Sādhanā (2018) 43:153
https://doi.org/10.1007/s12046-018-0923-x

Relative occurrences and difference of extrema for detection of transitions between broad phonetic classes

T V ANANTHAPADMANABHA[1], K V VIJAY GIRISH[2,*] and A G RAMAKRISHNAN[2]

[1]Voice and Speech Systems, Malleswaram, Bangalore, India
[2]Indian Institute of Science, Bangalore, India
e-mail: vijay.girish@gmail.com

MS received 31 March 2017; revised 31 December 2017; accepted 20 February 2018; published online 4 August 2018

Abstract. Detection of transitions between broad phonetic classes in a speech signal has applications such as landmark detection and segmentation. The proposed hierarchical method detects silence to non-silence transitions, sonorant to non-sonorant transitions and vice-versa. The subset of the extrema (minimum or maximum amplitude samples) above a threshold, occurring between every pair of successive zero-crossings, is selected from each frame of the bandpass-filtered speech signal. Locations of the first and the last extrema lie on either side far away from the mid-point (reference) of a frame, if the speech signal belongs to a non-transition segment; else, one of these locations lies within a few samples from the reference, indicating a transition frame. The transitions are detected from the entire TIMIT database for clean speech and 93.6% of them are within a tolerance of 20 ms from the phone boundaries. Sonorant, unvoiced non-sonorant and silence classes and their respective onsets are detected with an accuracy of about 83.5% for the same tolerance with respect to the labelled TIMIT database as reference. The results are as good as, and in some aspects better than, the state-of-the-art methods for similar tasks. The proposed method is also tested on the test set of the TIMIT database for robustness with respect to white, babble and Schroeder noise, and about 90% of the transitions are detected within a tolerance of 20 ms at the signal to noise ratio of 5 dB. On NTIMIT database, 62.7% of the transitions are detected, and 63.5% of the sonorant onsets, within 20 ms tolerance.

Keywords. Transitions; phonetic classes; silence index; occurrence of extrema; segmentation.

1. Introduction

1.1 *Segmentation problem*

During speech production, the articulators continually move, resulting in a speech signal with almost continuous formant tracks. Also, the source process is influenced by the preceding or the succeeding phone, as for example the glottal abduction during a vowel–consonant transition, the presence of frication noise following a burst or the presence of noise components at the onset of a vowel following a strong fricative. Thus, the adjacent phones have a considerable influence on the temporal and spectral properties of a short segment of speech corresponding to the so called current phone [1]. However, we perceive clean speech as if it is made up of a sequence of distinct sounds, thus evoking an expectation that the signal can be segmented into non-overlapping intervals corresponding to phones. Hence, speech segmentation is a challenging problem. Despite the lack of phone-wise segmentation property in a speech signal, there are clearly marked events or transitions or landmarks arising due to an abrupt change of source process (voiced/unvoiced) and/or an abrupt movement of an articulator (sudden release as in stops, switch over from oral to nasal output as for nasals). Detection of such events serves to guide semi-automatic segmentation, variable frame-rate analysis or analysis around landmarks to extract distinctive features (DFs) or manner classes [2] or phonetic features (PFs).

1.2 *Literature review*

There are three broad approaches to segmentation: (i) sequential, non-overlapping segmentation based on phones, (ii) parallel, multiple segmentations based on DFs and (iii) hierarchical segmentation based on PFs. Classification of a speech signal, phone-wise or feature-wise, can also be interpreted as performing segmentation, since it automatically divides the speech signal into distinct segments.

The first view, namely, the phone-based segmentation, is motivated by the perception of speech as a series of distinct units. As per this view, any speech signal is a sequence of non-overlapping intervals, each representing a phone. This

*For correspondence

assumption is widely used in manual labelling of the speech databases. Such a labelling scheme contradicts the acoustic-phonetics knowledge that a given frame of speech signal is strongly influenced by the neighboring phones. However, it is understood that the manual labelling of phones must be considered along with the surrounding context of phones. In phone level segmentation, abrupt changes in the short-time spectra are marked as transition events [3–9]. Various short-time spectral representations have been used: linear prediction smoothed spectral envelope, ensemble interval histogram, auditory sub-band filter outputs, mel frequency cepstral coefficients (MFCCs), weighted MFCCs, etc. In addition to the standard Euclidean and Mahalanobis distance measures [3], cross-correlation of short-time spectra [4], model fitting [6], maximum likelihood estimates and template matching [7] have also been used to detect segment boundaries.

The second approach to segmentation is based on DFs, a view based on phonology. It is postulated that each speech sound is a bundle of (about 16) binary DFs [10, 11]. In this model, speech signal consists of a parallel stream of DFs. The presence of each of the DFs extends over different, overlapping intervals with their own boundaries. Acoustic description of DFs given by Jakobson *et al* [10] has remained qualitative in nature since there is no robust automatic method to extract these descriptors. Chomksy and Halle [11] have proposed articulation-based DFs. In order to extract these DFs, King and Taylor [12] used frame-wise analysis with MFCCs and their derivatives (39 features) as the acoustic feature vector input to a neural network classifier trained for each DF separately, and obtained a high (>90%) frame-wise accuracy for the individual DF. However, the accuracy for the joint or simultaneous occurrence (all correct) of the DFs for a given phone is low (around 50%).

The third approach, based on the PFs, has two models. One of the models is based on the manner and place classification of speech sounds, a view inspired by the process of speech production. This is similar to the approach of DF but with multi-valued features and only two parallel streams (manner and place). In their work on DFs, King and Taylor [12] also reported on the identification of manner and place features. The reported frame-wise accuracy is about 90% for the individual features. A later extension of this study attempted to incorporate mutual dependences amongst DFs [13] and found a marginal improvement in the accuracy. Juneja and Wilson [14, 15] report manner class (silence, vowel, sonorant consonant, fricative and stop) segmentation accuracy of about 79% [15] on a part of the test set of the TIMIT database, using MFCCs as well as acoustic parameters and support vector machine classifier.

An issue related to the extraction of PFs is the landmark detection, landmark being an important transition dividing a speech signal into certain broad segments [16–18]. Conventionally, speech analysis for the extraction of acoustic features is carried out frame-wise. However, in this alternative approach, speech signal around the landmarks is analysed to extract the acoustic features, which are subsequently given as an input to a classifier to determine either the phones or PFs. Liu [18] has used the change of energy, over six sub-band signals, between two frames spaced 50 ms apart, for detecting four broadly defined landmarks. Salomon *et al* [17] have used a set of 12 temporal parameters to detect three landmarks as well as for manner classification.

