Int J Speech Technol (2011) 14:259–272
DOI 10.1007/s10772-011-9101-9

Recognition of consonant-vowel (CV) units under background noise using combined temporal and spectral preprocessing

Anil Kumar Vuppala · K. Sreenivasa Rao · Saswat Chakrabarti · P. Krishnamoorthy · S.R.M. Prasanna

Received: 24 May 2011 / Accepted: 18 July 2011 / Published online: 11 August 2011

Abstract This paper proposes hybrid classification models and preprocessing methods for enhancing the consonant-vowel (CV) recognition in the presence of background noise. Background Noise is one of the major degradation in real-time environments which strongly effects the performance of speech recognition system. In this work, combined temporal and spectral processing (TSP) methods are explored for preprocessing to improve CV recognition performance. Proposed CV recognition method is carried out in two levels to reduce the similarity among large number of CV classes. In the first level vowel category of CV unit will be recognized, and in the second level consonant category will be recognized. At each level complementary evidences from hybrid models consisting of support vector machine (SVM) and hidden Markov models (HMM) are combined for enhancing the recognition performance. Performance of the proposed CV recognition system is evaluated on Telugu broadcast database for white and vehicle noise. The proposed preprocessing methods and hybrid classification models have improved the recognition performance compared to existed methods.

Keywords Recognition of consonant-vowel (CV) units · Noisy speech recognition · Speech enhancement · Combined temporal and spectral processing · Support vector machine (SVM) · Hidden Markov models (HMM)

A.K. Vuppala (✉) · S. Chakrabarti
G. S. Sanyal School of Telecommunications, Indian Institute of Technology, Kharagpur, India
e-mail: anil.vuppala@gmail.com

S. Chakrabarti
e-mail: saswat@ece.iitkgp.ernet.in

K.S. Rao
School of Information Technology, Indian Institute of Technology, Kharagpur, India
e-mail: ksrao@iitkgp.ac.in

P. Krishnamoorthy
Samsung India Software Center, Noida, India
e-mail: krishna.m1@samsung.com

S.R.M. Prasanna
Department of Electronics and Communication Engineering, Indian Institute of Technology, Guwahati, India
e-mail: prasanna@iitg.ernet.in

1 Introduction

The goal of automatic speech recognition is to convert speech into text. Commonly used approach for speech recognition is based on segmenting speech into subword units and labeling them using a subword unit recognizer (Rabiner and Juang 1993). Phonemes are widely used subword units of speech for speech recognition, but recent studies reveal that syllables (combinations of phonemes) are the suitable subword units for speech recognition in Indian languages (Gangashetty 2004; Sekhar 1996). Context-dependent units such as syllables capture significant co-articulation effects and pronunciation variation compared to phonemes. In general, the syllable-like units are of type $C^m V C^n$, where C refers to consonant, V refers to a vowel, m and n refers to the number of consonants preceding and following in a syllable. Among these units, the CV units are the most frequently (around 90%) occurring basic units (Gangashetty 2004) in Indian languages, and hence CV units are considered to carry out this study.

The major issues involved in the recognition of CV units are the large number of CV classes and high similarity among several CV units. Hidden Markov models (HMM) are the commonly used classification models in speech

recognition, but in Gangashetty (2004), Sekhar (1996), Sekhar et al. (2003) authors have reported that multi layer feed forward neural network (MLFFNN) and support vector machines (SVM) work better for recognition of CV units in Indian languages compared to HMM. In this work, two level approach has been proposed for the recognition of CV units in Indian languages. In this proposed approach CV units are divided into five classes (see columns 3–7 in Table 2) based on vowel, to reduce the influence of vowel on recognition of CV units. In the first level of the proposed method vowel category of the CV unit will be recognized, and in the second level consonant category will be recognized. At both levels complimentary evidences from HMM and SVM models are combined for improving the recognition performance (Gangashetty et al. 2005a; Ho et al. 1994). HMMs are developed using maximum likelihood (ML) approach (Rabiner 1989) and SVMs are developed using discriminative learning approach (Burges 1998). Because of the differences in the training methods they may provide complimentary evidences for recognition of highly confusable and large number of CV classes (Gangashetty et al. 2005a; Ho et al. 1994). This complimentary information may be useful to improve recognition performance under noise.

Most of the present speech recognition systems are developed using clean speech. In practical applications of automatic speech recognition clean speech is often distorted by a background noise. Because of this distortion, the speech features are distorted and therefore there is a mismatch between the training (clean) and testing (noisy) conditions. This mismatch strongly degrades the performance of speech recognizers (Mokbel and Chollet 1991; Nolazco-Flores and Young 1993). Various methods have been proposed in the literature to overcome the noise effect on speech recognition. These methods can be grouped under three categories based on (a) compensation of noise, (b) robust feature extraction and (c) adaptation of models. Methods based on compensation of noise aim to enhance the noisy speech signals before feature extraction (Mokbel and Chollet 1991; Nolazco-Flores and Young 1993; Huang and Zhao 1997; Hermus et al. 2000; Hermus and Wambacq 2004; Kris and Patrick 2007). Such methods include spectral subtraction, minimum mean square error (MMSE) and subspace based speech enhancement techniques (Huang and Zhao 1997; Hermus et al. 2000; Hermus and Wambacq 2004; Kris and Patrick 2007). 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 (Hermanski et al. 1994; Viiki et al. 1998; Yu et al. 2008; Cui and Alwan 2005; Hilger and Ney 2006; de la Torre et al. 2005; Suh et al. 2007), e.g, RASTA filter (Hermanski et al. 1994), feature normalization (Viiki et al. 1998), MMSE based Mel-frequency cepstra (Yu et al. 2008), and histogram equalization (Hilger and Ney 2006; de la Torre et al. 2005; Suh et al. 2007) etc. In case of model adaptation approach, the parameters of the model are modified according to the characteristics of the background noise (Ohkura and Sugiyama 1991; Gales et al. 1996; Moreno 1996; Vaseghi and Milner 1997; Liao and Gales 2007; Ozlem et al. 2010; Kim and Gales 2011). Sum of the popular model adaptation methods include code book mapping (Ohkura and Sugiyama 1991), parallel model compensation (Gales et al. 1996), noise adaptive training (Liao and Gales 2007; Ozlem et al. 2010) etc.

This work aims to provide the robustness at the signal level by using combined temporal and spectral processing (Krishnamoorthy and Prasanna 2011) enhancement method as a preprocessing stage and hybrid classification models at modeling stage. In this paper, we demonstrated the effectiveness of noise reduction technique proposed in Krishnamoorthy and Prasanna (2011) for speech recognition task. For that, initially we analyzed the effect of noise on the recognition performance of CV units using white and vehicle noise at different signal-to-noise ratios (SNRs). CV units from Telugu broadcast database are used to carry out this study. Noise samples are collected from NOISEX-92 database. Later CV recognition performance under noise has shown to be improved by using proposed enhancement techniques and hybrid classification models. Temporal noise reduction technique used in this work uses speech specific characteristics, so present work is suitable only for background noises other than speech specific noises like babble noise.

Rest of the paper is organized as follows. Section 2 presents the combined temporal and spectral processing for enhancement of noisy speech. Proposed CV recognition system and experimental setup used for this work are presented in Sect. 3. Section 4 discuss the recognition performance of CV units under noise using hybrid models and preprocessing techniques. Summary and conclusions of the present work are mentioned in Sect. 5.

