

Improvements in the Detection of Vowel Onset and Offset Points in a Speech Sequence

Avinash Kumar¹ · S. Shahnawazuddin¹ · Gayadhar Pradhan¹

Received: 20 January 2016 / Revised: 25 August 2016 / Accepted: 27 August 2016 /
Published online: 8 September 2016
© Springer Science+Business Media New York 2016

Abstract Detecting the vowel regions in a given speech signal has been a challenging area of research for a long time. A number of works have been reported over the years to accurately detect the vowel regions and the corresponding vowel onset points (VOPs) and vowel end points (VEPs). Effectiveness of the statistical acoustic modeling techniques and the front-end signal processing approaches has been explored in this regard. The work presented in this paper aims at improving the detection of vowel regions as well as the VOPs and VEPs. A number of statistical modeling approaches developed over the years have been employed in this work for the aforementioned task. To do the same, three-class classifiers (vowel, nonvowel and silence) are developed on the TIMIT database employing the different acoustic modeling techniques and the classification performances are studied. Using any particular three-class classifier, a given speech sample is then forced-aligned against the trained acoustic model under the constraints of first-pass transcription to detect the vowel regions. The correctly detected and spurious vowel regions are analyzed in detail to find the impact of semivowel and nasal sound units on the detection of vowel regions as well as on the determination of VOPs and VEPs. In addition to that, a novel front-end feature extraction technique exploiting the temporal and spectral characteristics of the excitation source information in the speech signal is also proposed. The use of the proposed excitation source feature results in the detection of vowel regions that are quite different

✉ Gayadhar Pradhan
gdp@nitp.ac.in
Avinash Kumar
k.avinash@nitp.ac.in
S. Shahnawazuddin
s.yed@nitp.ac.in

¹ Department of Electronics and Communication Engineering, National Institute of Technology Patna, Patna 800005, India

from those obtained through the mel-frequency cepstral coefficients. Exploiting those differences in the obtained evidences by using the two kinds of features, a technique to combine the evidences is also proposed in order to get a better estimate of the VOPs and VEPs. When the proposed techniques are evaluated on the vowel–nonvowel classification systems developed using the TIMIT database, significant improvements are noted. Moreover, the improvements are noted to hold across all the acoustic modeling paradigms explored in the presented work.

Keywords Phone recognition system · Vowel onset point · Vowel end point · SGMM · DNN

1 Introduction

The vowel onset point (VOP) and the vowel end point (VEP) are the instants of starting and ending of a vowel region in a speech sequence, respectively [10, 28, 35, 40]. The change in the excitation source and the vocal tract system is predominantly reflected at these instants. The vowels are the prominent regions in a speech signal due to their larger amplitude, periodicity and longer duration [34]. Considering these aspects of speech production, several methods have been proposed in the literature to detect the vowel regions and their corresponding VOPs and VEPs. The vowels can be detected by anchoring these two events. On the other hand, the VOPs and the VEPs can be identified from the detected vowel regions by finding the starting and ending points. The former approach is mostly used in the explicit signal processing techniques [28, 33, 35–37, 40], while the latter in the statistical modeling approaches [4, 15, 30, 38]. The characteristics of both the vocal tract system and the excitation source are better manifested in the vowel regions [25]. The accurate detection of the vowel regions, the VOPs and the VEPs is employed in extracting different levels of features that are robust to the environmental degradation. Such features are preferred in the development of various speech-based applications [6, 18, 24–26, 35].

The features mostly employed for the detection of the VOPs, the VEPs and the vowel regions include various signal processing approaches like the difference in the energy of each of the peaks and their corresponding valleys in the amplitude spectrum [10], the zero-crossing rate, the energy and the pitch information of the speech signal [38], the wavelet scaling coefficients of the input speech signal [39], the Hilbert envelop of the linear prediction (LP) residual [29], the spectral peaks, the modulation spectrum energies [28] and the spectral energy present in the glottal closure region of the speech signal [35]. The statistical modeling methods like the Hierarchical neural network, the multilayer feed-forward neural network (MLFFNN) and the auto-associative neural network (AANN) have also been used [30, 38]. These models are generally trained on the features estimated using the speech frames around the VOPs.

The transition characteristics of the vowels vary with the context of the spoken utterance and the environmental conditions [27, 34]. For example, a fricative to vowel transition is completely different from that of a semivowel to vowel transition. Due to the similarities in the production characteristics of the vowels and the semivowels, most of the vowel detection algorithms fail to detect the semivowels, the VOPs and

the VEPs for the semivowel–vowel clusters and the diphthongs. Due to the aforementioned shortcomings, several signal processing methods [25,27] and statistical modeling techniques [4,15] have been explored in the literature for detecting the vowel-like regions (VLRs) instead of the vowel regions. The VLRs are defined as the regions corresponding to the vowel, the semivowel and the diphthong sound units [27].

The existing signal processing methods based on the transition characteristics are generally threshold dependent. In most of those methods, the VOPs and the VEPs are detected by convolving the features characterizing the temporal variations with a first-order Gaussian difference (FOGD) operator within a region that is 100 ms in duration [28,35–37,40]. In those works, it is generally assumed that for a continuous speech utterance, only one vowel will be present within a duration of 100 ms. The convolved output is then used as the evidence for the detection of the VOPs and the VEPs. In such approaches, the convolved output mainly depends on the 100 ms regions under consideration. As a result, most of the weak transitions are smoothed out. On the other hand, performing the convolution in a smaller region will lead to spurious detections. In order to overcome this issue, threshold-independent vowel detection systems should be developed by statistically modeling the vocal tract system, the excitation source and their transient behavior. Motivated by this, an excitation-based feature is proposed in this work to extract the temporal and spectral characteristics of the excitation source information.

In the presented work, an attempt is made to accurately detect the vowel regions by exploiting the different acoustic modeling techniques reported in the literature. For learning the acoustic model parameters, the hidden Markov model (HMM) is explored in this paper. At the same time, different techniques viz. the Gaussian mixture modeling (GMM), the subspace GMM (SGMM) [22] and the deep neural network (DNN) [3] are employed to model the observation densities for the HMMs. In addition to that, the feature and the speaker normalization techniques like the linear discriminant analysis (LDA) [12], the maximum likelihood linear transform (MLLT) [8] and the feature-space maximum likelihood linear regression (fMLLR) [5] are also explored. For the detection of the vowels, three-class classifiers (vowel, nonvowel and silence) are developed using the aforementioned acoustic modeling techniques. In the present work, the speech sound units excluding the vowel are termed as nonvowels. The classifiers are developed on the TIMIT database [9] using the proposed excitation features. In order to detect the vowel regions, the given speech sample is forced-aligned against a particular acoustic model. The first-pass hypothesis, generated by decoding the given data on the same acoustic model, is employed for forced alignment to simulate a realistic scenario. To determine the impact of semivowel and nasal sound units on the detection of vowel regions, the correctly detected and spurious vowel regions are also analyzed in detail.

The above-discussed studies are repeated on systems developed using the conventional mel-frequency cepstral coefficients (MFCC) for the sake of contrast. Interestingly, the vowel regions detected using the two kinds of features are observed to be quite different, i.e., the evidences do not completely overlap. Motivated by that, a novel scheme to combine the obtained evidences is also proposed in this work. The proposed approach for combining the evidences is found to significantly improve the

accuracy with which the vowel regions and their corresponding VOPs and VEPs are detected. The salient contributions of this study can be summarized as follows:

- (a) Exploring state-of-the-art statistical modeling approaches for the task of detecting the vowel regions in a speech sequence.
- (b) A novel front-end speech parameterization approach based on the temporal and the spectral characteristics of the excitation source information.
- (c) In order to enhance the accuracy of detecting the vowel regions, a technique to combine the evidences obtained with respect to the MFCCs and the excitation source features is also proposed.

The rest of the paper is organized as follows: The proposed excitation source features for the detection of vowel regions are discussed in Sect. 2. The experimental evaluation of the vowel detection and the detailed analysis of the detected vowel regions using the VOPs and VEPs are presented in Sects. 3 and 4, respectively. The proposed approach is compared with some of the existing VOP/VEP detection techniques in Sect. 5. Finally, the paper is concluded in Sect. 6.

2 Excitation Source Features for the Detection of Vowel Regions

It is well known that the vowels in the speech signal are mostly produced by the vibration of the vocal folds [34]. Due to a sudden closure of the vocal folds during the production of vowels, the excitation is observed to be impulse like. The strength of excitation in these regions is relatively higher when compared to other consonants. Taking this aspect of speech production into account, several signal processing methods have been proposed in the literature for the detection of the vowel regions by exploiting the excitation source information [18, 24, 25, 28]. Mostly, the variations in the energy associated with the linear prediction (LP) residual and the Hilbert envelop (HE) of the LP residual are used as the features. In general, the excitation strength at the start as well as at the end of the vowel regions in speech is characterized by a sudden change in energy. Consequently, the change in energy at the point where the vowel onset is observed can be used for detecting the vowels in speech [28]. Furthermore, the optimal threshold for the vowel–nonvowel classification generally varies with the context of the spoken utterance and the environmental conditions as mentioned earlier. For example, the variation in the excitation characteristics for a high energy voiced consonant to vowel transition is completely different from that for an unvoiced consonant to vowel transition. Therefore, a threshold-independent vowel–nonvowel classifier may be developed by statistically modeling the excitation characteristics of the vowel regions.