The work reported in this paper can be considered as an approach to segmentation of speech into broad classes, when the phone sequence is unknown. The superior speech perception performance of humans in degenerate conditions [19] suggests that humans employ a representation of speech, which has more mutual information with phonetic classes. In a recent study, Mesgarani *et al* [20] used high-density direct cortical surface recordings in humans and found response selectivity to distinct PFs in the superior temporal gyrus (STG), which shows acoustic-phonetic representation of speech in human STG.

1.3 *About this work*

Reddy [21] proposed the use of intensity differences (peaks and valleys) to detect certain broad classes of sounds for a limited vocabulary, speaker-dependent task. Stevens [16] has observed that certain landmarks may be located based only on abrupt amplitude changes in a speech signal. This paper proposes four different measures for detecting transitions between broad phonetic classes in a speech signal based on abrupt amplitude changes.

A measure is defined on the quantized speech signal to detect transitions between very low amplitude or silence (S) and non-silence (N) segments. These S-segments could be stop closures, pauses or silence regions at the beginning and/or ending of an utterance.

We propose two other measures to detect the transitions between sonorant and non-sonorant segments and vice-versa. We make use of the fact that most sonorants have higher energy in the low frequencies, than other phone classes such as unvoiced fricatives, affricates and unvoiced stops. For this reason, we use a bandpass speech signal (60–340 Hz) for extracting temporal features. For a transition within a sonorant (vowel to voiced consonant or vice-versa), the amplitude of the bandpass-filtered speech signal does not change appreciably. However, for a transition from a sonorant to any of the unvoiced consonants, the amplitude changes suddenly from a relatively high to a low value across the transition. The converse is also true. Thus, by tracking the relative locations of extrema in successive closely spaced analysis frames, we can detect the transitions between relatively high (H) and low (L) amplitude segments and hence the broad phonetic classes in a speech signal.

When the amplitude of the bandpass-filtered signal in a frame is very low, any change in the relative amplitude level is not reliable. Thus, when the mean difference between extrema amplitudes in a frame is very low, any transition is ignored.

The afore-mentioned rationale for the selection of features and the proposed algorithm for detecting transitions are based on the acoustic-phonetic knowledge of the different classes of speech sounds.

Combining the afore-mentioned types of transitions, the speech signal is divided into five broad homogeneous classes: silence (S), high (H), low (L), high-low (HL) and low-high (LH). Based on the homogeneous classes, the speech signal is classified into the broad phonetic classes of sonorants, non-sonorants and silence. The proposed method is validated using the TIMIT database under clean and noisy conditions. The accuracy of detection and the temporal accuracy of the onset of these classes are computed. The results are noted to be comparable to those of state-of-the-art methods.

The proposed method is clearly distinct from the 'old' approaches of 1960s as well as the later 'Rate-of-Rise' (ROR) approaches. These earlier approaches are based on 'short-time energy contour', where, if the duration of the analysis frame is too short, there will be rapid fluctuations in the energy contour and if the frame duration is large, energy contour smears the transitions. Such disadvantages are not present in the proposed method, where the positive and negative thresholds adapt to the statistics of the peaks and valleys in the analysis frame. The proposed method is based on a novel method of computing 'temporal envelope' properties based on the amplitudes of successive peaks and valley-to-peak amplitude, rather than the short-time energy.

2. Proposed temporal features

From the normalized speech signal, we derive temporal features that are independent of the amplitude level (gain) of the signal. The parameters used are derived from a development set.

2.1 *Pre-processing*

The speech utterance $s[n]$ is normalized after removing the mean value. Frames extracted from the normalized utterance $s_N[n]$, using a uniform frame shift of 5 ms, are used for deriving the temporal features.

2.2 *Silence index*

Usually a threshold on energy is used to detect a silent segment. Here, we propose an alternate method, by defining a new measure called silence index (SI).

The silence segments within an utterance have a much lower spectral dynamic range (6–12 dB) compared with a speech segment (20–40 dB or more). The low spectral dynamic range results in a time domain signal without heavy, sudden fluctuations. Hence, we gross quantize the signal by removing the seven least significant bits, after which samples in most silent segments have the same value. The normalized speech signal of 16-bit resolution is quantized to 9 bits by shifting right by 7 bits and a staircase signal is obtained. The size of the analysis frame is 10 ms. Whenever there are a minimum of three successive samples having the same value and whose absolute values are up to a threshold (two times the quantization level, 2^7), these samples are counted for the calculation of SI.

SI is a dimensionless ratio (between 0 and 1), defined as follows:

$$SI = \frac{\text{count of samples below threshold}}{\text{number of samples in the frame}}. \quad (1)$$

Since a frame shift of 5 ms is used, there is a new value of SI for every 5 ms segment.

Figure 1 shows the signal and its quantized counterpart, together with the SI values for three types of speech segments, each containing three overlapping frames: (a) silence containing a noisy impulse, (b) unvoiced and (c) a closure–burst transition segment. It may be noted that SI has a very high value, as desired for the silence segment, even in the presence of a large amplitude impulse. The SI is low for the unvoiced segment. During a closure–burst transition, there is a sharp decrease in the value of SI for two successive frames. We make use of such abrupt changes in the value of SI for detecting the transitions from/to silence segments.

2.3 *Features based on the extrema in a frame*

Features based on extrema are used to detect transitions from/to a sonorant segment. As sonorants have significant energy in the low frequencies below 500 Hz as compared with other segments, the normalized speech signal is bandpass filtered (BPF) using the following bell cosine shaped filter in the frequency domain:

$$h_B[f] = \begin{cases} 0.5 - 0.5\cos\left(\pi\left(\frac{f - f_1/2}{f_1/2}\right)\right) & f_1/2 \le f < f_1 \\ 1 & f_1 \le f \le f_2 \\ 0.5 + 0.5\cos\left(\pi\left(\frac{f - f_2}{f_2}\right)\right) & f_2 < f \le 2f_2 \\ 0 & \text{elsewhere} \end{cases}$$

$f_1 = 70$ Hz, $f_2 = 250$ Hz and $h_B[f]$ is the frequency response of the filter. This filter has a cosine rising function from 35 to 70 Hz, unit gain from 70 to 250 Hz and cosine falling function from 250 to 500 Hz. The 3 dB frequencies