2 Combined temporal and spectral processing for enhancement of noisy speech

Methods developed in literature for the enhancement of noisy speech are grouped into spectral processing and temporal processing. Spectral processing methods are based on the fact that spectral values of noisy speech will have both speech and noisy components (Bell 1979; Kamath and Loizou 2002; Ephrain and Malah 1984). The spectral characteristics of noise are therefore estimated and removed to obtain the enhancement. Spectral processing methods are popular due to simplicity and effectiveness. The demerit of spectral processing methods is the need for explicit modeling of spectral characteristics of noise. This is difficult for highly non-stationary noise cases. Temporal processing

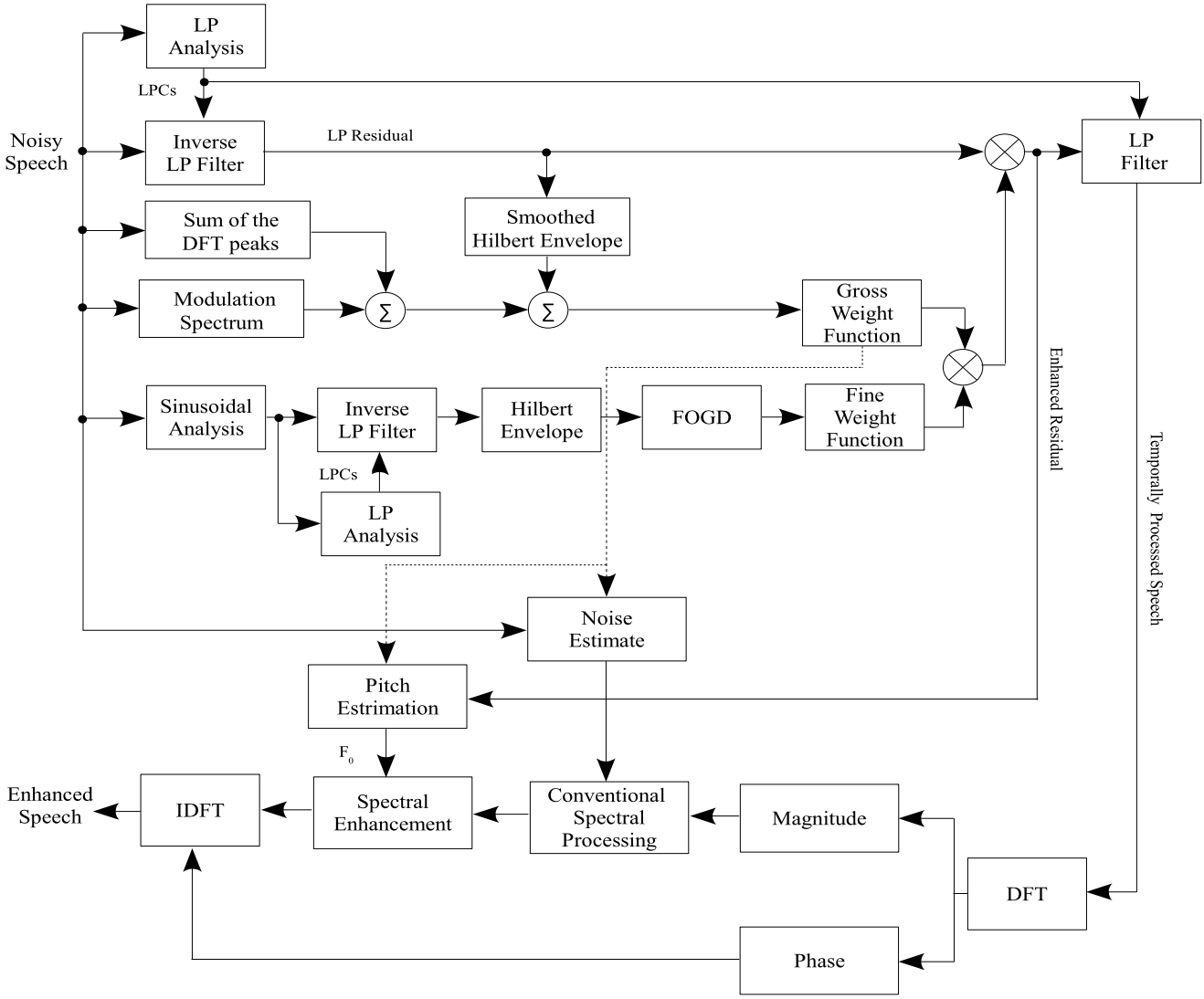

Fig. 1 Block diagram of the combined temporal and spectral noisy speech enhancement method (Krishnamoorthy and Prasanna 2011)

methods are based on identifying and enhancing the speech-specific regions of noisy speech (Yegnanarayana et al. 1999, 2005; Yegnanarayana and Murthy 2000). The merit of temporal processing is in the enhancement of speech specific regions and don't require explicitly modeling of degradation. The demerit may be the ineffectiveness in minimizing the degrading component, since it is not explicitly modeled. It may be possible that one domain of the processing may aid other domain of processing in minimizing the demerit. Therefore, one can effectively combine temporal and spectral processing approaches to obtain improved performance (Krishnamoorthy and Prasanna 2011).

In the proposed temporal and spectral processing methods the enhancement is achieved by identifying and enhancing speech-specific features from the noisy speech present both in the temporal and spectral domains. Block diagram of the combined temporal and spectral noisy speech enhancement method is shown in Fig. 1 (Krishnamoorthy and Prasanna 2011). The temporal processing involves identifying and enhancing the speech-specific features present at the gross and fine temporal levels. The main objective of the gross level processing is to identify and enhance the speech components at the sound units (100–300 ms) level and the objective of the fine level processing is to identify and enhance the speech-specific features at the segmental (10–30 ms) level. The high SNR speech regions at gross level are determined using speech-specific parameters like sum of 10 largest peaks in the discrete Fourier transform (DFT) spectrum, smoothed Hilbert envelope of the LP residual and modulation spectrum values from the noisy speech signal. The motivation behind using these three parameters is that they represent different aspects of the speech production mechanism. The sum of the peaks in the DFT spectrum represents predominantly the vocal tract information

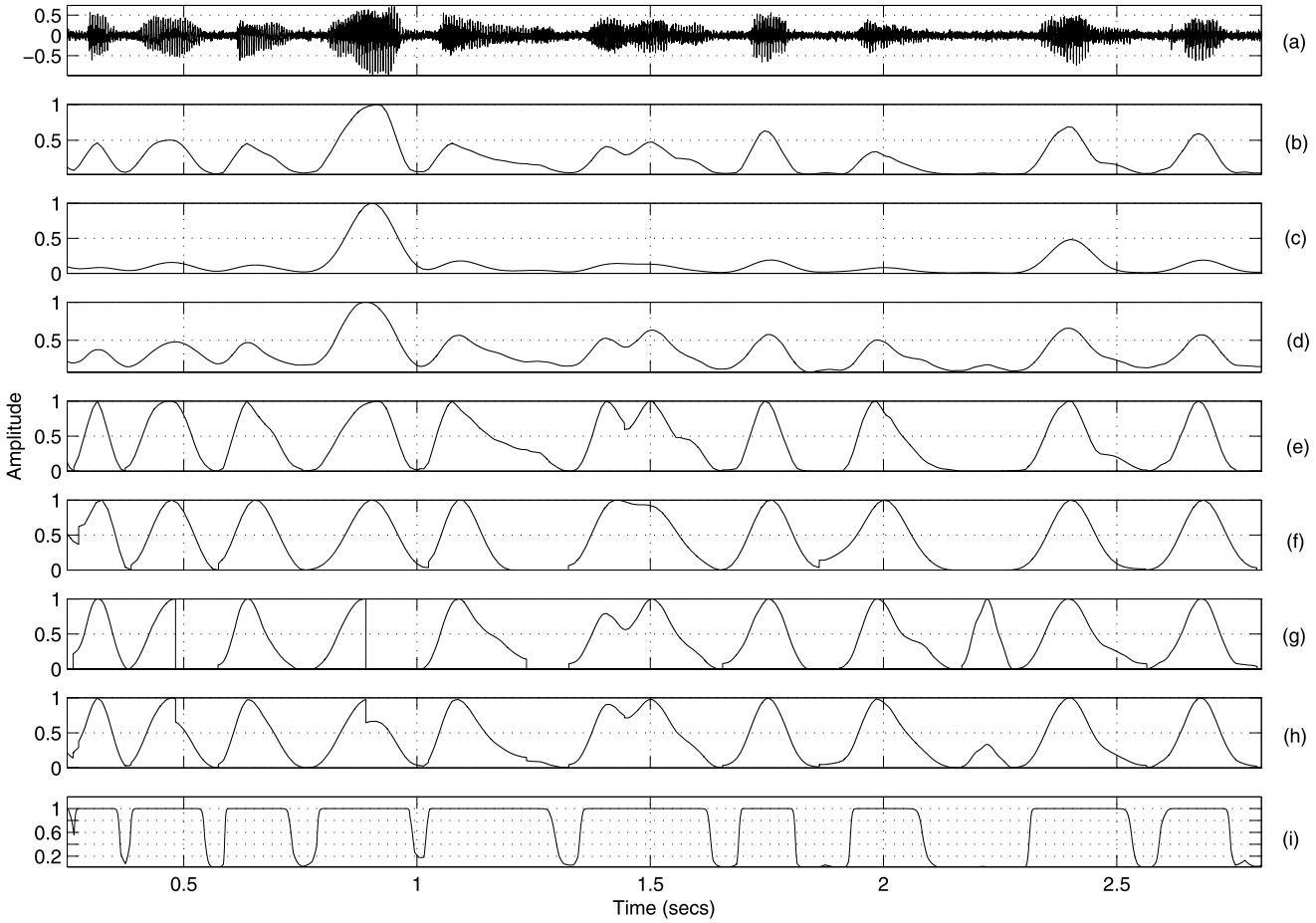

Fig. 2 Computation of gross weight function: (**a**) noisy speech, (**b**) sum of peaks in the DFT spectrum, (**c**) smoothed Hilbert envelope of LP residual, (**d**) modulation spectrum, (**e**) enhanced DFT spectrum values, (**f**) enhanced smoothed Hilbert envelope values, (**g**) enhanced modulation spectrum values, (**h**) normalized sum of (**e**), (**f**) and (**g**); and (**i**) Gross weight function

of speech production. The smoothed Hilbert envelope of the linear predictive (LP) residual represents predominantly the excitation source information of speech production. The modulation spectrum represents the long term (supra segmental) information of speech production. Since the origin of these three parameters is different, combining them may improve the robustness and also the detection accuracy as compared to any one of them.