In the existing approaches based on the excitation source, the composite LP residual signal is processed only in the temporal domain. The variation of the excitation strength in different frequency bands is completely neglected. It is to note that the temporal and the spectral characteristics of the excitation source vary in different frequency bands [20, 41]. The considered frequency bands in this work are derived by splitting the analysis range of 0–4 kHz into 8 nonoverlapping bands of bandwidth 500 Hz each. This is achieved by filtering the LP residual signal through a bank of band-pass filters each having a bandwidth of 500 Hz. Narrowing the bandwidth does not provide more

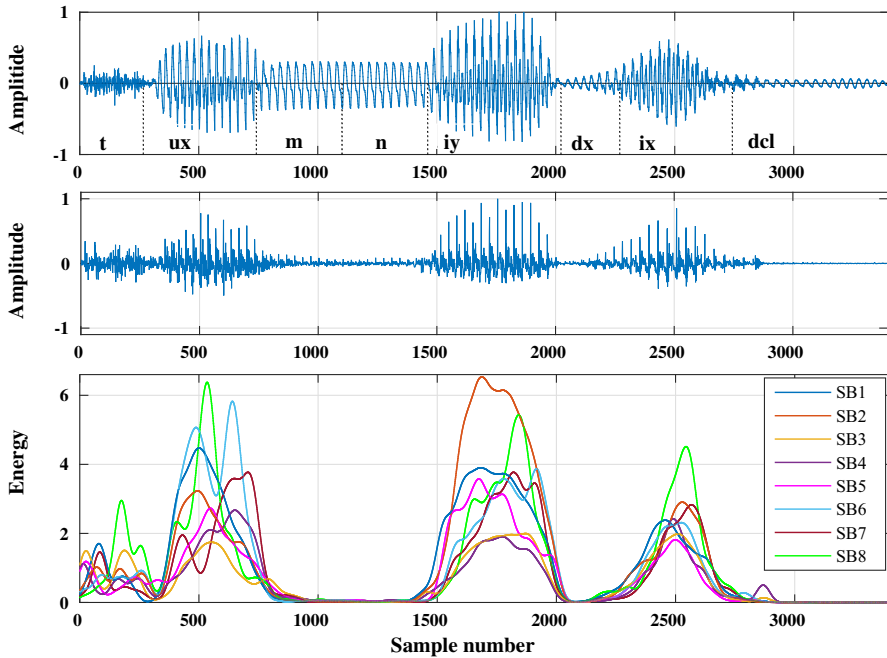


Fig. 1 The *top panel* shows a speech segment along with the reference marking of different sound units (*dotted lines*). The *middle panel* represents the corresponding LP residual signal. The *bottom panel* shows the spectral energy in the LP residual corresponding to the 8 frequency bands considered in this work. The frequency bands are derived by splitting the 0–4 kHz analysis range into 8 nonoverlapping bands of bandwidth 500 Hz each

discriminating information since there is not much variation in the dynamic range of the LP residual spectrum. At the same time, increasing the number of filters results in an increase in the number of coefficients in the feature vector. This, in turn, increases the complexity of the classifiers. On the other hand, increasing the bandwidth results in a degradation of the discriminative property of the features. For example, the energy of the nasal sound units lies prominently in the 0–500 Hz region. Consequently, if the chosen bandwidth is much greater than 500 Hz, the energy of some other sound unit will be captured along with that of the nasals. The feature in that case will represent not only the nasal, but also some other sound unit. This results in a degradation of discrimination due to the merging of the energies being analyzed. The choice of 500 Hz is found to be more suitable through preliminary experimental studies. It is to note that a slight variation in the bandwidth does not hamper the effectiveness of the derived feature vectors much.

The variation in the spectral energy in different frequency bands is shown in Fig. 1. In the case of vowel regions (*/ux/*, */iy/* and */ix/*), a greater degree of variation in the spectral energies for the considered frequency bands is evident from the figure. On the other hand, significantly less variation is noticeable for the nonvowel regions (*/t/*). Due to the nature of the LP residual signal (shown in the middle panel), the sub-band energies for the nasal units (*/m/* and */n/*) are observed to be insignificantly small

in comparison with the vowels. Therefore, the features for the statistical modeling of the excitation source information within vowels may be obtained by considering these variations in the different sub-bands. Consequently, such acoustic features are expected to be more discriminative.

The energy in the LP residual signal is mostly concentrated only in a part of the residual signal around the instant of glottal closure [41, 42]. Since the nature of the excitation in the vowel regions is different, a better finer feature for the vowel–nonvowel classification can be derived by processing only 2 ms portion of the residual signal around the significant excitation. This results in another set of features different from the one described above. Further, the performance of the statistical classifiers for the vowel–nonvowel segmentation may be enhanced by combining these features with those that better model the vocal tract system. Motivated by these factors, a front-end speech parameterization approach for the detection of vowel regions in speech is proposed in this work. The sequence of steps involved in the extraction of the proposed features are described in the following subsections.

2.1 Detecting Instants of Significant Excitation Using Zero Frequency Filtering Method

The front-end features proposed in this work rely on an accurate detection of the glottal closure instants (GCIs). Among the several existing GCI detection techniques, the zero frequency filtering (ZFF) method is noted to detect the GCIs with a much better accuracy. At the same time, the ZFF-based method is found to be robust to the variations in the environmental conditions [19, 35]. The ZFF-based GCI detection technique [19] exploits the fact that the excitation source exhibits impulse-like discontinuities. The discontinuities due to impulsive excitation are spread uniformly across all the frequencies, including the zero frequency. The ZFF method then filters the speech signal to preserve the energy around the zero frequency which is mainly due to the impulse-like excitation. The positive zero crossings of the ZFF signal give the location of GCIs. Using the ZFF approach, the location of the GCIs can be obtained from the speech signal $s(n)$ by the following sequence of steps [19]:

- Determine the first difference $x(n)$ of the speech signal $s(n)$ where

$$x(n) = s(n) - s(n - 1). \quad (1)$$

- Compute the output of a cascade of two ideal digital resonators at 0 Hz

$$y(n) = - \sum_{k=1}^4 a_k y(n - k) + x(n) \quad (2)$$

where $a_1 = 4$, $a_2 = -6$, $a_3 = 4$, $a_4 = -1$.

- Remove the trend, i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n) \quad (3)$$

where $\bar{y}(n) = \frac{1}{(2N+1)} \sum_{n=-N}^N y(n)$ and $2N + 1$ corresponds to the average pitch period computed over a longer segment of speech.

- The trend removed signal $\hat{y}(n)$ is called the ZFF signal.
- The positive zero crossings of the ZFF signal will give the location of GCIs.

2.2 Extraction of Excitation Source Information of Vowels from the LP Residual Signal

In the LP analysis of the speech signal, each sample is predicted as a linear combination of the past m samples as follows [1, 17, 41]:

$$\hat{s}(n) = - \sum_{k=1}^m b_k s(n-k) \quad (4)$$

where m is the order of prediction and the set of linear prediction coefficients (LPCs) are denoted by $\{b_k\}_{k=1}^m$. The LPCs are computed by minimizing the mean square error between the original and the predicted speech samples. The error between the predicted samples $\hat{s}(n)$ and actual speech samples $s(n)$ is referred to as the LP residual signal. The LP residual signal is obtained by passing the speech signal through a time-varying inverse filter constructed using the LPCs. This inverse filter is given by the following equation:

$$B(z) = 1 + \sum_{k=1}^m b_k z^{-k}. \quad (5)$$

The LP residual signal mostly contains the excitation source information [1, 41]. The accuracy in representing the excitation source by the LP residual signal, in turn, depends on the order of prediction. Most of the studies presented in the literature show that a 10th order prediction is sufficient for characterizing the excitation source information in the vowel regions for a speech signal that is sampled at 8 kHz rate [28, 35].

In Fig. 2, a segment of speech signal taken from a vowel region (Fig. 2a), its LP residual derived using 10th-order LP analysis (Fig. 2b), the corresponding ZFF signal (Fig. 2c) and the location of the GCIs derived from the ZFF signal (Fig. 2d) are shown. From Fig. 2b, it is evident that the LP residual energy is mostly concentrated in a small region around the GCIs. The GCI locations given in Fig. 2d match accurately with the high energy portions of the LP residual signal shown in Fig. 2b. Therefore, the temporal characteristics of the excitation source in the vowel regions can be modeled by considering the 2 ms regions around the GCIs. It is to note that, in the case of LP residual, the maximum energy is concentrated in the 2 ms region around the GCI [41]. These regions can be accurately detected by anchoring the location of the GCIs detected by the ZFF method.

