailed explanation for the existing VOP detection methods, CV recognition systems, and time scale modification methods, which are later used for comparing the performance with the proposed methods and systems. The chapter is organized as follows: Sect. 2.1 reviews existing methods for VOP detection. Section 2.2 briefly reviews the state-of-the-art speech and speaker recognition systems in mobile environment. Section 2.3 presents the review of CV recognition in Indian languages. Section 2.4 reviews the existing approaches for TSM. Section 2.5 summarizes the review and the major issues addressed in this book.

2.1 Approaches for Detection of Vowel Onset Points

There are various methods available in literature for the detection of VOPs [7–17]. The method presented in [7] detects VOPs based on rapid increase in the vowel strength. The vowel strength is calculated using the difference in the energy of the peaks and their corresponding valleys in the amplitude spectrum. This method requires detection of unvoiced and voiced regions in a speech signal. The method for VOP detection presented in [8] uses a product function generated by using wavelets. The values of product function for vowel regions are much larger than consonant regions. The methods presented in [9–11] use a hierarchical neural

K.S. Rao and A.K. Vuppala, *Speech Processing in Mobile Environments*, SpringerBriefs in Electrical and Computer Engineering, DOI 10.1007/978-3-319-03116-3_2,

network, multilayer feed-forward neural network (MLFFNN), and auto-associative neural network (AANN) models, respectively, for the detection of VOPs. These models are trained by using the trends in the speech signal parameters at the VOPs. The VOP detection using Hilbert envelope of excitation source information is presented in [12]. The acoustic cues such as formant transition, epoch intervals, strength of instants, symmetric Itkura distance, and ratio of signal energy to residual energy are explored in [4, 5] for the detection of VOP events in different categories of CV units.

The voice onset time (VOT) is the time delay between the burst onset and the start of periodicity, when it is followed by a voiced sound. In [13], automatic VOT is detected using phone model-based methods with forced alignment. VOT detection using reassignment spectra is presented in [14]. In [18], voice onset time detection method is presented for unvoiced stops (/p/, /t/ and /k/) using the nonlinear energy tracking algorithm (Teager energy operator). In [19], Bessel features are used for determining the voice onset time for stop consonant vowel units such as /ka/, /Ta/, /ta/, and /pa/.

Combination of the evidence from excitation source, spectral peaks, and modulation spectrum (COMB-ESM) has been explored in [15] for the detection of VOPs. Each of these evidence carries complementary information with respect to VOPs. The performance of COMB-ESM method is superior compared to existing methods. Hence, in this book COMB-ESM method is used for comparing the performance of the proposed VOP detection methods. Following subsection describes the details of the COMB-ESM method for VOP detection [15].

2.1.1 VOP Detection Using Excitation Source, Spectral Peaks, Modulation Spectrum, and Their Combination

2.1.1.1 VOP Detection Using Excitation Source Information

VOP detection using excitation source information is carried out in the following sequence of steps. Determine the Hilbert envelope (HE) of linear prediction (LP) residual (also known as excitation source) of speech signal. Smooth the HE of the LP residual by convolving with a Hamming window of size 50 ms. The change at the VOP present in the smoothed HE of the LP residual is further enhanced by computing its slope using first-order difference (FOD). These enhanced values are convolved with the first order Gaussian difference (FOGD) operator, and the convolved output is the VOP evidence using excitation source. VOP evidence using excitation source for speech signal */"Don't ask me to carry an"/* is shown in Fig. 2.1b.