Figure 2 shows the computation process of gross weight function. Figure 2(a) indicates the noisy speech spectrum. Figure 2(b)–Fig. 2(d) indicate the evidences derived from spectral peaks of the DFT spectrum, Hilbert envelope of the LP residual and modulation spectrum. Figure 2(e)–Fig. 2(g) indicate the enhanced evidences of Fig. 2(b)–Fig. 2(d). Combination of enhanced evidence Fig. 2(e)–Fig. 2(g) are shown in Fig. 2(h). The final gross weight function shown in Fig. 2(i) is derived from Fig. 2(h) by applying the sigmoid nonlinear function. The gross weight function indicate the higher value during speech regions and low values during nonspeech regions. By modifying the noisy speech LP residual with the gross weight function will enhance the residual by deemphasizing the nonspeech or noisy regions.

The high SNR speech-specific features at the fine level are identified by using the knowledge of the instants of significant excitation. The basis for the fine level temporal enhancement is that the voiced speech is produced as a result of excitation of quasi periodic glottal pulses and unvoiced speech is produced as a result of excitation of onset of events like burst and frication. The significant excitation in each glottal cycle takes place at the instant of glottal closure. By locating the instants of significant excitation, it is possible to enhance speech around the instants relative to other regions. A weight function is derived for the LP residual from the instants of significant excitation to enhance the excitation source information around these instants relative to other regions. A final weight function is derived by combining gross and fine weight functions, which is then multiplied with the LP residual of the noisy speech signal to enhance the speech-specific features in the temporal domain.

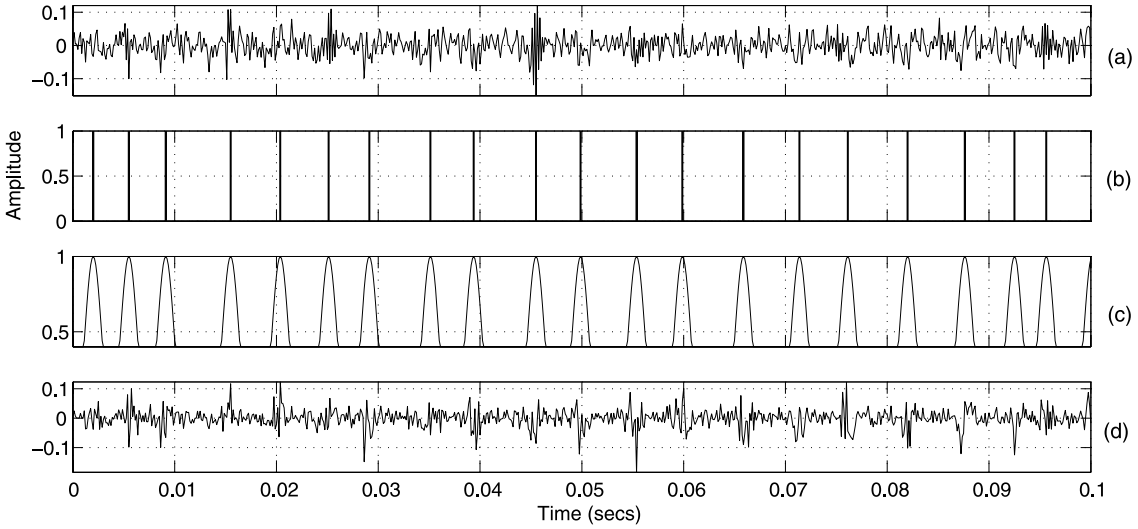

Fig. 3 Computation of fine weight function: (**a**) Degraded LP residual, (**b**) Instants of significant excitation, (**c**) Fine weight function, (**d**) Enhanced LP residual

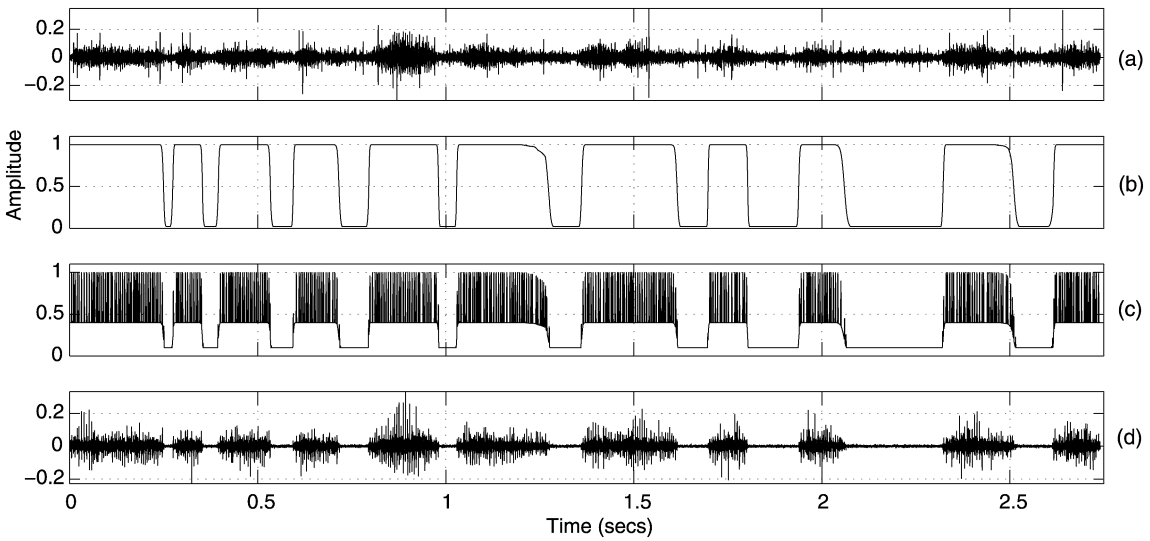

Fig. 4 Overall temporal processing steps (**a**) Degraded speech LP residual, (**b**) Gross weight function, (**c**) Final weight function, (**d**) Enhanced LP residual

Computation of fine weight function is shown in Fig. 3. Figure 3(a) shows the LP residual for a noisy speech segment. Figure 3(b) and Fig. 3(c) indicates the instants of significant excitation and fine weight function respectively for the noisy speech segment show in Figs. 3(a). Figures 3(d) shows the enhanced LP residual after applying the fine weight function. The overall temporal processing method for enhancing the speech is shown in Fig. 4. Figure 4(a) shows the segment of noisy speech. The gross weight function corresponds to the noisy speech segment shown in Fig. 4(a) is given in Fig. 4(b). Figure 4(c) indicate the final weight function consisting of combination of gross and fine weight function. Figure 4(d) shows the enhanced LP residual after applying the final weight function. From the enhanced LP residual, it is observed that LP residual is enhanced at both gross level and final level.