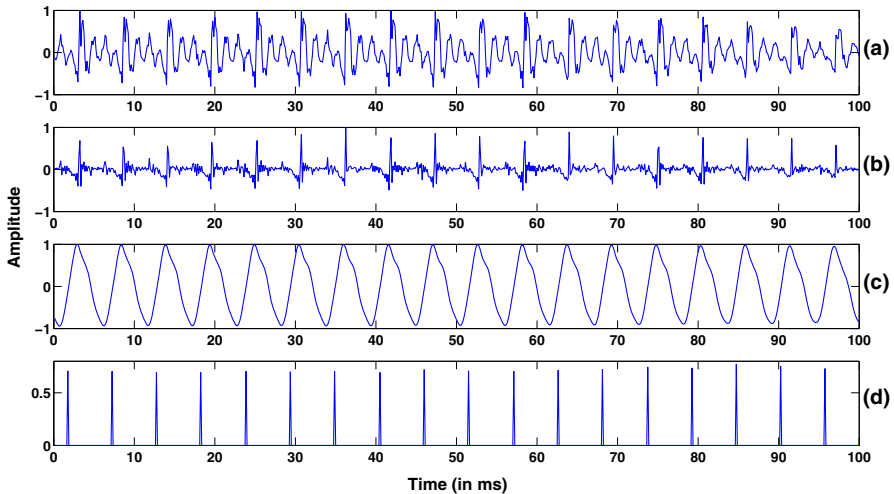


Fig. 2 Display of the LP residual and glottal closure instant of a segment of vowel speech. **a** The segment of the vowel sound unit considered, **b** the LP residual signal using 10th-order LP analysis, **c** the corresponding ZFF signal and **d** the GCIs derived from the ZFF signal

2.3 Extracting the Temporal and the Spectral Characteristics of the Excitation Source

Since the speech signal varies slowly with respect to time, short-time analysis considering a frame duration of 20–30 ms with an overlap of 50 % is generally employed during the front-end parameterization step. In this work, a frame size of 20 ms with a frame shift of 10 ms is considered during the computation of MFCC features. In order to have an equal number of frames for the proposed approach as well, the same frame size and frame shift are considered. The speech signal is processed through the following sequence of steps for estimating the temporal and the spectral characteristics of the excitation source information in the vowel regions.

- Step I : The instants of significant excitation (GCIs) are detected by using ZFF method.
- Step II : The speech signal is processed in frames that are 20 ms in duration with a frame shift of 10 ms. For each 20 ms block, 10th-order LP analysis is performed to estimate the LPCs. A time-varying inverse filter is constructed using these LPCs. The speech signal is passed through the inverse filter to extract the LP residual signal.
- Step III : Next, the LP residual signal is filtered through a bank of 8 nonoverlapping filters each having a bandwidth of 500 Hz. This splits the LP residual into 8 sub-bands. Feature vectors are then extracted by computing the spectral energy in each sub-band considering a frame size of 20 ms with 50 % overlap. This results in an 8-dimensional feature vector per frame capturing the spectral variations in the LP residual signal.

Step IV : Similarly, for each block of 20 ms, the GCI locations within that frame are identified using ZFF.

- If one or more GCIs are found within an analysis frame, the regions that are 2 ms to the right of each GCI for all the sub-bands of the LP residual are identified by anchoring these GCIs. For this, only those GCIs are considered for which such 2 ms regions exist to the right of the GCI within the analysis frame under consideration.
- The temporal energies are then computed within those regions. To do so, Hamming-windowed regions that are 2 ms in duration are considered. Finally, the average of the energy for all the GCIs within that frame is computed.
- If no GCIs are found within the analysis frame under consideration, the temporal energies for all the sub-bands are computed by considering the central portions of the analysis frame with a duration of 2 ms.
- This results in another 8-dimensional feature vector per frame capturing the temporal variations in the LP residual signal.

Step V : To reduce the dynamic variations, the logarithm of the spectral and temporal energies is taken.

Step VI : The derived features are then concatenated to obtain a 16-dimensional feature vector.

Step VII : The first- and the second-order temporal derivatives (the delta and delta–delta coefficients) are then computed for the current feature vector by using the two preceding and the two succeeding frames. The delta as well as the delta–delta coefficients will also be 16-dimensional. These coefficients are appended to the base features making the total feature dimension equal to 48 (16-dimensional base + 16-dimensional delta + 16-dimensional delta–delta).

Since the analysis frames considered for the computation of the spectral and the temporal energies are quite different in duration, the derived feature vectors turn out to be different.

3 Vowel–Nonvowel Classification

As mentioned earlier, different acoustic modeling approaches (GMM–HMM, SGMM–HMM and DNN–HMM) have been explored in this work for the task of correctly identifying vowels in speech signal. The Kaldi speech recognition toolkit [13,23] was used to develop the vowel–nonvowel detection systems employing the different acoustic modeling approaches. In the following, the details of the speech corpus used in the presented study are discussed. This is followed by the description of the vowel–nonvowel classification systems employed in this work for the detection of vowels in a speech sequence.

3.1 Speech Corpus Employed

The system development and evaluation were done on the TIMIT corpus [9]. The speech corpus was split into orthogonal sets following the standard Kaldi recipe.

In order to train the acoustic model parameters, the speech data from 462 speakers comprising of 3696 utterances were used. The test set comprised of 192 utterances from 24 speakers. All the experiments reported in this paper were performed on 8 kHz re-sampled data to simulate telephone-based speech interface. The training transcription was modified to represent the possible vowels in the database as a single class. The nonvowels were grouped together to represent the second class. The silence, the short pause and the other nonspeech units (fillers) were grouped together to represent the third class (silence). The objective of the presented study is to segment the given speech signal into vowel, nonvowel and silence regions. Therefore, a three-class classifier (*vowel, nonvowel and silence*) was trained on the TIMIT data using various statistical modeling approaches. The spectral parameterization approaches employed for feature extraction are discussed later.

3.2 Description of the Different Acoustic Modeling Techniques Used for Vowel–Nonvowel Classification

In the work presented in this paper, an attempt is made to detect the accurate vowel regions by exploring different acoustic modeling techniques. As mentioned earlier, the three statistical approaches for learning the acoustic model parameters reported in the literature are the GMM, the SGMM [22] and the DNN [3]. In all these approaches, the temporal variation is captured using the hidden Markov model (HMM). Acoustic modeling techniques based on subspace Gaussian mixture model (SGMM) and deep neural network (DNN) are very recent developments in the domain of speech recognition research. Both these techniques are reported to be superior to the GMM-based approach. Motivated by the success of SGMM- and DNN-based approaches, their effectiveness for the detection of the sequence of vowels and nonvowels in a given speech signal is explored in this paper. Further, the effectiveness of various acoustic modeling techniques in the determination of vowels, VOPs and VEPs in the given speech data is also compared. In the following subsections, a brief description of the SGMM- and the DNN-based acoustic modeling is presented. These discussions closely follow the works reported in [3, 11, 21, 22, 32].

3.2.1 Subspace Gaussian Mixture Modeling

In the case of acoustic modeling based on GMM–HMM, each HMM state is modeled using a dedicated mixture of multivariate Gaussians. Hence, there is no sharing of parameters between the states. On the other hand, the SGMM-based acoustic modeling approach facilitates parameter sharing among the states. As a result, the acoustic model parameters can be robustly learned even with a smaller amount of training data in the case of SGMM. The SGMM approach has some similarities with the joint factor analysis used in speaker verification [14] as well as the model space adaptation techniques like the eigenvoices [16] and the cluster adaptive training [7].

The acoustic model parameters in the case of SGMM represent a globally shared subspace. A set of low-dimensional state-specific vectors, referred to as *state projection vectors* $\{\mathbf{v}_j\}$, are trained from data to capture principal directions of acoustic variability.

The model means and the mixture weights are then derived using those state projection vectors. It is to note that the covariances in the SGMM are shared among all the HMM states and are represented by full covariance matrices unlike the diagonal matrices used in the case of GMM. Additional flexibility in the SGMM parameterization is provided by introducing the notion of a substate within a state. To train the model parameters, a single-state GMM called the universal background model (UBM) is learned on the data from all speech classes pooled together. The subspace parameters \mathbf{M} , \mathbf{v} and $\mathbf{\Sigma}$ are initialized in such a way that the means and covariances in each state for the first iteration are the same as that of the UBM. The usual expectation–maximization (EM) algorithm employing a maximum likelihood (ML) criterion is used to optimize the parameters of the SGMM in an iterative manner.

3.2.2 Deep Neural Network

A major drawback of GMM-based acoustic modeling approach, as suggested in [11], is its inefficiency in modeling the data that lie on or near a nonlinear manifold in the data space. The artificial neural network (ANN), on the other hand, is reported to have the potential to learn these models of data that lie on or near the nonlinear manifold. Consequently, the deep neural networks containing many layers of nonlinear hidden units and a very large output layer are now being used for modeling the acoustic variations in speech recognition systems [3]. In the DNN–HMM systems, the posterior probabilities of the senones (or the context-dependent tied state) are modeled using the DNN. These posterior probabilities are then used in a HMM-based classifier. The speech recognition systems based on the DNN–HMM modeling paradigm are reported to outperform the ones based on GMM–HMM.