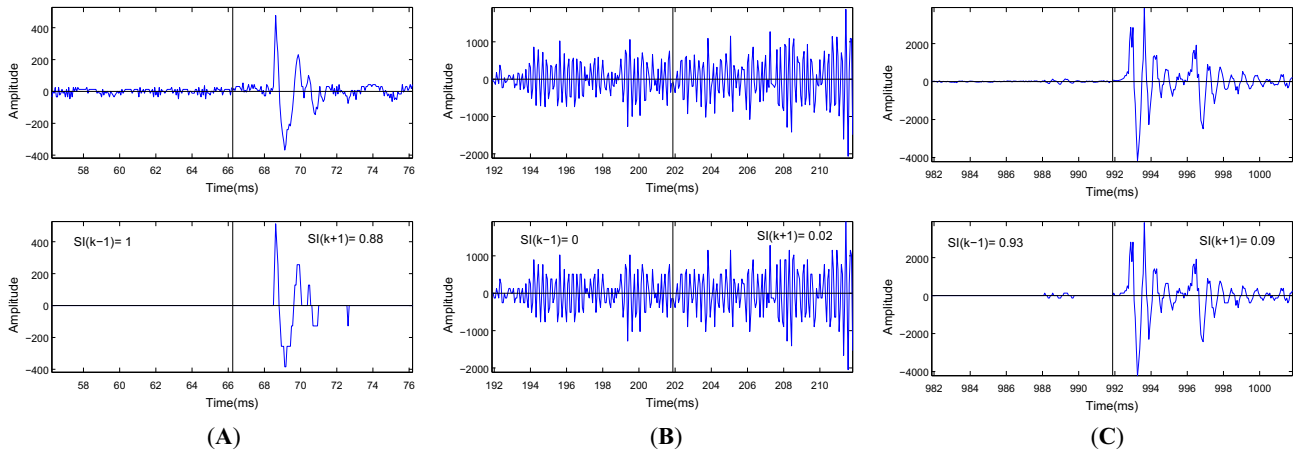


Figure 1. Variation of SI values with the variation in the nature of the signal across consecutive frames. Speech signal and the corresponding quantized signals for (**a**) presence of a high amplitude pulse in a silence segment. (**b**) An unvoiced segment. (**c**) A stop closure–burst transition. (Note that the *y*-scales are different for the three plots.).

of the bandpass filter are 60 and 340 Hz. This is close to 'Band 1' used by Liu [18] for landmark detection. The corresponding BPF signal $s_B[n]$ is given as

$$s_B[n] = s_N[n] * h_B[n] \tag{2}$$

where $h_B[n]$ is the impulse response corresponding to $h_B[f]$. The BPF signal $s_B[n]$ is analysed with a frame size of 40 ms, twice the pitch period corresponding to the assured minimum value of fundamental frequency of 50 Hz.

2.3a *Selection of extrema based on a dynamic two-pass threshold*: Let s_B^j denote the BPF signal between the first and the last zero crossings in the jth frame. We define features based only on those extrema in s_B^j remaining after a 2-pass, frame-adaptive threshold.

The first-pass positive threshold T_{P1}^j is defined as

$$T_{P1}^j = 0.5 \times mean(\{s_B^j[n]\})\ \forall s_B^j[n] > 0. \tag{3}$$

From s_B^j, all the positive peaks p_B^j between successive zero crossings are obtained. A subset of these peaks is selected as

$$p_{B1}^j = \{p_B^j,\ \forall\ p_B^j > T_{P1}^j\}. \tag{4}$$

The second-pass positive threshold T_{P2}^j is defined as

$$T_{P2}^j = 0.5 \times mean\{p_{B1}^j\}. \tag{5}$$

The set of peaks after the second pass is obtained as

$$p_{B2}^j = \{p_{B1}^j,\ \forall\ p_{B1}^j \geq T_{B2}^j\}. \tag{6}$$

The factor of 0.5 is applied to compute the threshold based on the intuition that in the case of a midframe transition, half of the peaks p_{B1}^j in the frame might lie below and the rest above T_{B2}^j. Similarly, from the valleys (negative peaks) between successive zero crossings, the set of valleys v_{B2}^j is obtained. Figure 2a shows a segment of voiced frame /ae/, its corresponding BPF output and the first zero crossing, the first and second-pass positive thresholds. It is to be noted that the peaks obtained after the first and second-pass are the same for this non-transition frame.

2.3b *Relative occurrences of first and last extrema in a frame*: The times of occurrence of the first extremum (OFE) and the last extremum (OLE) in p_{B2}^j or v_{B2}^j are measured with respect to the mid-sample of the frame as the relative time reference. Thus, occurrences ahead of the reference have a negative value. The values of OFE and OLE are treated as the features of the mid-5-ms segment of the frame.

For both the voiced and unvoiced speech segments shown in figure 2a and b, OFE and OLE occur long before and after the reference instant, i.e., $OFE \ll 0$ and $OLE \gg 0$. We notice that the extrema corresponding to OFE and OLE lie on either side, far away from the reference instant of the frame, i.e., $OFE \ll 0$ and $OLE \gg 0$. Thus, this property of $OFE \ll 0$ and $OLE \gg 0$ is satisfied whenever the speech signal within a frame corresponds to a homogeneous class. For a transition from a voiced to unvoiced signal and vice-versa, it is intuitive to observe an OLE or OFE close to zero, i.e., mid-sample of the frame. Figure 2c shows a transition from a voiced to an unvoiced segment. Here, OFE is highly negative and OLE has a low negative value. The converse is true for an unvoiced to voiced transition as shown in figure 2d. Thus, the algorithm works because there are abrupt changes between the segments before and after the transitions.

From these illustrations, we can deduce the following: (a) when $OFE \ll 0$ and $OLE \gg 0$, the frame corresponds