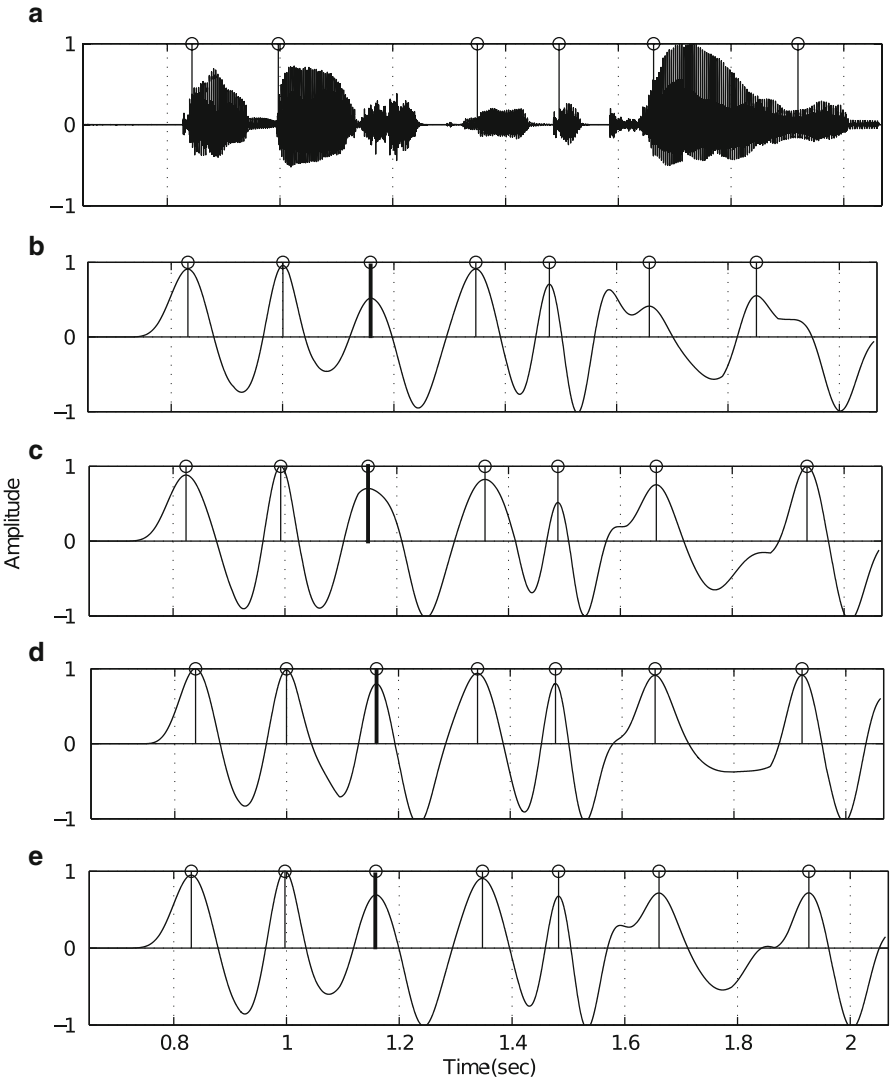

Fig. 2.1 VOP detection using combination of three evidences for a speech utterance */"Don't ask me to carry an"/.* (**a**) speech signal, VOP evidence plots for (**b**) excitation source, (**c**) spectral peaks, (**d**) modulation spectrum, and (**e**) COMB-ESM method

2.1.1.2 VOP Detection Using Spectral Peaks Energy

VOP detection using the spectral peaks energy is carried out in the following sequence of steps. The speech signal is processed in blocks of 20 ms with a shift of 10 ms. For each block, a 256-point discrete Fourier transform (DFT) is computed, and the ten largest peaks are selected from the first 128 points. The sum of these

spectral peaks is plotted as a function of time. The change at the VOP available in the spectral peaks energy is further enhanced by computing its slope using FOD. These enhanced values are convolved with FOGD operator. The convolved output is the VOP evidence using spectral peaks energy. VOP evidence plot using spectral peaks energy for speech signal */"Don't ask me to carry an"/* is shown in Fig. 2.1c.

2.1.1.3 VOP Detection Using Modulation Spectrum Energy

Slowly varying temporal envelope of speech signal can be represented by using modulation spectrum. VOP detection using modulation spectrum energy is carried out in the following sequence of steps. The temporal envelope of speech is dominated by low-frequency components. The VOP evidence due to 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. The change at the VOP available in the modulation spectrum energy is further enhanced by computing its slope using FOD. These enhanced values are convolved with FOGD operator and the convolved output is the VOP evidence using modulation spectrum energy. VOP evidence using modulation spectrum energy for speech signal */"Don't ask me to carry an"/* is shown in Fig. 2.1d.