Deep neural networks are created by stacking layers of restricted Boltzmann machine (RBM) which is a undirected generative model. The joint probability of a vector of observable variables (\mathbf{v}) and a vector of latent/hidden variables (\mathbf{h}) in the case of an undirected model is given by a single set of parameters (\mathbf{W}) via an energy function E . After training an RBM on the data, the output of the hidden units can be used as the input data for training another RBM. For each data vector \mathbf{v} , the vector of hidden unit activation probabilities \mathbf{h} is computed. These hidden activation probabilities are then used as the training data for a new RBM. Thus, each set of RBM weights can be used to extract features from the output of the previous layer. The initial values for all the weights of the hidden layers of the neural nets can thus be generated using RBM training (the number of hidden layers being equal to the number of RBMs trained). This is called pre-training of a deep belief network (DBN). A randomly initialized softmax output layer is then added, and all the weights in the network are discriminatively fine-tuned using backpropagation to create a DNN. In the case of an automatic speech recognition system, the softmax output layer has as many nodes as the number of classes, i.e., the number of senones. For speech recognition, the output unit j converts the total input x_j into a class probability p_j by using the softmax nonlinearity given by

$$p(x_j) = \frac{\exp(x_j)}{\sum_k \exp(x_k)} \quad (6)$$

where k is an index over all the classes. The senone likelihood $p(x_j)$ is used with the HMM modeling the sequential property of the speech. The DNNs can be trained by backpropagating derivatives of the cost function, i.e., the cross-entropy between the target probabilities and the output of the softmax function.

3.3 Vowel–Nonvowel Classification Using the MFCC Feature Vector

In the first level, the vowel detection systems using the MFCCs as the feature vector were developed. For the MFCC feature vector extraction, the speech data were analyzed using a Hamming window of length 20 ms with frame rate of 100 Hz and a pre-emphasis factor of 0.97. The 13-dimensional base MFCC features ($C_0 - C_{12}$) were computed employing 23-channel mel-filterbank. Since the speech data used in this work are sampled at 8 kHz rate, the number of filters is selected to be 23. In the case of higher sampling rates, the number of filters should be modified accordingly. The first- and the second-order temporal derivatives (the delta and delta–delta coefficients) were then appended, making the feature dimension equal to 39. The acoustic model parameters were then learned on those 39-dimensional MFCC feature vectors. The details of the employed statistical modeling techniques are given in the following.

3.3.1 Development of GMM–HMM, SGMM–HMM and DNN–HMM systems

The context-independent-monophone- and the context-dependent-triphone-based acoustic modeling was performed using the 39-dimensional features (GMM–HMM (monophone) and GMM–HMM (triphone)). For the context-dependent-triphone system, crossword modeling with decision-tree-based state tying was employed. For further improving the recognition performances, time splicing of the base MFCCs, considering a context size of 9 (4 frames to the left and to the right of the central frame), was done making the total feature dimension equal to 117. The dimensionality of the derived time-spliced features was then reduced to 40 using LDA and MLLT. These features were then employed in developing another context-dependent-triphone-based recognition system (GMM–HMM (triphone, LDA–MLLT)). Speaker normalization using fMLLR was also explored to further improve the performance. The fMLLR transformations were generated for the training and test data using speaker adaptive training (SAT) [2] approach as suggested in [31]. A revised recognition system was developed on the fMLLR-normalized features as well (GMM–HMM (triphone, fMLLR–SAT)).

In the case of SGMM-based vowel–nonvowel classification system, the number of Gaussians used for training the universal background model was selected as 400. The number of leaves and Gaussians in the SGMM was chosen to be 9000 and 7000, respectively. The LDA–MLLT features followed by fMLLR-based normalization were employed in the training of the model parameters. Discriminative training using boosted maximum mutual information (MMI) was also explored in combination with the SGMM.

For learning the DNN parameters, the 40-dimensional time-spliced features with fMLLR-based normalization were further spliced over 4 frames to the left and right

Table 1 Classification error rates for the vowel and nonvowel classifiers developed using different acoustic modeling approaches explored in this work

Explored acoustic modeling technique	Error (in %)		
	MFCCs	Prop.	Concat.
GMM–HMM (monophone)	20.14	22.83	20.32
GMM–HMM (triphone)	15.40	19.49	16.19
GMM–HMM (triphone, LDA–MLLT)	14.87	18.72	15.63
GMM–HMM (triphone, fMLLR–SAT)	14.55	18.81	15.84
SGMM–HMM	13.82	17.53	15.04
SGMM–HMM–MMI	13.56	17.96	14.57
DNN–HMM	15.61	18.79	16.42

Performances are given with respect to the systems developed using the conventional MFCC feature vectors, the proposed excitation feature vectors and the feature vectors obtained by their frame-level concatenation

of the central frame. The number of hidden layers was selected to be equal to 2 as the amount of data were quite less. Increasing the number of hidden layers was not found to be helpful. The number of nodes in the hidden layer was selected to be 300. The \tan/h function was used to model the nonlinearities in the hidden layers with cross-entropy being the optimization criterion. The initial learning rate was set to 0.015 which was then reduced to 0.002 in 15 epochs. After reducing the learning rate to 0.002, extra 5 epochs of training were employed [43]. The minibatch size for neural net training was fixed at 128.¹

The recognition performances for the three-class classifier developed on the MFCC features employing the discussed modeling techniques are given in Table 1. The error rates given in Table 1 are computed in the same way as the word error rates with the possible words being vowel, nonvowel and silence. The effect using different acoustic modeling techniques is quite evident from the enlisted error rates.

3.4 Vowel–Nonvowel Classification Using the Excitation Source Feature

Having successfully developed systems using the MFCCs feature vectors, the detection systems using the proposed excitation source features were developed next. These features are derived from the speech signal by following the sequence of steps as described in Sect. 2.3. The GMM–HMM, SGMM–HMM and DNN–HMM systems on the 16-dimensional base features extracted through the proposed method are developed following the same procedure as described in Sect. 3.3. The recognition performances for the various systems on the proposed features are also enlisted in Table 1. In addition to that, the effect of concatenating the MFCC and the proposed feature vectors is also explored in this work. To do so, for each frame, the 13-dimensional base MFCC features and the 16-dimensional proposed excitation features were concatenated to derive

¹ It is to note that, for all the discussed configuration parameters, the chosen values are taken from the Kaldi recipe.

a 29-dimensional feature vector. The statistical models were then trained on these 29-dimensional vector treating them as the base features. The recognition performances for the various systems trained on the concatenated features are also given in Table 1.

From the presented error rates, it can be concluded that the proposed excitation features can be used for developing the classification systems even though such systems will be inferior to those developed using the MFCCs. Furthermore, the frame-level concatenation of the two features is also helpful to a certain extent. Motivated by these results, the classification systems developed using the two kinds of features were employed for the detection of vowel regions in a given speech signal. It will be noted from the studies presented in the following section that two kinds of features result in different and, at times, nonoverlapping vowel regions for the same speech sequence. Consequently, the evidences obtained from the two features can be combined to get a better estimate of the vowel regions. In this regard, a novel approach of combining the evidences that enhances the performances is also proposed.

4 Detection of Vowel Regions in Speech Signal

The frame-level alignments required to detect the vowel regions, the VOPs and the VEPs were generated by the forced alignment of the test data with respect to the trained acoustic models under the constraints of the first-pass hypothesis. This first-pass hypothesis was obtained by decoding the test data on the trained acoustic models. The use of first-pass hypothesis represents the real testing scenario. For the purpose of evaluating the effectiveness of the proposed techniques, the true frame-level alignments for the vowel regions are determined using the hand-labeled transcription available with the TIMIT database. The metrics employed to compute the accuracy with which the vowel regions were detected is described in the next subsection. This is followed by a detailed analysis of the effectiveness of the two kinds of feature vectors in the estimation of the VOPs and the VEPs. Further, we also propose a novel scheme for combining the evidences obtained by using the two kinds of feature vectors in order to get an enhanced evidence of the VOPs and the VEPs.

4.1 Metrics for Evaluating the Accuracy of the Detected Vowel Regions

The performances of the developed systems, in order to determine the vowel regions and their VOPs and VEPs, are measured using the following metrics:

- *Identification rate (IR)* The percentage of reference vowel regions that exactly match with the detected vowel regions.
- *Spurious rate (SR)* The percentage of detected vowel regions that lie outside the reference vowel regions. The spurious rate is further broken into following three categories:
 - i. *SR for semivowels* The percentage of reference semivowel regions that exactly match with the detected vowel regions.
 - ii. *SR for nasals* The percentage of reference regions for the nasal unit that exactly match with the detected vowel regions.

- iii. *SR for others* The percentage of nonvowel regions (excluding semivowels and nasals) that exactly match with the detected vowel regions.