The temporally processed speech sounds to be perceptually enhanced. This is mainly due to the enhancement of speech-specific features in the noisy speech signal. This includes high SNR regions at gross level and regions around the instants of significant excitation. This is achieved by multiplying the LP residual of the noisy speech signal by the weight function. Even though the speech-specific features are emphasized in the temporally processed speech, the noise suppression is minimal mainly due to the use of all-pole filters derived from the noisy speech. To further improve the enhancement level, the speech-specific features corresponding to the all pole filter features are en-

Table 1 Steps in combined temporal and spectral processing method for the enhancement of noisy speech

Temporal Processing:

Gross Level Processing

- Compute Linear Prediction (LP) residual of noisy speech using a frame size of 20 ms, shift of 10 ms and 10^{th} order LP analysis.
- Compute the sum of 10 largest peaks in the Discrete Fourier Transform (DFT) magnitude spectrum.
- Compute the Hilbert Envelope (HE) of LP residual and mean smooth using 50 ms rectangular window.
- Compute the modulation spectrum of the noisy speech signal.
- Enhance the high Signal to Noise Ratio (SNR) regions of each of the above parameters using the First Order Differentiator (FOD).
- Sum all the enhanced parameters and normalize the sum with respect to maximum value.
- Nonlinearly map the normalized sum values by using a sigmoid nonlinear function

$$w_g(n) = \frac{1}{1 + e^{-\lambda(s_i(n)-T)}}$$

where slope parameter $\lambda = 20$ and T equal to average value of the normalized sum $s_i(n)$.

- The nonlinearly mapped values is termed as gross weight function.

Fine Level Processing

- Compute the DFT magnitude and phase spectra for the noisy speech using 1024 point DFT.
- Pick the largest 8 peaks in the DFT magnitude spectrum and corresponding phase values and synthesize the speech.
- Calculate the LP residual of the signal obtained.
- Compute the HE of LP residual.
- Compute the emphasized HE of the LP residual.
- Mean smooth the emphasized HE using 1 ms rectangular window.
- Obtain the First Order Gaussian Differentiator (FOGD) given by

$$g_d(n) = \frac{1}{\sigma\sqrt{2\pi}}\left[e^{-\frac{(n+1)^2}{2\sigma^2}} - e^{-\frac{n^2}{2\sigma^2}}\right], \quad 1 \le n \le L_g$$

where Gaussian window of length $L_g = 80$ samples and variance $\sigma = 8$.

- Convolve the negative of FOGD operator with the mean smoothed HE of the LP residual and determine negative to positive transitions.
- Convolve detected instants with 3 ms Hamming window. The resultant signal is termed as fine weight function.

Final Weight Function

- Multiply the two weight functions (gross and fine weight functions) to generate the final weight function.
- Multiply the LP residual signal of noisy speech by the final weight function.
- Excite the time-varying all-pole filter using weighted residual to obtain the temporally processed speech.

Spectral Processing:

- Update the noise magnitude spectrum if 5 consecutive frames are detected as non-speech regions.
- Process the temporally processed speech by any of the conventional spectral processing (e.g., multi-band spectral subtraction (Bell 1979) or MMSE estimator (Ephrain and Malah 1984) methods.
- Reconstruct the enhanced speech signal using IDFT and overlap-add (OLA) method.

hanced subjected to spectral processing. To improve the vocal tract response characteristics at the spectral level and to provide better noise suppression, the spectral processing is performed on the temporally processed speech which involves conventional spectral enhancement (spectral subtraction (Bell 1979) or minimum mean square error and short time spectral amplitude (MMSE-STSA) (Ephrain and Malah 1984) based methods) and the pitch and harmonics based spectral enhancement. Spectral subtraction based speech enhancement is performed by subtracting the average magnitude of the noise spectrum from the spectrum of the noisy speech (Bell 1979). In this method noise is assumed to be uncorrelated and additive to the speech signal. The noise estimation is obtained based on the assumption that the noise is locally stationary, so that the noise characteristics computed during the speech pauses are a good approximation to the noise characteristics.

$$|\hat{S}(k)| = |Y(k)| - |\hat{D}(k)| \tag{1}$$

where $\hat{D}(k)$ is the average magnitude of the noise spectrum. $Y(k)$ is the spectrum of noisy speech signal and $\hat{S}(k)$ is the estimated enhanced speech signal spectrum. The MMSE-STSA for speech enhancement aims to minimize the mean square error between the short time spectral magnitude of the clean and enhanced speech signal (Ephrain and Malah 1984). This method assumes that each of the Fourier expansion coefficients of the speech and of the noise process can be modeled as independent, zero-mean and Gaussian random variables. The basic steps in the combined temporal and spectral processing method for the enhancement of noisy speech are given in Table 1 (Krishnamoorthy and Prasanna 2011).

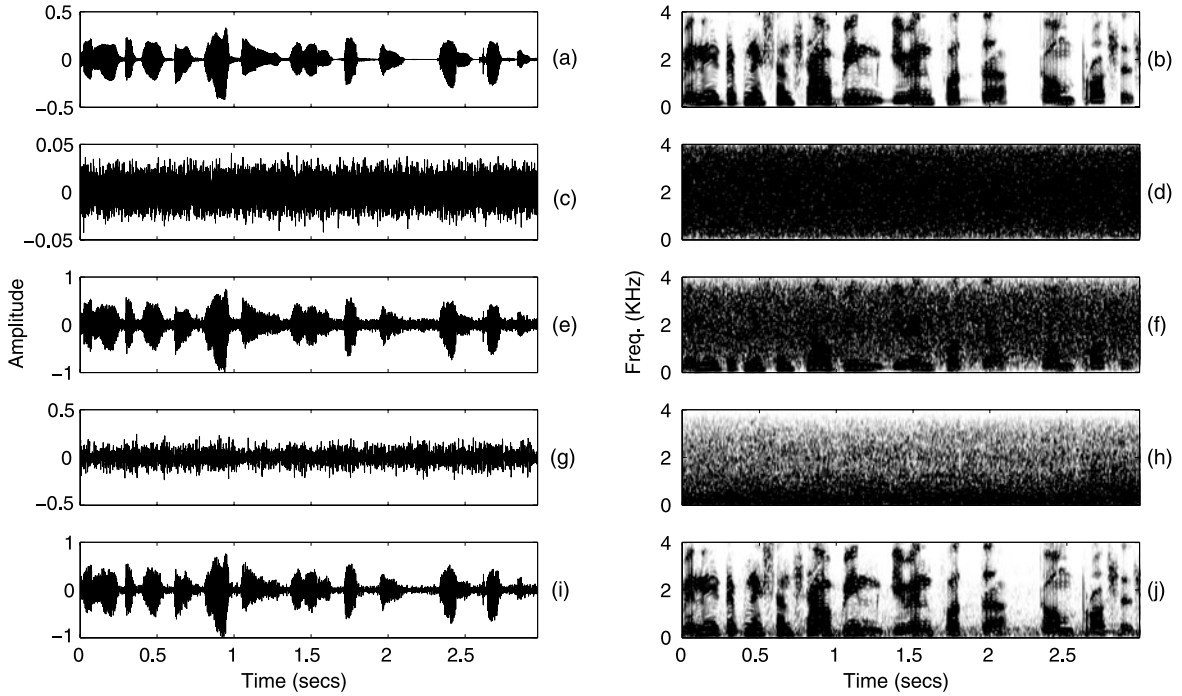

Fig. 5 Noisy speech: (**a**, **b**) clean speech and its spectrogram; (**c**, **d**) white noise and its spectrogram; (**e**, **f**) white noise (SNR 10 dB) added speech and its spectrogram; (**g**, **h**) vehicle noise and its spectrogram; (**i**, **j**) vehicle noise (SNR 10 dB) added speech and its spectrogram

The temporal and spectral details of clean speech, noisy speech and enhanced speech by the proposed temporal and spectral processing method are shown in Figs 5 and Fig. 6. Figure 5 shows the time and frequency domain details of clean speech, noise segments and noisy speech signals. From spectrogram plot of Fig. 5(d) it is observed that, in the case of white noise, noise is present at all frequencies (0–4 KHz) in the uniform manner. From Fig. 5(h), it indicates that the vehicle noise is dominant at low frequencies (i.e., less than 1 KHz), and at higher frequencies (i.e., beyond 1 KHz) its effect is less. Therefore, addition of the noises to clean speech effects more in case of white noise compare to vehicle noise. This is clearly observed from the spectrograms of noisy speech (see Fig. 5(f) and Fig. 5(j)).

Figure 6 shows noisy and enhanced speech signals and their respective spectrograms. In this case noisy speech is derived by adding the clean speech with white noise (SNR of 10 dB). By applying the temporal enhancement speech regions are enhanced by deemphasizing the nonspeech regions. It is observed from Fig. 6(f) that, since speech regions are preserved during temporal processing, the noise also remain along with speech components. Where as nonspeech regions are completely cleaned up because of deemphasizing the nonspeech regions. Spectral processing methods attempt to remove the noise spectral components from noisy speech. Therefore, the spectrograms of the spectrally enhanced speech (see Figs. 6(h) and (j)) indicate the enhanced noisy speech. In combined temporal and spectral processing methods the merits of individual methods are combined, so in Figs. 6 (l) and (n) we can observe the better enhancement compared to individual methods.