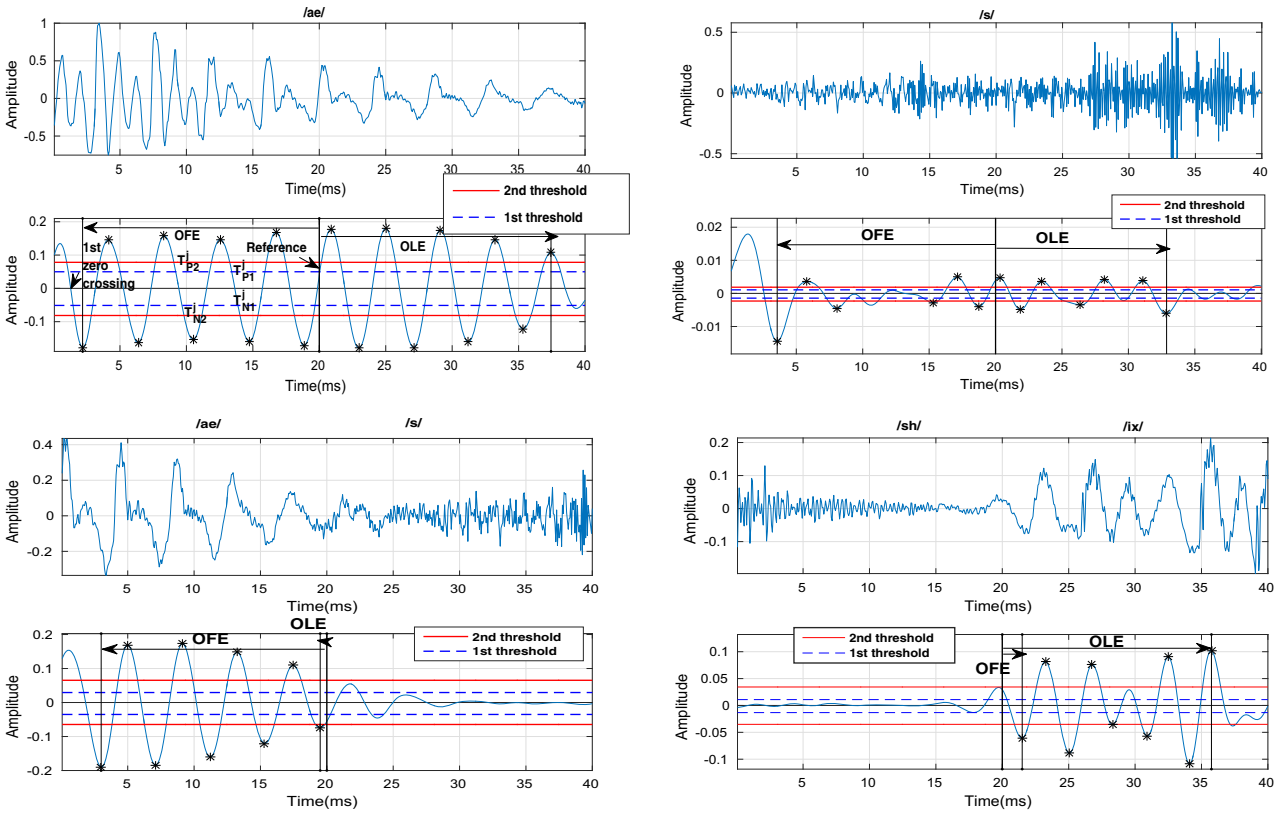


Figure 2. Speech signal (top) of some sample frames and their corresponding BPF versions (bottom). The first-pass threshold T^j_{P1} (T^j_{N1}) is half the mean of all the positive (negative) valued samples in the analysis frame. The second-pass threshold T^j_{P2} (T^j_{N2}) is half the mean of all peaks (valleys) above (below) the first-pass threshold. The extrema above the second-pass thresholds (horizontal lines above and below zero) as well as the occurrences of the first (OFE) and last extrema (OLE) are shown. (**a**) A homogeneous voiced segment (MADE = 0.32). (**b**) A homogeneous unvoiced fricative segment. Notice the very low amplitude of the bandpass signal in this case, and MADE = 0.01. (**c**) A voiced–unvoiced transition (MADE = 0.31). (**d**) An unvoiced–voiced transition (MADE = 0.15).

to a homogeneous class, (b) $OFE \ll 0$ and $OLE \approx 0$ for a frame with a transition from a relatively high to a low amplitude (H-L) and (c) $OFE \approx 0$ and $OLE \gg 0$ for a frame with a transition from a relatively low to high amplitude (L–H). Thus, we can divide the speech signal into a homogeneous class (H-class or L-class) and the two types of transitions H–L and L–H.

2.3c *Mean absolute difference between extrema (MADE) within a frame*: The peak values of the BPF signal in figure 2b are very low (≈0.005). The transitions in such frames are ignored, since the whole frame corresponds to a non-sonorant segment. For this purpose, another measure named mean absolute difference between extrema (MADE) is introduced, which is the mean of the absolute differences between successive peaks and valleys after the second thresholds. The caption for figure 2 also gives the values of MADE for each of the sample signals shown. Figure 3 shows the histogram of MADE for sonorant and non-sonorant frames for 20 randomly selected files from the TIMIT database, used as a development set. The histogram suggests an optimal threshold of 0.024 for sonorant/non-sonorant classification. Transitions corresponding to OFE and OLE are ignored when MADE is below this threshold. Spurious detections of transitions in unvoiced segments and due to frication noise following bursts are avoided and we detect only transitions from/to sonorant segments.

3. Algorithm for the detection of transitions

We refer to the proposed algorithm shown in figure 4 as AGR algorithm. We discuss the strategy used in the algorithm with an example. The first step divides the speech signal into silence and non-silence segments.

3.1 *Detection of transitions between silence and non-silence classes*

To arrive at a threshold, the histogram of SI values for silence (S) and non-silence (N) frames for the development

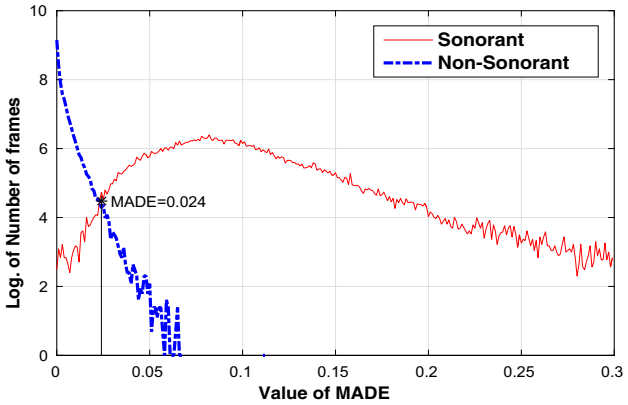


Figure 3. Histogram showing the distribution of computed values of framewise mean absolute difference between extrema (MADE) for 20 randomly selected files from TIMIT database.

set is computed and is plotted in figure 5. A threshold value of $SI = 0.5$ is chosen to distinguish between silence and non-silence frames. The temporal accuracy for the transition is improved by recomputing SI in the mid-5-ms region for non-overlapping sub-segments of 1 ms duration. The instant of transition is defined as the point when SI crosses the value of 0.5 in any direction. The samples between an N–S and a following S–N transitions are labelled as S-class and vice-versa.

SI values obtained for a speech segment containing several phones and the detected transitions are shown in figure 8a. The S–N transitions around 200 ms corresponding to the boundary between 'h#' and /sh/ and around 770 ms corresponding to the boundary between /dcl/ and burst /d/ are detected successfully. However, the /dcl/ segment (550–580 ms) between /eh/ and /jh/ is missed, since considerable energy is present in the corresponding segment. The phone /jh/, normally a voiced affricate, is realized as unvoiced in this utterance. It is not clear if a closure needs to be necessarily marked for an affricate realized as a fricative. Despite the presence of a noticeable impulse, /kcl/ segment from 920 to 980 ms is correctly identified as S-class.

3.2 *Detection of transitions between sonorant and non-sonorant classes*

A transition from a non-sonorant to a sonorant class is mostly detected as L–H and vice-versa.