4.2 Detection of Vowel Regions Using the MFCC and Excitation Source Features

As already mentioned, forced alignment of the test data with respect to the trained acoustic models was used to generate the frame-level state alignments. These state alignments were then employed to detect the vowel regions in the speech. The IR and the SR for the two feature vectors with respect to the acoustic models developed using the explored modeling approaches are enlisted in Tables 2 and 3. On analysis, the MFCCs and the proposed feature vectors are observed to result in somewhat different state alignments. Consequently, the possibility of combining the evidences in order to obtain an enhanced accuracy in the detection of vowel regions was explored. One of the ways to combine the evidences is to consider the union or the intersection of the vowel regions detected by these two features. The performance in terms of the IR and the SR for these two cases is also given in Tables 2 and 3. In the case of union, the IR rate improves significantly in comparison with the individual features at the cost of an increase in the SR. In this result, it is interesting to note that the improvement in IR is significantly more when compared to the overall increase in SR. In the case of intersection, on the other hand, the SR is reduced at the cost of a reduction in IR.

It is evident from these results that the vowel regions detected by these features are much different, and even in some cases, there may be no overlap between them. By a suitable combination of these features, the IR can be improved with a simultaneous reduction in the SR. Motivated by these observations, the features were concatenated at the frame-level in the feature domain. The concatenated feature vectors were then used to learn the statistical models following the procedure described in Sect. 3.3. Once the acoustic models were trained, forced alignment was used to detect the vowel regions. The performances obtained by the use of the concatenated feature vectors with respect to the different acoustic modeling methods in terms of the IR and the SR are also given in Tables 2 and 3. The feature concatenation is observed to be superior to the excitation source features but, at the same time, somewhat poorer when compared to the MFCCs. This may be due to the significant diversity between the two kinds of feature vectors.

In order to quantify the differences in the two kinds of acoustic features, canonical correlation analysis (CCA) was performed. In the case of CCA, the sample canonical coefficients for the two kinds of feature matrices are computed. The matrices consist of same number of observations (rows), but number of columns (dimensions) may be different. To derive the matrices, the feature vectors corresponding to the entire TIMIT database were collected together. Separate matrices were created for the two kinds of acoustic features, i.e., the MFCCs and the proposed features. The canonical correlation coefficients between the MFCCs and the proposed features are shown in Fig. 3. Except for the first coefficient, the canonical correlation turns out to be quite low. This means that the proposed feature vectors capture information that is not represented through

Table 2 Performances of the vowel detection systems developed using the different GMM–HMM-based acoustic modeling approaches explored in this work

Acoustic modeling technique		IR in %	SR in %		
			Semivowel	Nasal	Other
GMM–HMM (monophone)	MFCC	88.53	11.52	3.93	4.14
	Excitation	82.15	10.32	5.11	7.10
	Union	93.32	11.77	5.53	8.31
	Intersection	77.36	7.06	2.24	1.34
	Feat. Concat.	84.29	10.48	6.05	4.84
	Prop. Comb.	89.82	10.97	4.10	3.53
GMM–HMM (triphone)	MFCC	85.67	10.47	2.94	3.77
	Excitation	81.03	9.87	4.35	5.98
	Union	92.16	11.17	4.61	7.27
	Intersection	74.55	6.25	1.62	1.04
	Feat. Concat.	83.13	9.80	5.63	4.22
	Prop. Comb.	86.95	9.99	3.18	3.14
GMM–HMM (LDA–MLLT)	MFCC	84.63	10.55	2.30	3.54
	Excitation	82.05	10.26	4.37	3.85
	Union	91.85	11.57	4.44	5.33
	Intersection	74.82	6.42	1.32	1.06
	Feat. Concat.	83.34	9.46	4.91	3.65
	Prop. Comb.	86.31	10.00	2.66	2.87
GMM–HMM (fMLLR–SAT)	MFCC	84.32	10.58	1.97	3.54
	Excitation	82.55	10.49	4.04	3.39
	Union	91.61	11.71	4.06	5.11
	Intersection	75.26	6.58	1.21	0.94
	Feat. Concat.	82.87	9.55	4.68	3.87
	Prop. Comb.	86.28	10.12	2.24	2.53

Performances are given for the conventional MFCC feature vectors, the proposed excitation feature vectors, the feature vectors obtained by their frame-level concatenation and the proposed method of combining the evidences. Performances are also given for the union and the intersection of the vowel evidences obtained by using the MFCC feature vectors and the proposed excitation feature vectors. The frame-level alignments employed in evaluating the performances are obtained by the forced alignment of test data with respect to the considered acoustic models under the constraints of the first-pass transcription

the MFFCs. Since the first coefficient represents the signal energy in the case of both the kinds of feature vectors, the canonical correlation turns out to be high.

4.3 Combining the Evidences

For all the acoustic modeling methods considered in this work, the MFCC features are noted to perform better than the proposed excitation features. Furthermore, even the frame-level concatenation of the features does not result in great improvements in

Table 3 Performances of the vowel detection systems developed using the SGMM–HMM- and the DNN–HMM-based acoustic modeling approaches explored in this work

Acoustic modeling technique		IR in %	SR in %		
			Semivowel	Nasal	Other
SGMM–HMM	MFCC	83.18	10.32	1.52	3.93
	Excitation	81.15	10.59	3.55	2.95
	Union	90.94	11.70	3.58	5.04
	Intersection	73.40	6.33	0.88	0.91
	Feat. Concat.	82.63	9.38	3.97	4.09
	Prop. Comb.	85.06	9.80	1.83	3.15
SGMM–HMM–MMI	MFCC	83.10	10.32	1.52	3.70
	Excitation	81.07	10.59	3.47	3.00
	Union	91.00	11.72	3.94	4.97
	Intersection	73.28	6.29	0.85	0.87
	Feat. Concat.	82.37	9.33	3.87	3.93
	Prop. Comb.	85.12	9.87	1.78	2.74
DNN–HMM	MFCC	83.76	10.93	1.80	4.32
	Excitation	83.06	11.15	4.14	4.25
	Union	91.28	12.30	4.16	6.02
	Intersection	75.54	6.85	1.01	1.41
	Feat. Concat.	82.92	10.08	4.38	4.27
	Prop. Comb.	85.90	10.31	2.18	3.94

Performances are given for the conventional MFCC feature vectors, the proposed excitation feature vectors, the feature vectors obtained by their frame-level concatenation and the proposed method of combining the evidences. Performances are also given for the union and the intersection of the vowel evidences obtained by using the MFCC feature vectors and the proposed excitation feature vectors. The frame-level alignments employed in evaluating the performances are obtained by the forced alignment of test data with respect to the considered acoustic models under the constraints of the first-pass transcription

the IR and the SR. The IR and the SR for the vowel detection may be improved by combining the vowel evidences obtained from both the kinds of features. Since the IR and SR for the MFCCs are better than the excitation features, a higher weighting must be given to the evidence obtained using the MFCC features. To achieve this, a method is proposed for combining the evidences. In the proposed method, the detected evidences are first classified into two categories, i.e., the overlapping and the nonoverlapping categories. This is followed by modifying the starting and ending points that is done as follows:

- (a) If the vowel evidences obtained by using the MFCCs and the excitation feature exhibit a minimum overlap of 70 %, then those are considered as overlapping evidences. On the other hand, in the case of nonoverlapping evidences, the overlap is less than 70 %. In the case of overlapping evidences, the starting and ending points of the combined evidence are selected as follows:

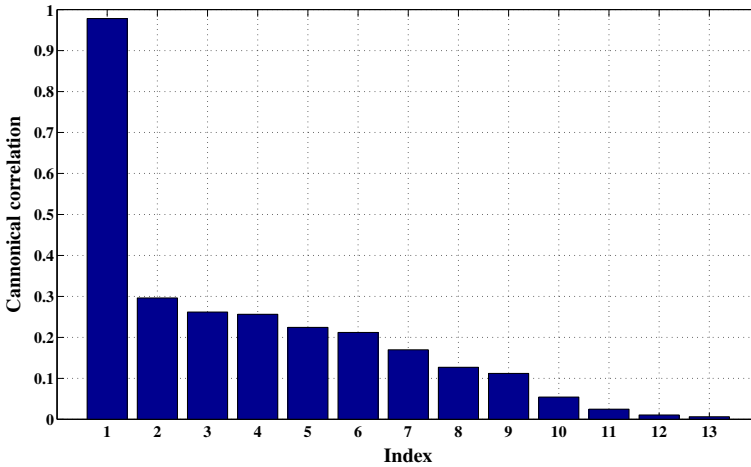


Fig. 3 Bar graph showing the canonical correlation coefficients between the two kinds of acoustic features explored in this work

- If the starting point of the vowel evidence detected by the proposed feature falls within two analysis frames (240 samples for 20 ms frame size with 10 ms frame shift) of the starting point of the vowel evidence detected by the MFCC feature, then the starting point of final evidence is considered as the mean of these two locations. Otherwise, the starting point of the vowel evidence detected by the MFCC is considered as the starting point for that particular case.
 - Similar steps are also followed for deciding the end points of the overlapped vowel evidences.
- (b) For both kinds of features, the nonoverlapping evidences that are a minimum of 100 ms in duration are identified and preserved in the final evidences without any modification. Those nonoverlapping evidences that are less than 100 ms in duration are treated as spurious detections and are eliminated.