3 Proposed consonant-vowel (CV) units recognition system

In this work CV recognition is carried out using the proposed two level approach with the help of hybrid models at each stage. The performance of the CV recognition system is evaluated using Telugu broadcast database. Details of the database and proposed CV recognition system are described in the following subsections.

3.1 Database

The database used in this work contains the broadcast news corpus developed at speech and vision lab, Indian Institute of Technology, Madras, India (Sekhar 1996; Gangashetty 2004; Gangashetty et al. 2005b; Hegde et al. 2004; Rao and Yegnanarayana 2009a, 2009b). Broadcast news corpus is recorded for 3 Indian languages, namely Telugu, Tamil and Hindi. In literature, this broadcast database is used for performing several speech tasks like continuous speech recognition (Hegde et al. 2004), CV recognition (Sekhar 1996; Gangashetty 2004; Gangashetty et al. 2005b), speech synthesis (Rao 2011), duration modeling (Rao and Yegna-

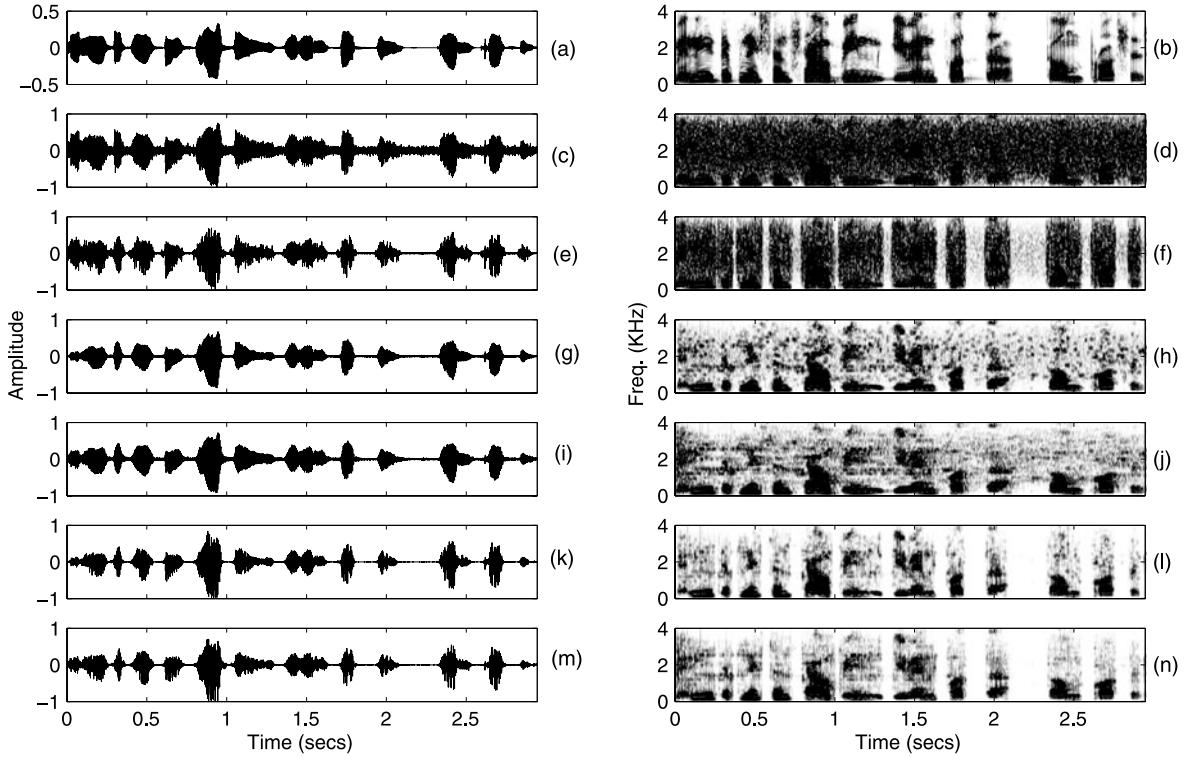

Fig. 6 Noisy speech enhancement:(**a**, **b**) clean speech and its spectrogram; (**c**, **d**) white noisy (SNR of 10 dB) speech and its spectrogram; (**e**, **f**) white noisy speech processed by temporal processing (TP) and its spectrogram; (**g**, **h**) white noisy speech processed by spectral subtraction (SS) method and its spectrogram; (**i**, **j**) white noisy speech processed by MMSE method and its spectrogram; (**k**, **l**) white noisy speech processed by combining TP and SS methods, and its spectrogram; (**m**, **n**) white noisy speech processed by combining TP and MMSE methods, and its spectrogram

narayana 2009b), and intonation modeling (Rao and Yegnanarayana 2009a). Among three languages we consider Telugu language news corpus for this work. Duration of Telugu database is about five hours, collected over 20 sessions by 11 male speakers and 9 female speakers. Among 20 sessions, 15 sessions (8 male + 7 female) are used for training and 5 sessions (3 male + 2 female) are used for testing the CV recognizer. Manually marked syllable boundaries available in database are used for picking the CV units from continuous speech. In the context of Indian languages there are 145 (29 consonants and 5 vowels) most frequently occurred CV units (Gangashetty 2004). In this work, among 145 CV units, 95 CV classes whose frequency of occurrence in the database is more than 50 are considered for the analysis, and their contribution is more than 95% of CV units present in the Database. CV units considered in this work are shown in Table 2. Among 95 CV units, a, e, i, o and u vowel groups contains 26, 16, 22, 10, and 21 consonant classes respectively. We consider both short and long vowels as one vowel only, as it is very difficult to recognize short and long vowels in the continuous speech. Simple language model will take care of short and long vowels during speech recognition. In Fig. 7 an example CV unit $/ka/$ is shown. Different regions of significant events in the production of the CV unit $/ka/$ along with vowel onset point (VOP) (Vuppala et al. 2010) are also shown in Fig. 7. In Fig. 7, VOP is vowel onset point, it is the instant at which onset of vowel takes place in the speech signal. VOP plays anchor role in identifying consonant, transition and vowel regions. In CV segment, the region before VOP is consonant region, and the short duration after VOP is transition region. From transition region to end of CV segment corresponds to vowel region.

3.2 Proposed CV recognition method

High similarity and large number of CV classes are the major issues involved in CV recognition. In this work, we proposed two level approach for the recognition of CV units in Indian languages. In the first level vowel will be recognized, and in the second level consonant will be recognized. In both levels, evidences from HMM and SVM models are combined with appropriate weights. The motivations for the proposed method are described below.

- Monolithic (single level) models may not be appropriate for classification of large number and highly confusable CV classes. Therefore, proposed method suggested two level approach.

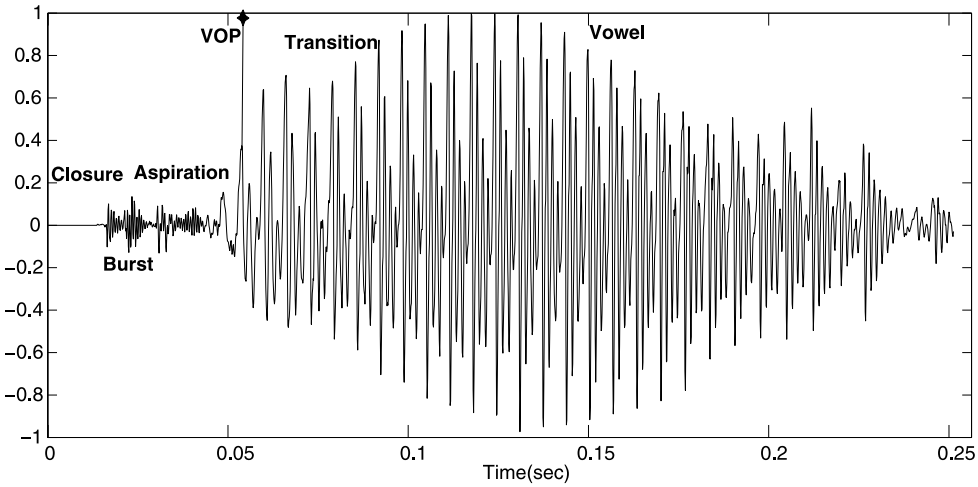

Fig. 7 Regions of significant events in the production of the CV unit /*ka*/ (Vuppala et al. 2010)