2.1.1.4 VOP Detection Using COMB-ESM Method

Each of the above three methods uses complementary information about the VOP, and hence they are combined for the enhancement of VOP detection performance. In combined method, the evidences from excitation source, spectral peaks, and modulation spectrum energies are added sample by sample. VOP detection using individual and combination of all three evidences for speech signal */"Don't ask me to carry an"/* is shown in Fig. 2.1.

Figure 2.1a shows the speech signal with manually marked VOPs for an utterance */"Don't ask me to carry an"/*. Figure 2.1b–d shows the VOP evidence corresponding to excitation source, spectral peaks, and modulation spectrum, respectively. Figure 2.1e shows the VOP evidence by combining the evidence. The peaks in the combined VOP evidence signal (Fig. 2.1e) are marked as the VOPs obtained from COMB-ESM method. From Fig. 2.1, it is observed that a spurious VOP is present in third position in all VOP evidence plots. The performance of COMB-ESM method for VOP detection is around 96 % within 40 ms deviation and only 45 % within 10 ms deviation [15]. A summary of the discussion related to the detection of VOP is given in Table 2.1.

Table 2.1 Summary of the review of VOP detection methods

• Existing methods for VOP detection have low accuracy • Most of the existing VOP detection methods are based on block processing of speech signals • Information present in glottal closure regions may also be used for the detection of the VOP events in the presence of degradations such as coding and background noise

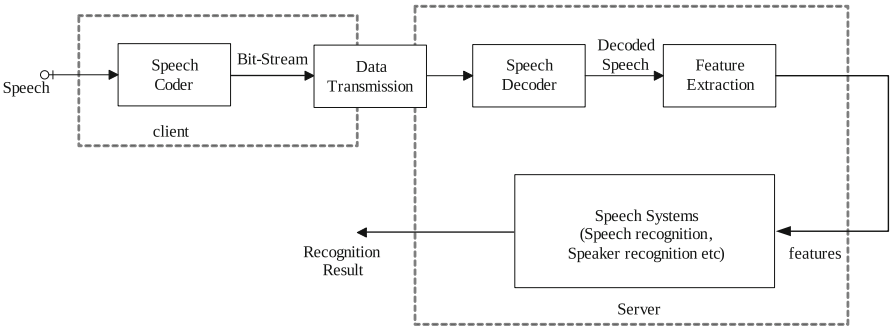

Fig. 2.2 Block diagram for network-based speech system

2.2 Speech Processing in Mobile Environment

In mobile environment, speech systems are developed in three different configurations. They are: (1) embedded speech systems (client-based), (2) network speech systems (server-based), and (3) distributed speech systems (client-server-based) [20–24]. These configurations are characterized according to the location where processing is taking place. In embedded speech systems, the speech task is performed in the terminal device itself. Due to cost-sensitive nature of the terminal device, constraints are imposed on computational and memory resources. Therefore, this approach is aimed to limited vocabulary applications. In network speech systems, speech is transmitted to remote server over a communication channel, and speech task is performed at the server. In distributed speech systems, features required for speech task are extracted from the speech at the client side, and task is performed at the remote server. The applications such as large continuous speech recognition in mobile environment is possible through network and distributed based speech systems. In this book, network-based configuration is considered for analyzing the proposed methods for speech and speaker recognition. The block diagram for network-based speech system is shown in Fig. 2.2. Survey of speech recognition techniques for mobile devices is presented in [20–23, 25].

There are three major challenges in speech processing in mobile environment [21–23, 25]: (1) Degradations due to different speech coders used for speech transmission. Speech coding is a compact way of representing speech by exploiting speech production and perception characteristics. In the process of speech coding, speech and speaker-specific information present in speech will be degraded.