Suppose there is a vowel, followed by an unvoiced or voiced closure. Figure 6 shows three consecutive frames of a speech segment where a strong H–L transition occurs from a vowel /ah/ to a unvoiced closure /kcl/. It is seen that OLE decreases from a positive value (figure 6a), crosses zero (figure 6b) and then further decreases to a negative value (figure 6c). Figure 7 shows three consecutive frames of a speech segment where a weak H–L transition occurs from a vowel /ih/ to a voiced closure /dcl/. It is seen that OLE decreases from a positive value (figure 7a) to a minimum near zero (figure 7b), and then without any zero crossing again increases (figure7c).

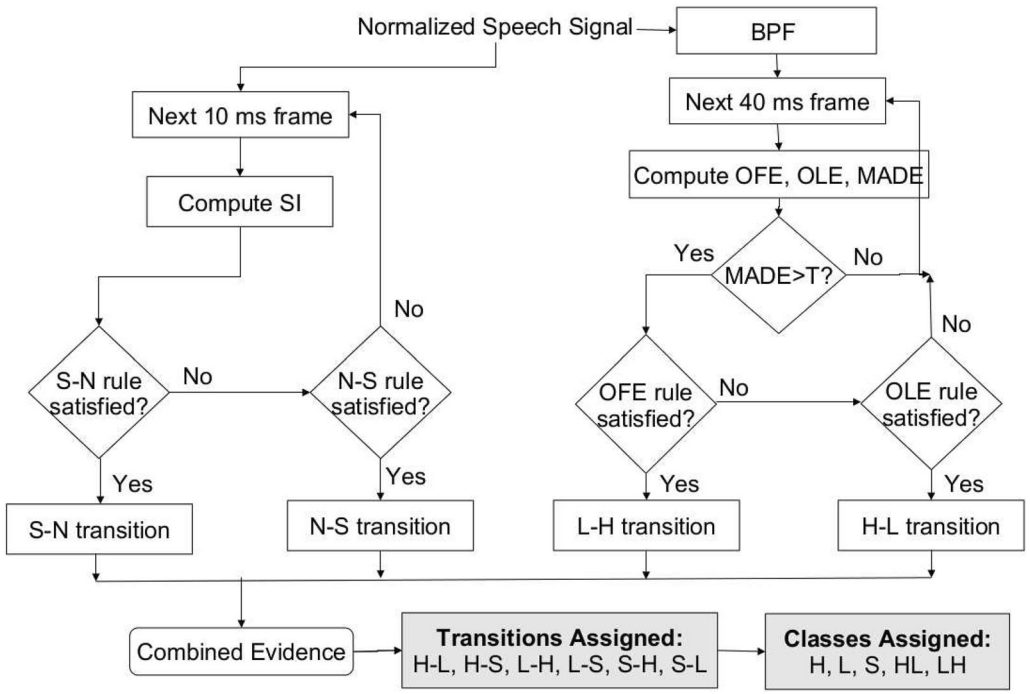


Figure 4. Flowchart for the detection and class assignment of transitions (T is the threshold for MADE).

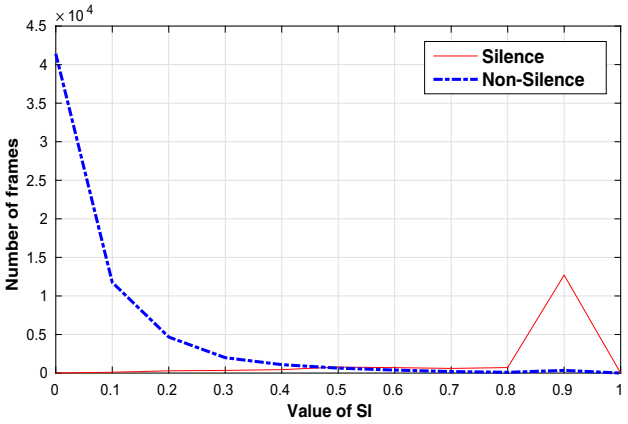


Figure 5. Histogram of framewise silence index of 20 randomly selected files from TIMIT database.

As long as the analysis window contains only the vowel part, most peaks and valleys in the BPF signal have comparable, high amplitudes. Hence, the first and the last extrema occur at the beginning and end of the analysis frame; this makes the values of OFE and OLE highly negative and positive, respectively, with respect to the centre of the frame. Once the closure region enters the analysis window, the occurrence of the last extremum (still from the vowel region only) slowly moves towards the centre, reducing the OLE value from a high positive value towards zero (see figures 6 and 7a and b). However, OFE remains highly negative, since there is a part of the vowel still at the beginning of the analysis window. Now, since nearly half the analysis window contains the closure signal, the first-pass threshold reduces.

In the case of unvoiced stop, there are no peaks in the closure interval and hence the second-pass threshold remains almost the same, being decided only by the extrema of the vowel. Thus, after the next frame shift, OLE moves further to the left, beyond the centre (see figure 6c). Thus, OLE reduces from a high positive value, becomes zero and then goes negative. Thus, OFE having a consistently high negative value and OLE having a zero crossing from positive to negative value denote a high (low-frequency) amplitude phone (say, a vowel) to a low amplitude phone (say, unvoiced stop or fricative) transition or a strong H–L transition. Similarly, OLE having a high positive value and OFE having a zero crossing from positive to negative value denote a low to high amplitude transition.

In the case of voiced stop (following a vowel), there are small but definite peaks in the closure interval, which bring down the second-pass threshold also. After the next frame shift, the voiced closure region enters the first half of the analysis window, further bringing down the second-pass threshold (see figure 7c). Now, most of the peaks and valleys in the closure region survive the low second-pass threshold. Thus, OLE again becomes highly positive, rather than becoming negative. Thus, OFE having a consistently negative value and OLE going through a minimum within 5 ms (the duration of a frame shift) from the centre also denotes a H–L transition. However, to distinguish it from the afore-mentioned scenario, we call this as a weak H–L transition. Thus, in a weak H–L transition, the OLE, rather than going through a PZC to NZC, actually goes through a minimum and again increases. Similarly, OLE having a high positive value and OFE going through a maximum near zero and again decreasing denotes a weak L–H transition.