The evidences for the vowel regions detected by using the MFCCs, the excitation features, their frame-level concatenation and the proposed combination with respect to the acoustic models trained via the GMM–HMM (fMLLR–SAT) and the SGMM–HMM–MMI methods are given in Fig. 4. By comparing the detected vowel evidences with the references, it can be observed that for both the modeling approaches, the proposed method of combining the evidences helps in detecting the vowel regions with far more accuracy when compared to that detected using each of the individual features as well as their frame-level concatenation. As discussed earlier, in the case of the excitation features, the VOPs are detected only after 1–3 glottal cycles. On the other hand, in the case of the MFCC features, the evidences are detected before few glottal cycles. For these two features types, not only the obtained evidences are different, but also the confusion between the voiced region and the vowel is also different. The observed differences may probably be attributed to that fact that the proposed

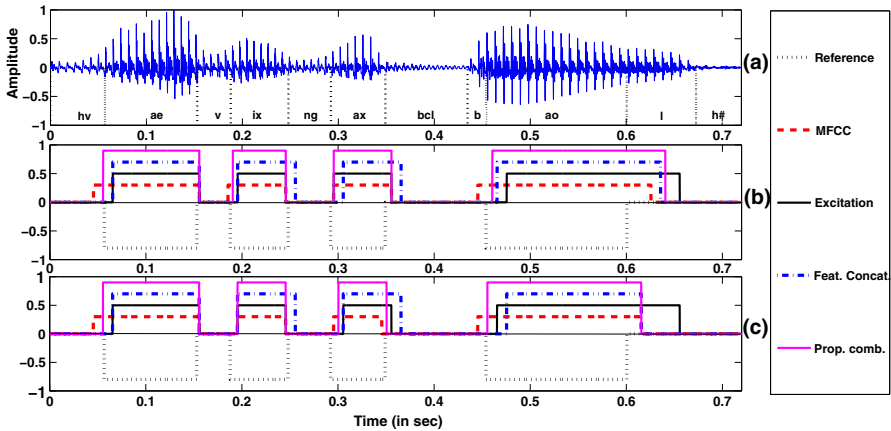


Fig. 4 Illustrations depicting the effectiveness of the proposed method for combining the evidences. **a** A segment of speech, “having a ball,” with reference markings for the sound units is shown. **b** The detected vowel evidences obtained with respect to the GMM–HMM (fMLLR–SAT) models using the conventional MFCC feature vectors, the proposed excitation feature vectors, the feature vectors obtained by their frame-level concatenation and proposed combination at evidence level are shown. **c** The same set of detected regions obtained by using the SGMM–HMM–MMI models are shown

excitation features and the MFCCs model the few frames at the vowel transitions quite differently. Consequently, as it is evident from Fig. 4, the VOPs and VEPs obtained using the proposed combination of evidences are more accurate when compared to those detected by using each of the individual features. The performances in terms of IR for the proposed combination scheme, obtained with respect to the explored acoustic modeling techniques, are given in Tables 2 and 3, respectively. The results show that for each of the acoustic modeling techniques, the proposed combination scheme provides better IR with an overall reduction in SR.

4.4 Detection of Vowel Onset and End Points

In this section, the effectiveness of the explored acoustic modeling methods for the detection of the vowel onset points (VOPs) and the vowel end points (VEPs) is presented. The signal processing methods proposed in the literature suggest that an accurate detection of VOPs and VEPs in the cases of semivowel–vowel transitions and diphthongs is quite difficult [25, 28]. Also, the performances for most of the explicit signal processing techniques are relatively poor in the case of the VEPs [25, 40]. The signal characteristics at the VEP are significantly different from that at the VOP [25, 40]. Unlike the VOP, the signal strength decreases slowly at the VEP. Due to this, detecting the VEPs is more challenging than that of the VOPs [25]. This is mainly due to the explicit use of the transition in the excitation strength as a feature vector in the earlier reported works. The proposed approach, on the other hand, does not explicitly depend on the transition in the excitation strength. In the presented work, since the statistical properties of the excitation source and the vocal tract system are used for detection, the proposed approach will detect these events with an enhanced accuracy. This also

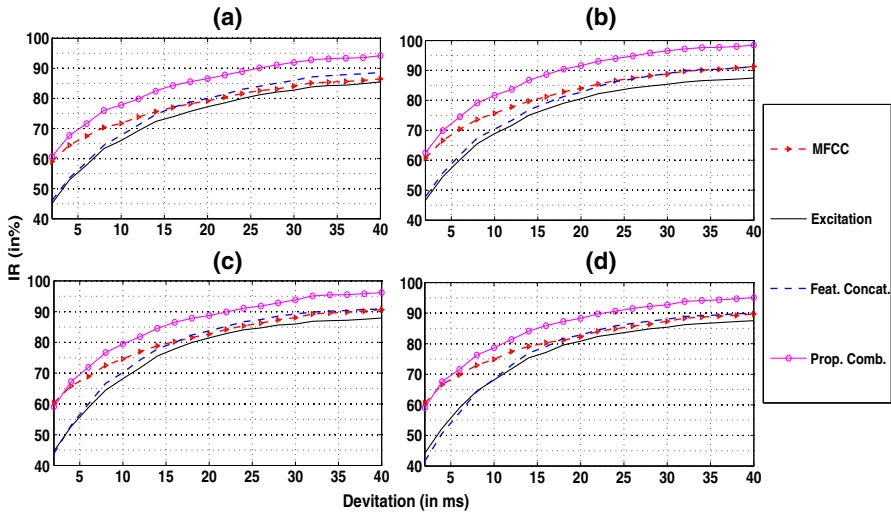


Fig. 5 The IR profiles for the VOPs with respect to the different GMM-HMM systems. The predefined deviation is varied from 2 to 40 ms in steps of 2 ms. The IR is given for the MFCCs, the excitation features, feature concatenation and the proposed combination of evidences. **a** IR for GMM-HMM (monophone) system, **b** IR for GMM-HMM (triphone) system, **c** IR for GMM-HMM (LDA-MLLT) system and **d** IR for GMM-HMM (fMLLR-SAT) system

helps in analyzing the missing and the spurious cases for the vowel evidences detected by different acoustic modeling methods.

4.4.1 Metrics for Performance Evaluation

The starting and the ending points of the detected vowel regions are marked as the VOPs and the VEPs, respectively. Using the manual markings given in the database as the reference, the performances of the detected VOPs and VEPs are measured using the following metrics:

- *Identification rate (IR)* The percentage of the reference VOPs/ VEPs that match with the detected VOPs/ VEPs within the pre-defined deviation (in ms).
- *Spurious rate (SR)* The percentage of detected VOPs/ VEPs, which are detected outside the vowel regions.

4.4.2 Performances for the VOP Detection

The identification rate (IR) for the VOP detection with respect to the explored GMM-HMM-based acoustic modeling techniques is summarized in Fig. 5. The IR for the SGMM-HMM and the DNN-HMM systems is given in Fig. 6. The pre-defined deviation values considered in this study are varied from 2 ms to 40 ms in steps of 2 ms. Among the different acoustic modeling approaches explored in this work, the SGMM-HMM-MMI system provides the best IR. It is evident from these results that, for all the deviations considered, the proposed method of combining the evidences provides

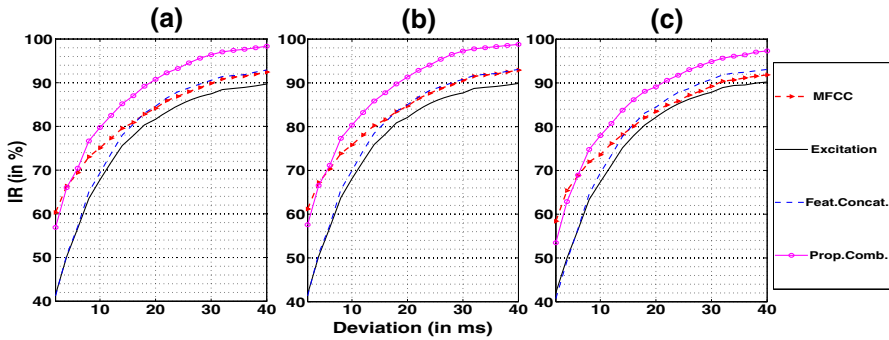


Fig. 6 The IR profiles for the VOPs with respect to the SGMM–HMM and DNN–HMM systems. The other details remain the same as that for Fig. 5. **a** IR for SGMM–HMM system, **b** IR for SGMM–HMM–MMI system and **c** IR for DNN–HMM system

a significant improvement in IR for all the explored acoustic modeling techniques. For an instance, the IR with respect to the SGMM–HMM–MMI system for the case when the deviation is 10 ms is 75.82, 68.09, 69.92 and 80.35 % for the conventional MFCC feature vectors, the proposed excitation feature vectors, the feature vectors obtained by their frame-level concatenation and proposed combination, respectively. Similarly, the IR for the 40 ms deviation case is 92.90, 89.87, 93.07 and 98.80 %, respectively. The spurious rate (SR) of the VOP detection for all the explored acoustic models is given in Table 4. It is to note that the SGMM–HMM–MMI system provides the minimum SR for the MFCC feature vectors. The proposed combination further reduces the SR along with a significant improvement in IR. Even though the feature vectors obtained by the frame-level concatenation provide slightly lesser SR, the IR in this case is significantly less when compared to that obtained by the proposed combination.