Table 2 List of 95 CV units from Telugu broadcast news corpus

Vowel class	CV units
/*a*/	ka, cha, Ta, ta, pa
	kha, Tha, tha, pha
	ga, ja, Da, da, ba
	gha, dha, bha, na, ma
	ya, ra, la, va, ha, sha, sa
/*e*/	ke, che, Te, te, pe
	phe, je, De, de, ne, me
	ye, re, le, ve, se
/*i*/	ki, chi, Ti, ti, pi
	thi, gi, ji, Di, di, bi
	dhi, bhi, ni, mi, yi, ri
	li, vi, hi, shi, si
/*o*/	ko, cho, to, po, do
	mo, yo, ro, lo, so
/*u*/	ku, chu, Tu, tu, pu
	thu, gu, ju, Du, du, bu
	dhu, bhu, nu, mu, yu
	ru, lu, vu, shu, su

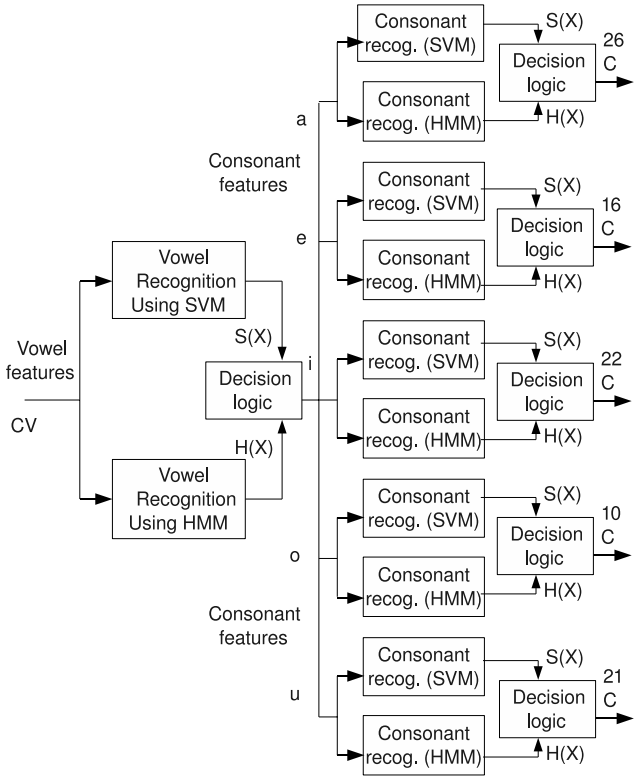

Fig. 8 Proposed CV recognition system using HMM and SVM

- For enhancing the recognition performance, hybrid models are explored at each level to capture the CV characteristics in different ways. In this work the hybrid model consists of combination of HMM and SVM. It is known that HMMs capture distribution and sequential knowledge from the feature vectors of the specific class, where as SVMs capture discriminative characteristics between the desired class and rest of classes by using positive and negative examples from the desired class and rest of the classes respectively. Since these models capture the class specific knowledge based on different modalities, combining the evidence from these models may improve the recognition performance.
- In existed multilayer CV recognition approach (Gangashetty 2004), same features are used at each level for developing the acoustic models. But, in the proposed approach different features are used at each level by using VOP as an anchor point. At the first level, vowel recognition models are developed using features extracted from VOP to end of CV segment (i.e., only vowel region) and at the second level consonant models are developed using features extracted from consonant and transition regions.

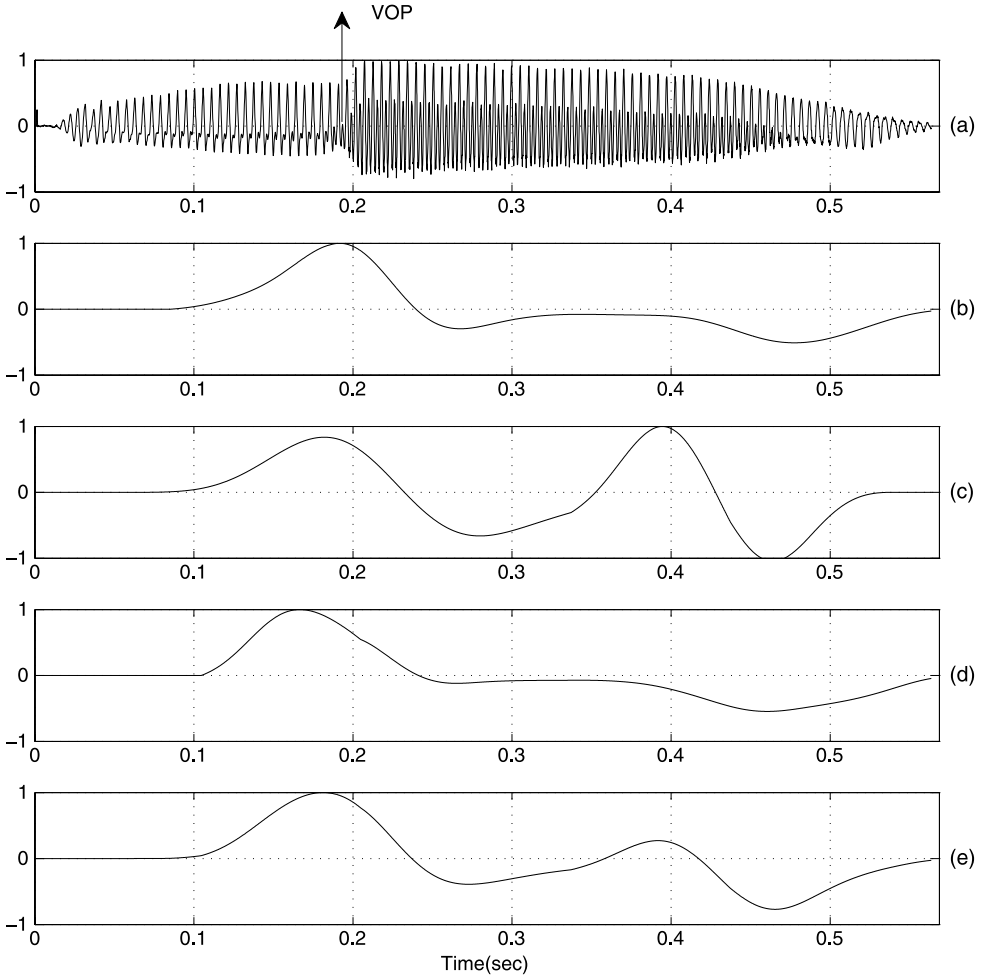

Fig. 9 VOP detection using combination of all three evidences. (**a**) Speech signal of *ni*. VOP evidence plot for (**b**) excitation source. (**c**) Modulation spectrum. (**d**) Spectral peaks. (**e**) Combined VOP evidence plot

In this work, transition region is assumed as 40 ms speech segment to the right of VOP.

Proposed two level CV recognition model is shown in Fig. 8. In the proposed method, at each level decision is taken by combining evidences from SVM and HMM using below equation.

$$Class(X) = \max(w1 * S(X) + w2 * H(X)) \quad (2)$$

where $S(X)$ and $H(X)$ corresponds to normalized evidence scores from the SVM and HMM models respectively for test utterance X. $w1$ and $w2$ ($w2 = 1 - w1$) are the weights given for SVM and HMM evidence scores respectively. In our study, $w1$ is varied in steps of 0.02 from 0 to 1.

Vowel onset point algorithm presented in Prasanna et al. (2009) is used for detecting the VOP in this study. This VOP algorithm uses combined evidences from excitation source, spectral peaks, and modulation spectrum energies. The Hilbert envelope (HE) of the linear prediction (LP) residual represents the excitation source information. The sum of ten largest peaks in the discrete Fourier transform (DFT) spectrum represents the vocal tract shape. The modulation spectrum represents the slowly varying temporal envelope. Thus, each of these three features represents a different aspects of speech production, it may be possible that they contain complementary information about the VOP. The individual evidences are therefore combined for detecting the final VOPs. Location of maximum value in the combined VOP evidence plot of CV unit is considered as final VOP. This Combined VOP algorithm works better compared to individual evidences collected from source, spectral peaks, and modulation spectrum energies (Prasanna et al. 2009; Hermes 1990; Prasanna 2004; Prasanna and Yegnanarayana 2005). VOP detection using individual and combination of all three evidences for speech signal of /*ni*/ is shown in Fig. 9. From the combined VOP evidence plot in Fig. 9 it is observed that, the prediction of VOP by modulation spectrum is eliminated and accuracy of prediction also enhanced. From the experiments it is observed that, combined VOP detection method is detecting VOPs with 15 ms average deviation. From the experimental results it is also observed that, the effect of noise is negligible on detection of VOP from CV units available in the database.