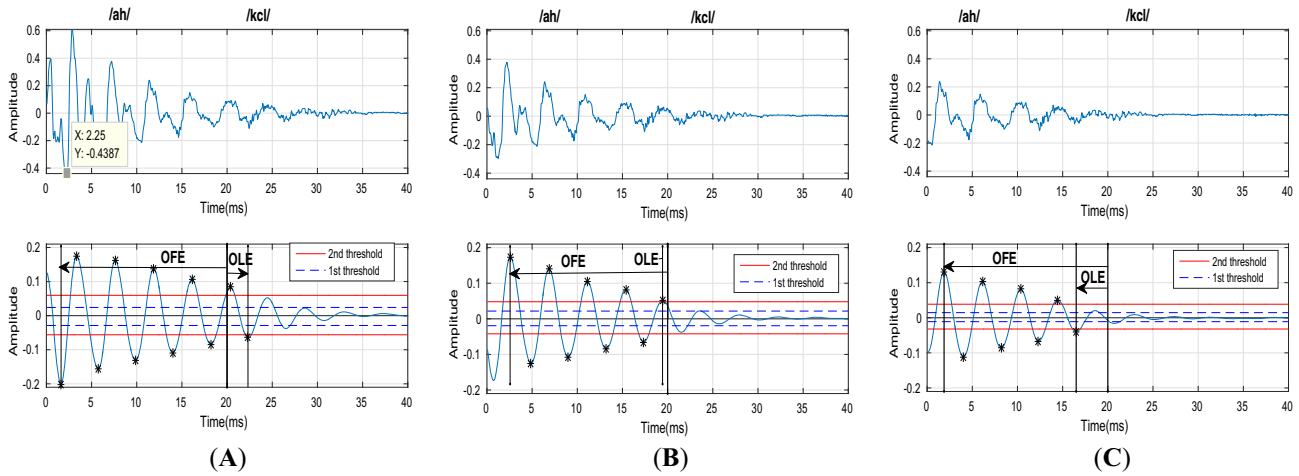


Figure 6. The signals and their bandpass-filtered versions of three consecutive frames of speech containing a strong H–L transition from the vowel (/ah/) to an unvoiced closure (/kcl/). The occurrences of the first (OFE) and last extrema (OLE), and the first and second-pass thresholds are shown, in each case. Plots of OFE and OLE show that OLE goes through a positive to negative zero crossing.

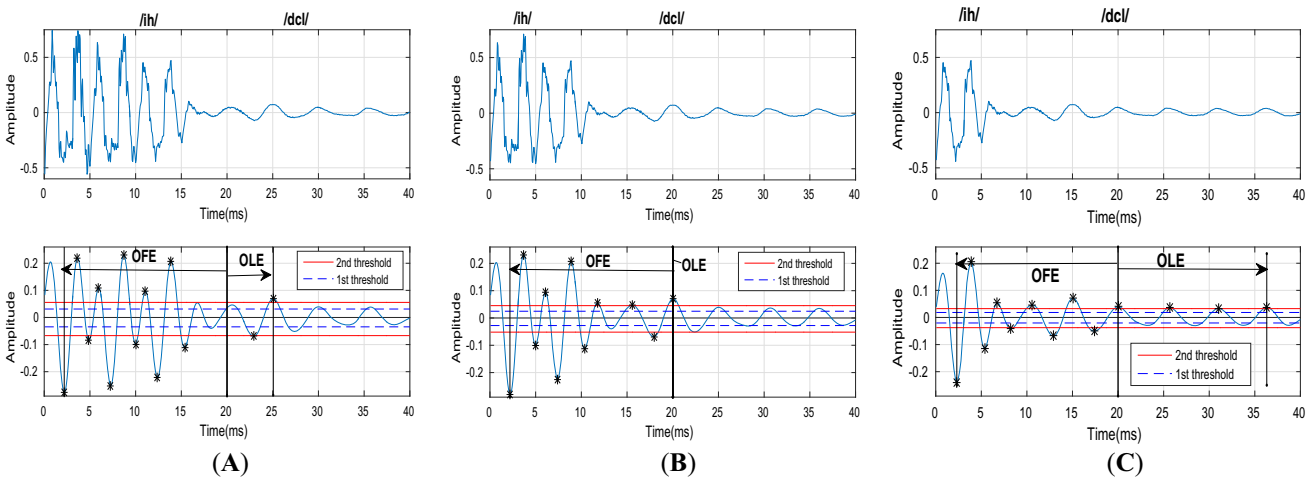


Figure 7. The signals and their bandpass-filtered versions of three consecutive frames of speech containing a weak H–L transition from the vowel (/ih/) to the voiced closure (/dcl/). The occurrences of the first (OFE) and last extrema (OLE), and the first- and second-pass thresholds are shown, in each case. Plots of OFE and OLE show that OLE goes through a minimum near zero.

An L–H transition is detected near the PZC to NZC of OFE as seen in figure 2d. Appropriate class labels are assigned to the segments between successive transitions. For example, in the simplest case, a segment that lies between H–L and L–H transitions would naturally be labelled as L-class.

Figure 8b shows the bandpass version of the segment of speech signal shown in figure 8a. The values of OFE and OLE obtained for successive frames are scaled and plotted as a function of time in figure 8b. Towards the end of /sh/, just before 300 ms, OFE rapidly increases from a negative to a positive value and returns to a negative value for the next phone. The PZC to NZC in OFE marks a strong L–H transition. We choose the zero-crossing in the BPF signal closest to this NZC as the transition instant.

During the H–L transition from /ih/ to /dcl/ around 720 ms, there is an abrupt decrease in amplitude (unlike /eh/ to /dcl/). OLE decreases rapidly to a minimum, close to the base line, without a sign change. If this minimum value of OLE is within 5 ms, then it is considered a weak H–L transition. The value of 5 ms arises because of the frame shift. In the next analysis frame, the voiced closure enters the first half of the window, thus reducing the second-pass threshold. This renders the extrema of the voiced closure region to go above the threshold, thus taking the value of OLE back to a high positive value. A similar weak L–H transition due to OFE is seen between /dcl d/ and /ah/ around 800 ms. Thus the so called weak transitions are also genuine transitions. The distinction between a strong and weak transition is noted only for the sake of further analysis, if required.

It is not necessary that L- and H-classes always alternate. It may be noted that across /kcl-k/ and /s-ux/, there are two consecutive L–H transitions due to OFE. In order to distinguish such transitions, the segment between two consecutive L–H transitions is denoted as HL-class. Similarly the signal between two consecutive H–L transitions is labelled as LH-class. Occurrences of LH- and HL-classes are rare. However, this specific example of HL-class is an exception. Though the segment /k s/ should have been labelled as HL-class, the first transition across /kcl-k/ due to OFE is ignored since MADE is below threshold and hence the label happens to be L. The transition /kcl k/ is still captured as S–N transition based on SI as shown in figure 8a. Thus, the class label of a speech segment needs to be decided by combining the information provided by SI and OFE/OLE.

3.3 *Class assignment based on combined evidence*

Since the transitions between H and L are detected independent of the transitions between N and S, these two evidences are combined to get a single stream of transitions. For example, a silence followed by a sonorant gives rise to both L–H and S–N transitions. Such simultaneous transitions are merged into a single transition. Hence, decisions need to be made on the temporal spacing allowed between the two types of transitions to merge them into one and the same transition and on the location of the new, merged transition. Further, the segment before an N–S transition is labelled as H- or L-class if the preceding transition is an L–H or H–L transition, respectively. Thus, the following five classes result after combining the evidences from the two types of transitions: (a) H, (b) L, (c) S, (d) HL and (e) LH.