(2) Effect of varying background conditions in mobile environment on the performance of speech systems. Background conditions like crowd of people, vehicle, restaurant, street, etc. are common in mobile environment, and they will degrade the performance of speech systems. (3) Effect of wireless channels on the performance of speech systems. Due to unreliable nature of radio-frequency channel, transmission errors will affect the performance of speech systems. Distortions due to speech coding and channel errors like packet loss are also common issues in voice over internet protocol (VoIP). In addition to these distortions, jitter is an issue in VoIP technology.

Currently speech coders are coming in two different versions. The basic version, also called narrowband, which is mainly intended for use by GSM, and wideband (for example AMR-WB), which is mainly intended for use by Universal Mobile Telecommunications System (UMTS). UMTS is one of the third generation (3G) mobile telecommunications systems. Wideband coders uses a speech bandwidth of 50–7,000 Hz, whereas the bandwidth of narrowband AMR is 300–3,400 Hz. This gives wideband AMR a more natural speech quality. We consider different narrowband speech coders to observe coding effect on the performance of speech systems. In this work, issues related to speech and speaker recognition under coding and speech recognition under background noise are addressed. Following subsections discuss the background work related to those issues.

2.2.1 Speech and Speaker Recognition Under Coding

Speech recognition is the process of converting spoken words to a machine readable input (text). The effect of speech coders such as GSM and CELP coders on digit recognition performance by using HMM models has been discussed in [26, 27]. Juan Huerta has presented weighted acoustic models to reduce the effect of GSM full rate coder on the speech recognition performance [22]. From his study it is evident that all phonemes in a GSM-coded speech corpus are not distorted to the same extent due to coding. Alternative front-end for speech recognition in GSM networks is presented in [28]. In this approach features are extracted directly from the encoded speech to avoid source coding distortion.

Speaker recognition is the process of automatically recognizing the identity of speaker from speech. Speaker recognition can be divided into speaker identification and speaker verification. In speaker identification, the task is to identify the speaker from the speech signal. The task of a speaker verification system is to authenticate the claim of a speaker based on the test speech. In literature the effect of coding on speaker recognition performance was analyzed in two ways. In the first case features required for speaker recognition are extracted from resynthesized speech [29–31], and in the second case features are extracted directly from the codec parameters [29]. The effect of GSM (12.2 kbps), G.729 (8 kbps), and G.723.1 (5.3 kbps) coders on speaker recognition is studied in [29]. This study indicated

Table 2.2 Summary of the review of speech systems in mobile environment

- There is no systematic study carried out on speech recognition for Indian languages in mobile environment
- Combined temporal and spectral preprocessing techniques can be used for speech recognition under background noise
- Information present around VOPs may be used for improving the performance of speech systems in mobile environment

that the performance of recognition system decreases with decreasing coding rate. The effect of speech coding on automatic speaker recognition is presented in [30] with matched and mismatched training and testing conditions. Matched condition (training and testing with the same coder) shows increase in the recognition performance. In [31], performance of speaker recognition under coding is improved using score normalization. The effect of GSM-EFR coder on the performance of speaker identification is presented in [32]. In [33], SVM-based text-independent speaker identification using a linear GMM supervector kernel was presented for coded speech.

2.2.2 Speech Recognition Under Background Noise

In practical applications of automatic speech recognition, speech is often distorted by a background noise. Because of this distortion, speech features are distorted, and therefore there is a mismatch between the training (clean) and testing (noisy) conditions. This mismatch severely degrades the performance of speech recognizers [34, 35]. Various methods have been presented in the literature to overcome the effect of noise on speech recognition. These methods can be grouped under three categories based on (1) compensation of noise, (2) robust feature extraction, and (3) adaptation of models. Methods based on compensation of noise aim to enhance the noisy speech signals before feature extraction [34–39]. Such methods include spectral subtraction, minimum mean square error (MMSE), and subspace-based speech enhancement techniques [36–39]. Methods based on robustness at the feature level are designed in such a way that the proposed features are less sensitive to the noisy degraded conditions [40–46], e.g., RASTA filter [40], feature normalization [41], MMSE-based mel-frequency cepstra [42], and histogram equalization [44–46], etc. In case of model adaptation approach, the parameters of the model are modified according to the characteristics of the background noise [47–53]. Some of the popular model adaptation methods include code book mapping [47], parallel model compensation [48], noise adaptive training [51, 52], etc. A summary of the discussion related to the speech systems in mobile environment is given in Table 2.2.