Table 3 Performance of vowel recognition from CV units using SVM, HMM and SVM + HMM acoustic models under different background noise cases

Models	Recognition performance (%)					
		White noise (different SNR levels in dB)				
	Clean	0	5	10	20	30
SVM	86.10	20.12	28.11	36.31	72.66	79.52
HMM	87.22	21.08	33.08	43.8	76.82	81.13
SVM + HMM	**91.82**	**24.2**	**37.52**	**47.72**	**80.12**	**86.12**
		Vehicle noise (different SNR levels in dB)				
SVM	86.10	68.88	73.34	80.42	82.12	84.52
HMM	87.22	72.66	77.41	81.66	82.72	85.31
SVM + HMM	**91.82**	**77.14**	**81.42**	**85.22**	**85.46**	**88.66**

Thirteen Mel-frequency cepstral coefficients (MFCC) (Picone 1993) extracted from a frame of 20 ms with frame shift of 5 ms are used for developing the acoustic models. HMM models are developed using maximum likelihood approach using HMM tool kit (htk) (Young et al. 2000). Feature vectors of size 39 dimension (13 MFCC + delta + delta-delta coefficients) are used for developing of HMM models. In the proposed method HMM models are developed using 3 states and 64 mixtures.

SVM models are developed using one against the rest approach using open source SVMTorch (Collobert and Bengio 2001). Fixed pattern length of 10 and Gaussian kernel with standard deviation of 40 are used to build SVM models. Fixed pattern length is obtained by using below equation.

$$s = (p * SL)/PL, \quad p = 0, 1, \ldots, PL - 1, \text{ and}$$
$$s = 0, 1, \ldots, SL - 1. \tag{3}$$

Where PL is pattern length, and SL is segment length. If segment length SL is greater than PL, few frames of the segment are omitted. If the segment length SL is smaller than PL, few frames of the segment are repeated. So, from each CV utterance a fixed 130 ($10PL * 13$ MFCC $= 130$) dimension feature vector is extracted for developing the SVM models.

4 Consonant-vowel (CV) units recognition under background noise

Effect of noise on the performance of CV recognition system is studied by using Telugu broadcast news database at different signal-to-noise ratios (SNRs) for two different noise types. Total number of CV utterances considered in this study are 52,703, among those (38,729 are used for training, and 13,974 are used for testing). Noises considered in this study are white and vehicle noises from NOISEX-92 database. Proposed two-level CV recognition method is used to carried out this study. In the case of noisy speech recognition, first speech enhancement is carried out, and the recognition is performed using enhanced speech. In first level of proposed method vowel will be recognized and in second level consonant will be recognized. Table 3 shows the vowel recognition performance from CV units by using HMM, SVM and HMM + SVM acoustic models. In this work models are trained with clean speech and tested with noisy speech. From the results (see Table 3), it is observed that the performance of HMM models seems to be better compared to SVM models for vowel recognition. This is because HMMs are good at capturing the state sequence corresponds to the sequence of vocal tract shapes. The sequences of vocal tract shapes are unique to each vowel. Other reason is, we are building 5 vowel classes by using training data from all 95 CV classes, so amount of training data available for vowel classes is enough to capture the distributions present in vowel features by HMM. From the results, we can also observe that the combination of evidences with appropriate weights has shown the improvement over individual evidences. It is observed that, recognition performance has effected significantly due to noise, and its impact is more at lower SNRs (higher noise levels). Among the two noises considered, effect of white noise is more compared to vehicle noise. It is due to the fact that, white noise effects entire speech spectrum, where as vehicle noise effects only at low frequencies. Table 4 shows the overall consonant recognition. From the results, it is observed that the performance of SVM models seems to be better compared to HMM models for consonant recognition. This is because, SVM models are trained using one against rest approach to capture the discriminative information present in highly similar consonant classes, and also SVMs are known for capturing the discriminative information with less number of examples compared to HMM. The amount of training data available for consonant recognition is less, so SVMs are giving higher performance for consonant recognition compared to HMMs.

Performance of the proposed CV recognition system is compared using existed monolithic SVM method (Gangashetty 2004; Gangashetty et al. 2005b) and results are shown in Table 5. From the results, it is observed that there is

Table 4 Performance of consonant recognition from CV units using SVM, HMM and SVM + HMM acoustic models under different background noise cases

Models	Recognition performance (%)					
		White noise (different SNR levels in dB)				
	Clean	0	5	10	20	30
SVM	66.38	15.96	20.76	27.06	38.75	44.98
HMM	58.9	8.9	16.06	23.46	36.39	42.17
SVM + HMM	**72.57**	**18.44**	**23.07**	**31.88**	**46.23**	**51.20**
		Vehicle noise (different SNR levels in dB)				
SVM	66.38	28.17	36.88	44.78	54.23	58.38
HMM	58.29	27.44	34.81	41.20	49.75	53.34
SVM + HMM	**72.57**	**32.84**	**42.07**	**50.59**	**63.33**	**68.21**

Table 5 Overall CV recognition using monolithic SVM and proposed acoustic models under different background noise cases

Models	Recognition performance (%)					
		White noise (different SNR levels in dB)				
	Clean	0	5	10	20	30
Monolithic SVM (Gangashetty 2004; Gangashetty et al. 2005b)	53.70	5.85	10.12	15	28.36	41.97
Proposed	66.64	4.46	8.65	15.21	37.01	44.09
		Vehicle noise (different SNR levels in dB)				
Monolithic SVM (Gangashetty 2004; Gangashetty et al. 2005b)	53.70	15.13	23.52	32.20	43.72	47.50
Proposed	66.64	25.33	34.25	43.11	54.12	60.47

12% improvement in recognition performance by using proposed two level CV recognition method compared to monolithic SVM method (see Table 5). It is observe that, performance of proposed CV recognition system is superior compared to monolithic SVM even under noisy case except low SNR values of white noise.

Table 6 shows the overall CV recognition performance using proposed CV recognition method under different background noise cases using speech enhancement techniques. In table abbreviations DEG, SS, MMSE, TP, TSP1 and TSP2 refer to degraded speech, multi band spectral subtraction, MMSE-STSA estimator, temporal processing, combined temporal and multi-band spectral subtraction and combined temporal and MMSE-STSA estimator, respectively. From Table 6, it is observed that spectral processing methods provided much better improvement in the recognition performance compared to temporal processing method. This is because, spectral methods enhance the noisy speech by filtering out the noise spectrum. Hence, the enhanced spectrum mostly contains speech characteristics. Where as temporal processing methods enhance the speech by processing high SNR speech regions, and attenuate all other regions. With this effect, the noisy speech is perceptually enhanced, but the presence of noise with in speech regions will significantly degrade the performance. Hence, in our study we have observed an improvement of 2–7% and 4–22% in the recognition performance using temporal method and spectral methods respectively. From the results it is also evident, combined temporal and spectral processing techniques are giving better performance compared to individual enhancement methods. Because merits of both temporal and spectral methods are combined in combined TSP.

5 Summary and conclusion

In this paper, we proposed hybrid classification models and preprocessing methods for enhancing the CV units recognition performance under background noise. Proposed CV recognition method consists of two levels, and at each level the combination of complimentary evidences from SVM and HMM are used to improve the recognition performance. In this study, the proposed CV recognition system is evaluated using Telugu broadcast database. Initially, the effect of noise on CV recognition is observed by conducting experiments on white and vehicle noise added test data us-

Table 6 Overall CV recognition using proposed CV recognition method under different background noise cases using speech enhancement techniques. In table abbreviations DEG, SS, MMSE, TP, TSP1 and TSP2 refer to degraded speech, multi band spectral subtraction, MMSE-STSA estimator, temporal processing, combined temporal and multi-band spectral subtraction and combined temporal and MMSE-STSA estimator, respectively

	Recognition performance (%)				
	Clean performance is **66.64**				
	White noise (different SNR levels in dB)				
Enhancement	0	5	10	20	30
DEG	4.42	8.66	15.12	36.85	43.90
SS (SP1)	24.12	30.25	34.11	44.36	51.97
MMSE (SP2)	26.81	32.20	38.42	46.06	54.76
TP	7.33	10.52	22.66	41.32	47.41
TSP1	27.46	36.39	40.43	48.38	57.16
TSP2	29.72	35.31	41.10	49.12	57.71
	Vehicle noise (different SNR levels in dB)				
DEG	25.12	33.85	43.07	53.85	60.08
SS (SP1)	34.23	41.34	47.81	56.20	62.21
MMSE (SP2)	36.07	42.59	49.75	58.31	62.08
TP	26.08	35.12	44.57	55.12	60.88
TSP1	37.24	45.46	52.39	59.24	63.06
TSP2	38.59	47.52	54.20	61.72	63.91

ing proposed CV recognition method and existed monolithic SVM method. From the results it is observed that proposed CV recognition method is giving better performance compared to monolithic SVM method. Further, performance of CV recognition system under background noise is improved by using combined temporal and spectral processing based preprocessing methods. The recognition results show that combined TSP methods gives relatively higher performance compared to individual methods. Future research may be carried out to study this combined TSP preprocessing approach for more challenging large vocabulary speech recognition tasks. It will also be interesting to study the combination of combined TSP preprocessing methods with conventional feature compensation and model adaptation methods.