4.4.3 Performances for the VEP Detection

The identification rate (IR) of the VEP detection with respect to the explored acoustic modeling techniques is summarized in Figs. 7 and 8. Similar to the case of VOP detection, the SGMM–HMM–MMI system provides the best IR compared to all other acoustic modeling approaches considered in this work. The proposed method of combining the evidences provides significant improvement in IR for all the acoustic modeling methods. Compared to the VOP, the proposed combination provides relatively more improvements in IR for smaller deviations in the case of the VEP. For an instance, the IR with respect to the SGMM–HMM–MMI system for the case when the deviation is 10 ms is 73.00 and 82.22 % for the MFCCs and the proposed combination, respectively. In the case of the MFCCs and the proposed combination, the IR for the VEPs is slightly poorer as compared to that of the VOPs. For the proposed excitation features and the feature vectors obtained by frame-level concatenation, the IR for the VEP detection are significantly less when compared to the IR of VOP. The SR of the VEP detection methods is also given in Table 4. For all the acoustic modeling techniques, the SR for the proposed excitation feature vectors is relatively

Table 4 Spurious rate (SR) of VOP and VEP detection using different acoustic modeling approaches explored in this work

Acoustic modeling technique		SR in %			
		MFCC	Excitation	Feat. Concat.	Prop. Comb.
GMM–HMM (monophone)	VOP	7.17	9.19	5.48	6.06
	VEP	8.12	14.88	12.57	8.34
GMM–HMM (triphone)	VOP	6.86	9.12	5.46	6.15
	VEP	7.64	13.68	11.57	8.08
GMM–HMM (LDA–MLLT)	VOP	6.98	6.37	4.07	5.95
	VEP	6.69	11.39	9.86	7.03
GMM–HMM (fMLLR–SAT)	VOP	7.23	5.85	4.39	6.43
	VEP	6.72	11.03	10.40	6.92
SGMM–HMM	VOP	6.78	5.12	4.43	5.66
	VEP	6.04	9.21	9.10	6.20
SGMM–HMM–MMI	VOP	6.54	5.65	4.42	5.48
	VEP	6.13	9.77	8.85	6.40
DNN–HMM	VOP	7.03	5.17	4.35	6.08
	VEP	6.30	9.48	8.67	6.58

Performances are given with respect to the systems developed using the conventional MFCC feature vectors, the proposed excitation feature vectors, the feature vectors obtained by their frame-level concatenation and the proposed combination of evidences. Performances are evaluated using the forced alignment under the constraints of the first-pass transcription with respect to the models trained on 8 kHz re-sampled data

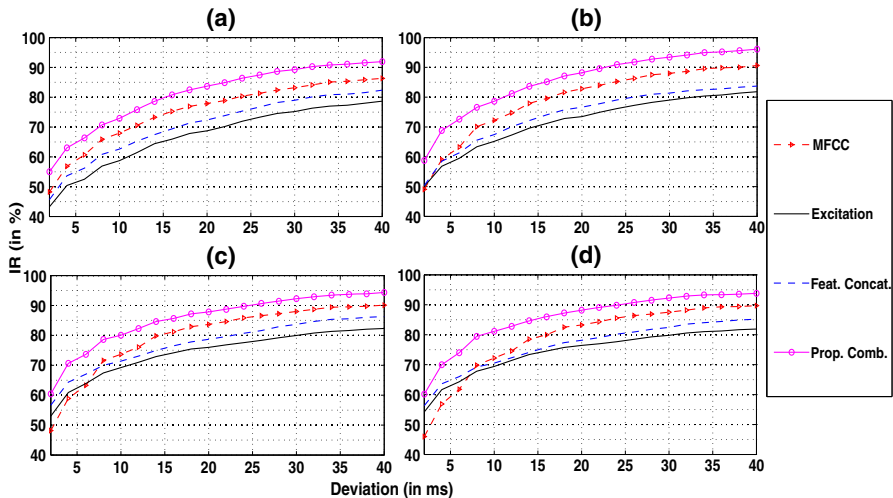


Fig. 7 The IR profiles for the VEPs with respect to the different GMM–HMM systems. The other details remain the same as that for Fig. 5. **a** IR for GMM–HMM (monophone) system, **b** IR for GMM–HMM (triphone) system, **c** IR for GMM–HMM (LDA–MLLT) system and **d** IR for GMM–HMM (fMLLR–SAT) system

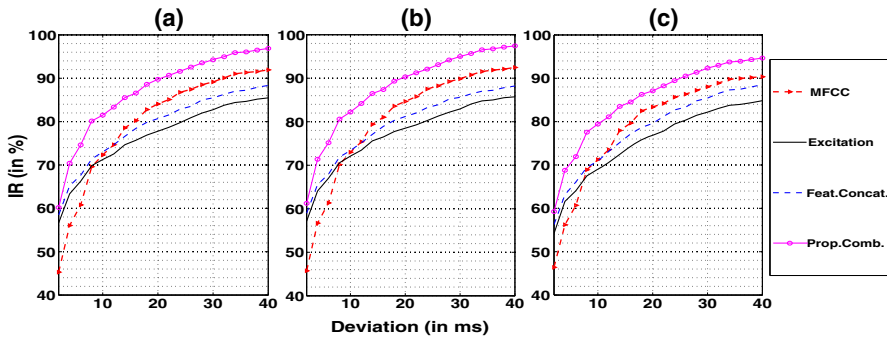


Fig. 8 The IR profiles for the VEPs with respect to the SGMM–HMM and DNN–HMM systems. The other details remain the same as that for Fig. 5. **a** IR for SGMM–HMM system, **b** IR for SGMM–HMM–MMI system and **c** IR for DNN–HMM system

higher in comparison with that of the MFCC feature vectors. This may be due to the lesser differences in the excitation characteristics between the vowel and similar high energy voiced consonants. The poor performance of the proposed feature for the vowel–nonvowel classification compared to that of the MFCCs is mainly due to the improper detection of the VEPs. Again, for all the acoustic modeling methods, the proposed combination of evidences provides significant improvement in the IR in comparison with that obtained by using MFCC features. At the same time, the SR remains nearly the same.

5 Comparison with Existing Methods

As mentioned earlier, a number of explicit signal processing approaches have been proposed over the years for the detection of VOPs/VEPs [28, 35, 40]. In this section, the identification and the spurious rates obtained by the proposed approach are compared with some of the existing techniques.

For comparing the proposed VOP detection approach with the existing approaches, two state-of-the-art methods [28, 35] are considered. The front-end speech parameterization employed in the first method (Method I) [28] consists of the following three features viz. the Hilbert envelope (HE) of the LP residual signal, the sum of the ten largest peaks in the discrete Fourier transform (DFT) and the modulation spectrum energy of the input speech signal. As suggested in that work, the features are smoothed and then enhanced by computing the slope using the first-order Gaussian difference. The evidences for each of those features are obtained by individually convolving with a first-order Gaussian difference (FOGD) operator. Finally, the respective evidences are combined sample by sample to obtain the final evidence for the VOPs.

In the second existing VOP detection approach considered in this work (Method II) [35], first the GCIs are determined using the ZFF method. Next, the DFT is computed for the speech samples present in 30% of glottal cycle starting from the GCI. This is followed by the computation of spectral energy within the frequency band of 500–2500 Hz. Mean smoothing is performed to smoothen out the fluctuations in the spectral

Table 5 Comparison of the IR obtained by the use of existing techniques for determining the VOPs/VEPs and the proposed approach

VOP / VEP	Method	IR in %				SR in %
		± 10 ms	± 20 ms	± 30 ms	± 40 ms	
VOP	Method I [28]	60.02	73.84	81.69	89.27	8.53
	Method II [35]	64.50	77.04	85.43	91.42	6.60
	Prop.	80.35	91.32	97.26	98.80	5.48
VEP	Method I [40]	56.69	71.16	78.80	84.37	7.57
	Prop.	82.22	90.30	95.01	97.43	6.40

The predefined deviation is varied from 10 ms to 40 ms in steps of 10 ms. Also shown is the spurious rate (SR) for the existing as well as the proposed techniques. In the case of the proposed approach, the enlisted performances are with respect to the SGMM–HMM–MMI models

energy contour. In order to enhance the change at the VOP present in the smoothed spectral energy, first-order difference is computed. Significant changes in spectral characteristics are then detected by convolving with the FOGD operator.

In the case of VEP detection, the existing approach considered (Method I) [40] for comparison happens to be slightly different from the VOP detection method described above [35]. In that VEP detection technique, the unwanted peaks with smaller values for the slope in the smoothed spectral energy contour are eliminated using a pre-determined threshold. Moreover, the spectral characteristics are locally enhanced within the region bounded by negative to positive zero crossing points. The valleys are then detected after convolving with the FOGD operator.