2.3 Recognition of CV Units of Speech in Indian Languages

Phones, diphones, and triphones are widely used subword units for speech recognition. But recent studies reveal that syllables are the suitable subword units for speech recognition [54,55] in Indian languages. In general, the syllable-like units are of type $C^m VC^n$, where C refers to consonant, V refers to a vowel, and m and n refer to the number of consonants preceding and following the vowel in a syllable. Among these units, the CV units are the most frequently (around 90 %) occurring units [9] in Indian languages. Different regions of significant events in the production of the CV unit $/ka/$ are shown in Fig. 2.3. The major issues involved in the recognition of CV units are the large number of classes and high similarity among those classes [54–57].

Hidden Markov models (HMMs) are the commonly used classification models in speech recognition, but in [54, 55, 57] authors have reported that MLFFNNs and support vector machines (SVMs) work better for recognition of CV units in Indian languages compared to HMM. In [55], modular neural networks are used for recognition of stop consonant-vowel (SCV) units. Separate neural networks (subnets) are trained for subgroups of classes. It has been reported in [55] that the performance of the conventional modular networks is poor, and a constraint satisfaction model (CSM) is presented to improve the recognition performance of SCV units. In CSM the outputs of the subnets are combined using the constraints that represent the similarities among the SCV classes. The constraints are derived from the acoustic phonetic knowledge of the classes and the performance of the subnets. In [54], constraint satisfaction neural network models are extended for recognition of isolated CV units that correspond to all categories of consonants. Features extracted around VOPs are used for recognition of CV units. In their study, VOPs are detected using AANNs and dynamic time warping (DTW)-based methods [54]. Further, CV units are recognized from continuous speech by using SVMs.

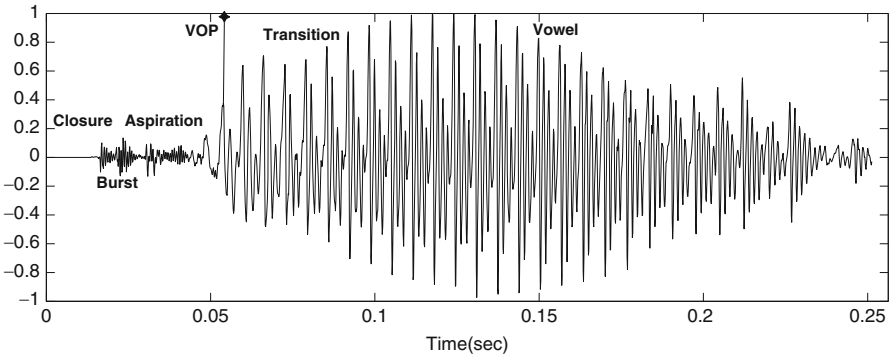

Fig. 2.3 Regions of significant events in the production of the CV unit $/ka/$

Table 2.3 Summary of the review of CV recognition in Indian languages

• VOP detection methods used in existing CV recognition systems are suffering with low accuracy
• Accurate VOP detection may improve the CV recognition performance
• The recognition performance of CV units by using single level hybrid approach presented in [58] can be improved by using multi-level hybrid approach

An approach based on combination of SVM and HMM evidences for enhancing the CV recognition performance is presented in [58]. A summary of the discussion related to CV recognition is given in Table 2.3.

2.4 Time Scale Modification

The purpose of time scale modification (TSM) is to change the rate of speech while preserving the characteristics of the original speech such as formant structure and pitch periods. There are various applications of time scale modification. For example, time scale compression can be used at the input side of the speech coder and transmission, followed by time scale expansion at the receiver to get back the original speech [59]. In some other applications, time scale expansion can be used to enhance the intelligibility of rapid or degraded speech [60]. Time scale compression is also useful in message playback systems for fast scanning of recorded messages [61]. Recently, adaptive TSM is used in VoIP applications for handling the network congestion [62]. Pitch and time scale modification was attempted in real time by focusing on processing only voiced regions of speech utterance [63].