Figure 8c shows the class labels assigned after the evidence combination. An S–N transition around 200 ms is followed by an L–H transition around 280 ms. Since the two transitions are spaced beyond 10 ms, both are retained. The N-class segment between S–N and L–H is assigned to the L-class.

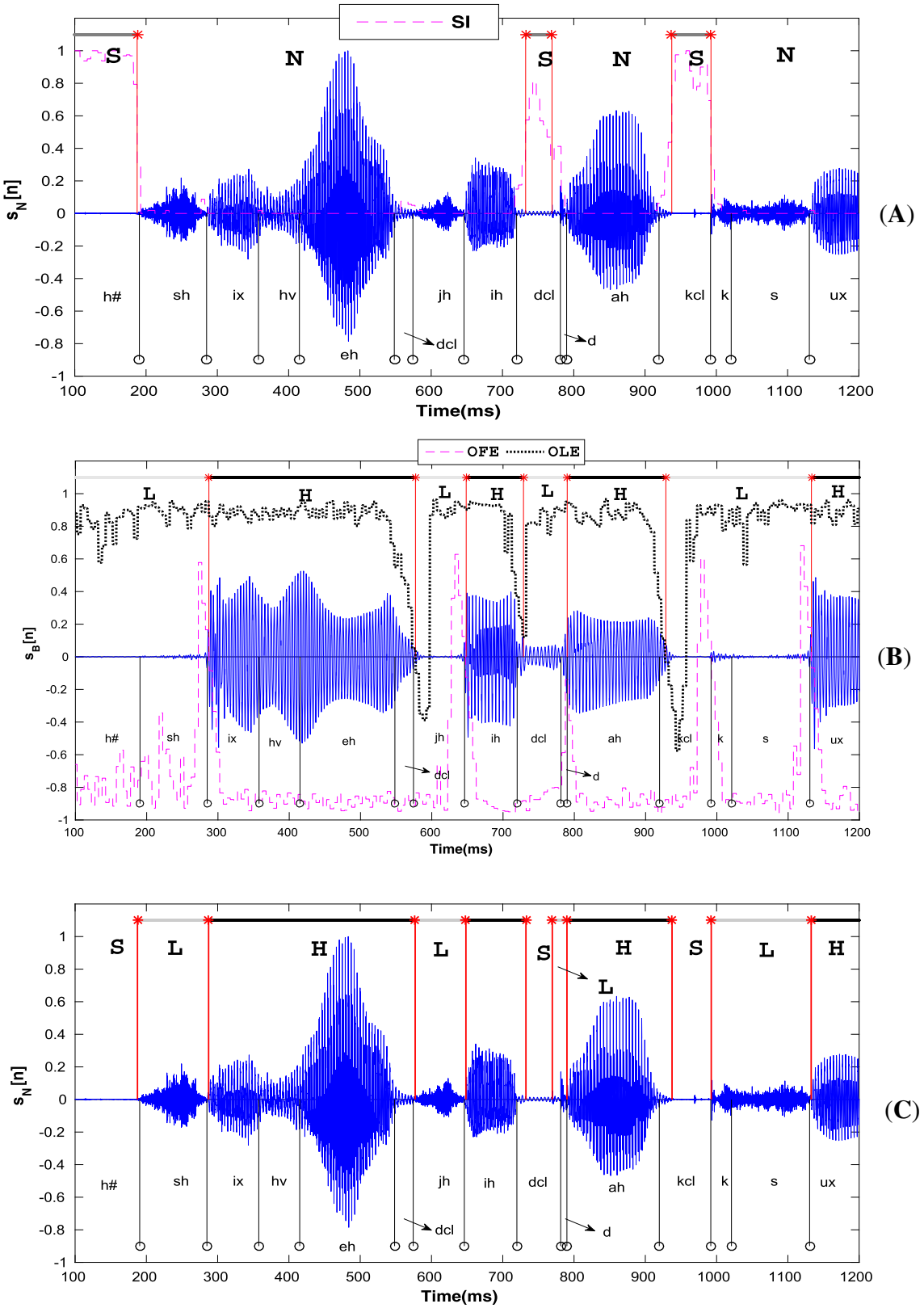


Figure 8. (**a**) S–N and N–S transitions (starred markers) detected using SI values derived from the original signal. (**b**) L–H and H–L transitions detected using the OFE/OLE values derived from the BPF signal. (**c**) The merged transitions.

4. Experimental details and evaluation

The proposed AGR algorithm has been validated on the entire TIMIT database [23], i.e., both training and test databases for clean speech. It consists of several dialects of North American English, totaling 6300 utterances spoken by 630 speakers. For evaluation on noisy speech, only the test database is used, as listed in table 1. We have also evaluated on telephone bandwidth speech from NTIMIT database. The TIMIT database has hand labels at the phone level and the closure durations of stops have been explicitly marked. Accordingly, the class 'stops' denotes 'stop bursts'.

Every detected transition is uniquely assigned to the nearest TIMIT boundary. The statistics of the temporal differences between the labelled boundaries and the assigned transition instants are computed. The boundary detection accuracy is measured for different values of temporal tolerance.

In order to study the relationship between the manner of articulation and the homogeneous segments, the distribution of each class of phones among the five classes is computed. Phonetic grouping given in TIMIT database is used as the reference for assigning the class of phones.

For every sonorant and non-sonorant onset in the labelled database, we verify if there is a detected transition within a specified temporal tolerance. If no transition has been detected within the tolerance for an onset, then it is a case of miss or deletion. This measures the accuracy of detection of onsets relative to the type of transition. A detected transition for which there is no associated labelled boundary is counted as an insertion. The ratio of the number of insertions to the total number of transitions detected is another performance measure.

5. Results and discussion

We first present results on clean speech. The results presented correspond to the total number of frames of 3,818,197 and the total number of detected transitions of 144,715.

Table 1. Databases used for evaluation of performance on clean, noisy and telephone bandwidth speech.

	Clean speech	Noisy speech	Telephone bandwidth speech
Database used	TIMIT training and test set	TIMIT test set: 168 speakers, 1344 sentences	NTIMIT training and test set

5.1 *Temporal accuracy of detection*

Figure 9a shows the histogram of the temporal deviations of the detected transitions from the hand-labelled boundaries, using a bin size of 5 ms. The mean and standard deviation are –1.62 and 17.05 ms, respectively. We observe that 36.4% of the detections are within $\pm$2.5 ms.