The IR and SR for the VOPs/VEPs detected using the existing and the proposed approaches are given in Table 5. In the case of explored existing techniques, for the sake of proper comparison, the respective parameters for the computation of the features and the evidences are chosen to be the same as described in those original works [28,35,40]. In the case of the proposed approach, the enlisted performances are with respect to the SGMM–HMM–MMI models employing the evidence level combination. For both the cases (VOPs/VEPs), the proposed technique is noted to be much superior to the existing explicit signal processing approaches. It is to note that the performances for the existing approaches are observed to be poorer than those reported in the respective original works [28,35,40]. This may be mainly due to the fact that a different test set is employed in the studies presented in this paper than that applied in those works. As already mentioned earlier, the TIMIT dataset was split into orthogonal sets for learning the model parameters and testing using the standard Kaldi recipe. It is quite likely that the chosen orthogonal test set is a much tougher set.

6 Summary and Conclusions

The work presented in this paper deals with the detection of vowel regions in a speech signal. To do the same, a given speech signal is forced-aligned on trained acoustic models to generate the frame-level state alignments. In this regard, seven different

acoustic modeling approaches have been explored in this work. Furthermore, a front-end feature extraction method is also proposed to extract the temporal and the spectral characteristics of the excitation source. The acoustic models are developed using the conventional MFCCs, the proposed excitation feature vectors and the feature vectors obtained by their frame-level concatenation. In addition to that, the vowel regions detected using all the explored features and acoustic models are analyzed in detail. Finally, a novel method is proposed to combine the evidences obtained using the MFCCs and proposed excitation feature vectors to enhance the detection of the vowel regions and their corresponding VOPs and VEPs. The proposed method of combining the evidences is noted to provide significant improvements in the performance for all the acoustic modeling approaches explored in this work. In future, signal processing techniques will be explored to extract the features that will further improve the discrimination of the vowels from the semivowel and the nasal sound units.

References

1. T.V. Ananthapadmanabha, B. Yegnanarayana, Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoust. Speech Signal Process.* **27**(4):309319 (1979)
2. T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul, A compact model for speaker-adaptive training. *Int. Conf. Spoken Lang. Process.* **2**, 1137–1140 (1996)
3. G. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Trans. Audio Speech Lang. Process.* **20**(1), 30–42 (2012)
4. B. Dev Sarma, S.R.M. Prasanna, Analysis of spurious vowel-like regions (VLRs) detected by excitation source information, in *Annual IEEE India Conference*, pp. 1–5 (2013)
5. V. Digalakis, D. Rtischev, L. Neumeyer, Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Trans. Speech Audio Process.* **3**(5), 357–366 (1995)
6. N. Fakotakis, J. Sirigos, A high performance text independent speaker recognition system based on vowel spotting and neural nets. *Int. Conf. Acoust. Speech Signal Process.* **2**, 661–664 (1996)
7. M. Gales, Cluster adaptive training of hidden Markov models. *IEEE Trans. Speech Audio Process.* **8**(4), 417–428 (2000)
8. M.J.F. Gales, Semi-tied covariance matrices for hidden Markov models. *IEEE Trans. Speech Audio Process.* **7**(3), 272–281 (1999)
9. J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, V. Zue, TIMIT acoustic-phonetic continuous speech corpus LDC93S1, vol. 33. Linguistic Data Consortium (1993)
10. D.J. Hermes, Vowel onset detection. *J. Acoust. Soc. Am.* **87**(2), 866–873 (1990)
11. G.E. Hinton, L. Deng, D. Yu, G. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Process. Mag.* **29**(6), 82–97 (2012)
12. Q. Jin, A. Waibel, Application of LDA to speaker recognition, in *INTERSPEECH*, pp. 250–253 (2000)
13. Kaldi Toolkit: <http://kaldi.sourceforge.net>
14. P. Kenny, P. Ouellet, N. Dehak, V. Gupta, P. Dumouchel, A study of interspeaker variability in speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **16**(5), 980–988 (2008)
15. B.K. Khonglah, B.D. Sarma, S.R.M. Prasanna, Exploration of deep belief networks for vowel-like regions detection, in *Annual IEEE India Conference*, pp. 1–5 (2014)
16. R. Kuhn, J.C. Junqua, P. Nguyen, N. Niedzielski, Rapid speaker adaptation in eigenvoice space. *IEEE Trans. Speech Audio Process.* **8**(6), 695–707 (2000)
17. J. Makhoul, Linear prediction: a tutorial review. *Proc. IEEE* **63**(04), 561–580 (1975)
18. L. Mary, B. Yegnanarayana, Extraction and representation of prosodic features for language and speaker recognition. *Speech Commun.* **50**(10), 782–796 (2008)
19. K.S.R. Murthy, B. Yegnanarayana, Epoch extraction from speech signals. *IEEE Trans. Audio Speech Lang. Process.* **16**(8), 1602–1613 (2008)

20. D. Pati, S. Prasanna, Speaker information from subband energies of linear prediction residual. In: National Conference on Communication, pp. 1–4 (2010)
21. D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. Rose, P. Schwarz, S. Thomas, Subspace gaussian mixture models for speech recognition, in *International Conference on Acoustics, Speech and Signal Processing*, pp. 4330–4333 (2010)
22. D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R.C. Rose, P. Schwarz, S. Thomas, The subspace gaussian mixture model—a structured model for speech recognition. *Comput. Speech Lang.* **25**(2), 404–439 (2011)
23. D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The Kaldi speech recognition toolkit, in *Workshop on Automatic Speech Recognition and Understanding* (2011)
24. G. Pradhan, S.R.M. Prasanna, Speaker verification under degraded condition: a perceptual study. *Int. J. Speech Technol.* **14**(4), 405–417 (2011)
25. G. Pradhan, S.R.M. Prasanna, Speaker verification by vowel and nonvowel like segmentation. *IEEE Trans. Audio Speech Lang. Process.* **21**(4), 854–867 (2013)
26. S.R.M. Prasanna, S.V. Gangashetty, B. Yegnanarayana, Significance of vowel onset point for speech analysis, in *International Conference on Signal Processing and Communications*, pp. 81–88 (2001)
27. S.R.M. Prasanna, G. Pradhan, Significance of vowel-like regions for speaker verification under degraded condition. *IEEE Trans. Audio Speech Lang. Process.* **19**(8), 2552–2565 (2011)
28. S.R.M. Prasanna, B.V.S. Reddy, P. Krishnamoorthy, Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. *IEEE Trans. Audio Speech Lang. Process.* **17**(4), 556–565 (2009)
29. S.R.M. Prasanna, B. Yegnanarayana, Detection of vowel onset point events using excitation source information, in *INTERSPEECH*, pp. 1133–1136 (2005)
30. J.Y.S.R.K. Rao, C.C. Sekhar, B. Yegnanarayana, Neural network based approach for detection of vowel onset points, in *International Conference on Advances in Pattern Recognition and Digital Techniques*, vol. 1, pp. 316–320 (1999)
31. S.P. Rath, D. Povey, K. Vesel, Cernock, J.: Improved feature process. for deep neural networks, in *INTERSPEECH*, pp. 109–113 (2013)
32. R.C. Rose, S.C. Yin, Y. Tang, An investigation of subspace modeling for phonetic and speaker variability in automatic speech recognition, in *International Conference on Acoustics, Speech and Signal Processing*, pp. 4508–4511 (2011)
33. B. Sarma, S. Prajwal, S.M. Prasanna, Improved vowel onset and offset points detection using bessel features, in *International Conference on Signal Processing and Communications*, pp. 1–6 (2014)
34. K.N. Stevens, *Acoustic Phonetics* (The MIT Press Cambridge, Massachusetts, 2000)
35. A. Vuppala, J. Yadav, S. Chakrabarti, K.S. Rao, Vowel onset point detection for low bit rate coded speech. *IEEE Trans. Audio Speech Lang. Process.* **20**(6), 1894–1903 (2012)
36. A.K. Vuppala, K.S. Rao, Vowel onset point detection for noisy speech using spectral energy at formant frequencies. *Int. J. Speech Technol.* **16**(2), 229–235 (2013)
37. K. Vuppala, K.S. Rao, S. Chakrabarti, Improved vowel onset point detection using epoch intervals. *AEU-Int. J. Electron. Commun.* **66**(8), 697–700 (2012)
38. J. Wang, C. Hu, S. Hung, J. Lee, A hierarchical neural network based C/V segmentation algorithm for Mandarin speech recognition. *IEEE Trans. Signal Process.* **39**(9), 2141–2146 (1991)
39. J.H. Wang, S.H. Chen, A C/V segmentation algorithm for Mandarin speech using wavelet transforms. *Int. Conf. Acoust. Speech Signal Process.* **1**, 417–420 (1999)
40. J. Yadav, K.S. Rao, Detection of vowel offset point from speech signal. *IEEE Signal Process. Lett.* **20**(4), 299–302 (2013)
41. B. Yegnanarayana, C. Avendano, H. Hermansky, P.S. Murthy, Speech enhancement using linear prediction residual. *Speech Commun.* **28**(1), 25–42 (1999)
42. B. Yegnanarayana, P.S. Murthy, Enhancement of reverberant speech using LP residual signal. *IEEE Trans. Speech Audio Process.* **8**(3), 267–281 (2000)
43. X. Zhang, J. Trmal, D. Povey, S. Khudanpur, Improving deep neural network acoustic models using generalized maxout networks, in *International Conference on Acoustics, Speech and Signal Processing*, pp. 215–219 (2014)