References

Bell, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*, 113–120.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2).

Collobert, R., & Bengio, S. (2001). Svmtorch: support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, *1*, 143–160.

Cui, X., & Alwan, A. (2005). Noise robust speech recognition using feature compensation based on polynomial regression of utterance snr. *IEEE Transactions on Speech and Audio Processing*, *13*(6), 1161–1172.

de la Torre, A., Peinado, A. M., Segura, J. C., Perez-Cordoba, J. L., Benitez, M. C., & Rubio, A. J. (2005). Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, *13*(3), 355–366.

Ephrain, Y., & Malah, D. (1984). Speech enhancement using minimum mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*, 1109–1121.

Gales, M., Young, S., & Young, S. J. (1996). Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, *4*(5), 352–359.

Gangashetty, S. V. (2004). *Neural network models for recognition of consonant-vowel units of speech in multiple languages*. Ph.D. dissertation, IIT Madras, October.

Gangashetty, S. V., Sekhar, C. C., & Yegnanarayana, B. (2005a). Combining evidence from multiple classifiers for recognition of consonant-vowel units of speech in multiple languages. In *Proc. of ICISIP* (pp. 387–391).

Gangashetty, S. V., Sekhar, C. C., & Yegnanarayana, B. (2005b). Spotting multilingual consonant-vowel units of speech using neural networks. In *An ISCA tutorial and research workshop on nonlinear speech processing* (pp. 287–297).

Hegde, R. M., Murthy, H. A., & Gadde, V. (2004). Continuous speech recognition using joint features derived from the modified group delay function and mfcc. In *Proc. INTERSPEECH-ICSLP* (pp. 905–908).

Hermanski, H., Morgan, N., & Hirsch, H. G. (1994). Recognition of speech in additive and convolutional noise based on rasta spectral processing. In *Proc. IEEE int. conf. acoust., speech, signal process*.

Hermes, D. J. (1990). Vowel onset detection. *The Journal of the Acoustical Society of America*, *87*, 866–873.

Hermus, K., & Wambacq, P. (2004). Assessment of signal subspace based speech enhancement for noise robust speech recognition. In *Proc. IEEE int. conf. acoust., speech, signal process* (pp. 945–948).

Hermus, K., Verhelst, W., & Wambacq, P. (2000). Optimized subspace weighting for robust speech recognition in additive noise environments. In *Proc. of 6th international conference on spoken language processing* (pp. 542–545).

Hilger, F., & Ney, H. (2006). Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(3), 845–854.

Ho, T. K., Hull, J. J., & Srihari, S. N. (1994). Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *16*(1), 66–75.

Huang, J., & Zhao, Y. (1997). Energy-constrained signal subspace method for speech enhancement and recognition. *IEEE Signal Processing Letters*, *4*, 283–285.

Kamath, S., & Loizou, P. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Proc. IEEE int. conf. acoust., speech, signal process*, Orlando, USA.

Kim, D. K., & Gales, M. J. F. (2011). Noisy constrained maximum-likelihood linear regression for noise-robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(2), 315–325.

Kris, H., Patrick, W., & ham Hugo, V. (2007). A review of signal subspace speech enhancement and its application to noise robust speech recognition. *EURASIP Journal on Applied Signal Processing*, 195–209.

Krishnamoorthy, P., & Prasanna, S. R. M. (2011). Enhancement of noisy speech by temporal and spectral processing. *Speech Communication*, *53*, 154–174.

Liao, H., & Gales, M. J. F. (2007). Adaptive training with joint uncertainty decoding for robust recognition of noisy data. In *Proc. IEEE int. conf. acoust., speech, signal process* (pp. 389–392).

Mokbel, C., & Chollet, G. (1991). Speech recognition in adverse environments: speech enhancement and spectral transformations. In *Proc. IEEE int. conf. acoust., speech, signal process*.

Moreno, P. J. (1996). *Speech recognition in noisy environments*. Ph.D. dissertation, Carnegie Mellon University.

Nolazco-Flores, J. A., & Young, S. (1993). *CSS-PMC: a combined enhancement/compensation scheme for continuous speech recognition in noise* (Technical Report). Cambridge University Engineering Department.

Ohkura, K., & Sugiyama, M. (1991). Speech recognition in a noisy environment using a noise reduction neural network and a codebook mapping technique. In *Proc. IEEE int. conf. acoust., speech, signal process*.

Ozlem, K., Michael, L. S., Jasha, D., & Alex, A. (2010). Noise adaptive training for robust automatic speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 1889–1901.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, *81*(9), 1215–1247.

Prasanna, S. M. (2004). *Event-based analysis of speech*. Ph.D. dissertation, IIT Madras, March.

Prasanna, S. R. M., & Yegnanarayana, B. (2005). Detection of vowel onset point events using excitation source information. In *Proc. of interspeech* (pp. 1133–1136).

Prasanna, S. M., Reddy, B. S., & Krishnamoorthy, P. (2009). Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(4), 556–565.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proc. of IEEE* (pp. 257–286).

Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs: Prentice Hall.

Rao, K. S. (2011). Application of prosody models for developing speech systems in Indian languages. *International Journal of Speech Technology*, *14*(1), 19–33.

Rao, K. S., & Yegnanarayana, B. (2009a). Intonation modeling for Indian languages. *Computer Speech & Language*, *23*(2), 240–256.

Rao, K. S., & Yegnanarayana, B. (2009b). Duration modification using glottal closure instants and vowel onset points. *Speech Communication*, *51*, 1263–1269.

Sekhar, C. C. (1996). *Neural network models for recognition of stop consonant-vowel (scv) segments in continuous speech*. Ph.D. dissertation, IIT Madras.

Sekhar, C. C., Lee, W. F., Takeda, K., & Itakura, F. (2003). Acoustic modeling of subword units using support vector machines. In *Proceedings of WSLP*.

Suh, Y., Ji, M., & Kim, H. (2007). Probabilistic class histogram equalization for robust speech recognition. *IEEE Signal Processing Letters*, *14*(4), 287–290.

Vaseghi, S. V., & Milner, B. P. (1997). Noise compensation methods for hidden Markov model speech recognition in adverse environments. *IEEE Transactions on Speech and Audio Processing*, *5*(1), 11–21.

Viiki, O., Bye, B., & Laurila, K. (1998). A recursive feature vector normalization approach for robust speech recognition in noise. In *Proc. IEEE int. conf. acoust., speech, signal process*.

Vuppala, A. K., Chakrabarti, S., & Rao, K. S. (2010). Effect of speech coding on recognition of consonant-vowel (CV) units. In *Proc. int. conf. contemporary computing*. *Springer communications in computer and information science* (pp. 284–294).

Yegnanarayana, B., & Murthy, S. (2000). Enhancement of reverberant speech using lp residual signal. *IEEE Transactions on Speech and Audio Processing*, *8*, 267–281.

Yegnanarayana, B., Avendano, C., Hermansky, H., & Murthy, S. (1999). Speech enhancement using linear prediction residual. *Speech Communication*, *28*, 25–42.

Yegnanarayana, B., Prasanna, S. R. M., Duraiswami, R., & Zotkin, D. (2005). Processing of reverberant speech for time-delay estimation. *IEEE Transactions on Speech and Audio Processing*, *13*, 1110–1118.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2000). *The HTK book version 3.0*. Cambridge: Cambridge University Press.

Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y., & Acero, A. (2008). A minimum-mean-square-error noise reduction algorithm on Mel-frequency cepstra for robust speech recognition. In *Proc. IEEE int. conf. acoust., speech, signal process* (pp. 4041–4044).