There are number of approaches available in the literature for time scale modification. Some of them use sinusoidal model, pitch synchronous overlap and add (PSOLA), and phase vocoders [60, 64, 65]. In [59], the authors have presented an epoch-based time scale modification method, where the duration of speech signal is modified using the knowledge of epochs. In this method, TSM is performed in residual domain. Linear prediction pitch synchronous overlap and add (LP-PSOLA) approach also performs TSM in the residual domain similar to the epoch-based method [66]. The approaches mentioned above basically perform time scale modification uniformly, for entire speech signal. But, fast or slow speech produced by humans may not vary uniformly across all the speech segments. In [16], fast and slow speech produced by human beings was analyzed, and observed that durations of consonant and transition regions remain the same in fast or slow speech, and only the vowel and pause regions will vary according to speech rate. Based on this observation, authors have presented a nonuniform time scale modification method, where consonant and transition regions of speech are kept unaltered, and only vowel and pause segments are modified according to desired speaking rate [16].

Table 2.4 Summary of the review of TSM methods

• Majority of the existing TSM methods modify all speech segments with same modification factors
• Modifying different speech segments with different modification factors based on their production and articulatory constraints may improve the quality of speech

Attempts to incorporate nonuniform duration modification are reported in the literature [64, 67, 68]. Speech adaptive TSM method presented in [64] modify the speech rate based on voicing probability derived from sinusoidal pitch estimator. The voicing probability is close to unity during steady voicing, decreases during transition, and close to zero during unvoiced speech and pauses. The assumption is that changes in speaking rate for compression or expansion do not take place in sounds which are not voiced, but they occur mostly in voiced sounds. A nonuniform time scaling method has been developed along with spectral shape and pitch modification for automatic morphing of one sound to another sound [67]. Another method for speech adaptive TSM is presented, which allows slowing down the speech without compromising the quality or naturalness of the slowed speech [68]. In this method, different scaling factors are applied to different types of speech segments. Transient detection in music and audio signals has been studied for different applications such as segmentation and editing of audio recordings [69] and improving audio effects [70] through TSM [71–73]. Different methods use different cues of audio signal for the detection of transient audio segments. Sum of significant spectral peaks is used in [74] for discriminating the transients from steady segments. Variance of the spectrum and time offset of the center of gravity are used in [75] for classifying the transients. In most of the studies, transient detection was used to improve the quality of audio for different speaking rates. Bonada has proposed a frequency domain method for processing the fast changes in the signal in a different way compared to other components [71]. Roebel has proposed a new approach for processing transients in the phase vocoder, where transient peaks are preserved during stretching [73]. Recently, a nonuniform TSM method based on waveform similarity overlap-and-add (WSOLA) technique is presented for time scale modification of music signals [76]. In this approach, the perceptually significant transient sections (PSTs) such as temporal envelope changes and significant spectral transitions will be preserved from modification. A summary of existing approaches for TSM is given in Table 2.4.

2.5 Summary

In this chapter, we have reviewed some of the existing methods for VOP detection, speech systems in mobile environment, CV recognition in Indian languages, and time scale modification. Existing methods for VOP detection are suffering with poor detection accuracy. Therefore, accuracy issues in the detection of VOP

are the main focus in this work. In contrast to the existing block processing approaches, the methods proposed in this work enhance VOP detection performance by exploiting the spectral energy in glottal closure regions. The goal of this book is to demonstrate the significance of accurate VOP detection for CV recognition, speaker identification, and nonuniform time scale modification.

http://www.springer.com/978-3-319-03115-6

Speech Processing in Mobile Environments

Rao, K.S.; Vuppala, A.K.

2014, XII, 121 p. 35 illus., 2 illus. in color., Softcover
ISBN: 978-3-319-03115-6