The detection accuracy is computed for different values of the temporal tolerance, namely, 5–40 ms in steps of 5 ms. The ratio of successful detections to the total number of transitions, excluding insertions, is computed as the detection accuracy of transitions and is shown in figure 9b as a function of temporal tolerance; 57.8% of the transitions lie within $\pm$5 ms and 98% of the transitions lie within $\pm$40 ms. Thus, the temporal resolution of detection is higher than those of related previous works to be presented in section 5.5.

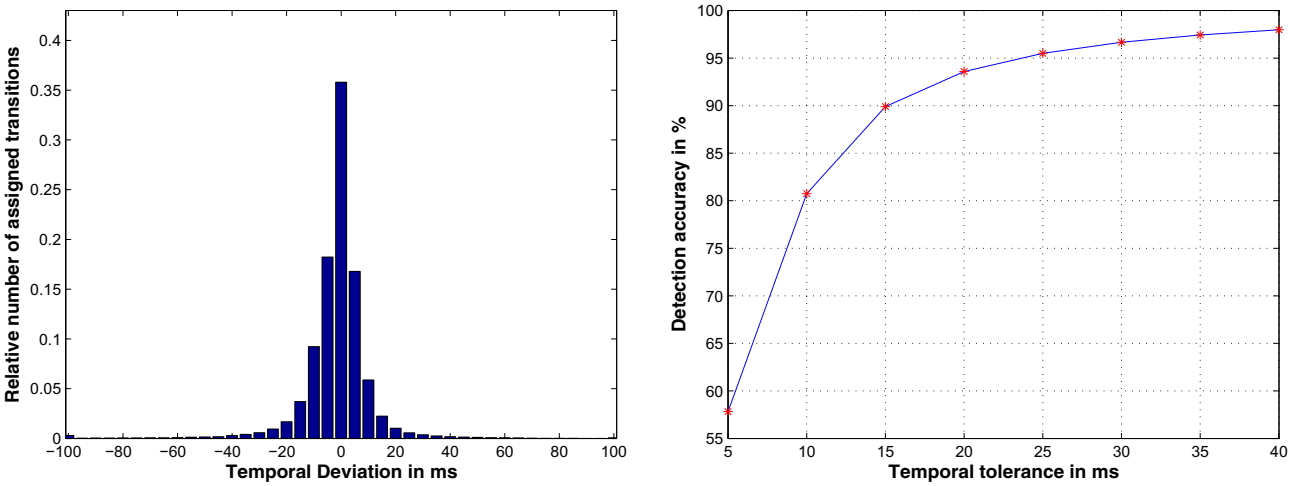
5.2 *Classes of phones detected across each type of transition*

It is of interest to know the distribution of various classes of phones (vowels, semivowels, etc.) that belong to the five broad classes obtained: H, L, S, HL and LH. This distribution is listed in table 2, for a temporal tolerance of 20 ms. More than 91% of vowels belong to H-class. But we note that there are about 4.8% of vowels in L-class and 0.5% in S-class. About 41% of the 'ax-h' phones lie in L-class, since it has the characteristics of unvoiced speech, with a very low amplitude in the BPF signal. Amongst the semivowels, 74.0% of 'hh' lie in the L-class, since this phone also has characteristics similar to those of unvoiced speech. It is seen that in any nasal-fricative segment, there is a short interval of silence at the end of the nasal, which is however not hand labelled as silence. This explains the occurrence of about 6.8% of nasals in S-class. A short silence segment is not unexpected since there is a change of source process as well as a drastic shift in the articulatory positions.

About 91% of affricates and unvoiced fricatives lie in L-class. Affricates include the voiced affricate /jh/, which also lies in L-class. Amongst the fricatives, 20% of 'th' lies in S-class, since it sometimes manifests as a burst with a closure interval.

In all, 56.9% of voiced fricatives also lie in L-class, whereas 28.6% lie in H-class and 8.3% go to S-class. The presence of voicing in voiced stops gives rise to a large amplitude BPF signal and when these classes follow a silence or an L-class phone, they go to H-class; 72% of /z/ lies in L-class despite being a voiced fricative. The phone 'dh' sometimes behaves like a stop with a closure and /v/ is realized both as voiced and unvoiced.

The phone labels of the TIMIT database are mapped to the phonetic classes, sonorants and non-sonorants. All vowels, semi-vowels and nasals are assigned to the sonorant class. Others ('h#', 'epi', 'pau') and the closures of



(A) **(B)**

Figure 9. (**a**) Histogram of the temporal deviations of the detected transitions from the TIMIT hand-labelled boundaries and (**b**) detection accuracy in percentage (%) of transitions as a function of temporal tolerance.

Table 2. Relative distribution of each class of phones among the broad five classes. Results on the entire TIMIT data, containing both training and test data.

Segment type	H	L	S	HL	LH
Vowels	91.8	4.8	0.5	2.2	0.7
Semivowels	86.2	9.2	1.3	2.5	0.9
Nasals	82.0	7.2	6.8	3.3	0.8
Unvoiced fricatives	4.2	91.1	2.0	2.2	0.4
Voiced fricatives	28.6	56.9	8.3	5.0	1.2
Voiced stops	48.0	37.3	12.1	1.6	1.1
Unvoiced stops	16.3	70.5	11.3	1.3	0.5
Affricates	6.3	91.0	0.1	1.8	0.7
Others	4.5	13.3	82.1	0.1	0.0
Voiced closures	10.6	8.7	78.4	1.7	0.7
Unvoiced closures	2.1	5.6	92.0	0.2	0.1

Table 3. Distribution of each broad class of phones in the TIMIT database among the five classes.

Phone class	H	L	S	HL	LH
Sonorant	89.8	5.8	1.3	2.4	0.7
Non-sonorant	15.2	75.5	6.2	2.4	0.7
Silence	4.8	10.5	84.2	0.4	0.1
Voiced non-sonorant	34.8	50.6	9.5	3.9	1.2
Unvoiced non-sonorant	8.4	84.2	5.0	1.9	0.5

stops are assigned to the silence class. Non-sonorants include all the phones except the sonorants and silence. The relative distribution as per the phonetic classes, 'sonorant', 'non-sonorant' and 'silence', is shown in table 3. About 89.8% of sonorants lie in H-class. About 75.5% of non-sonorants are in L-class. If we remove voiced fricatives and voiced stops from non-sonorants, then the unvoiced non-sonorants in L-class increase to 84%. This suggests that we need two groups of non-sonorants: 84.2% of 'silence' segments lie in S-class with 10.5% in L-class. Once again, this may arise due to some so called silence phones like 'h#' and 'epi' having a high amplitude.

Table 4. Percentage of onsets of broad phonetic classes detected, as a function of temporal tolerance in the TIMIT database.

Onset of	Type	20	30	40
Sonorants+	L–H, S–H	92.0	94.0	94.7
Unvoiced fricatives/affricates*	H–L, S–L	83.0	85.4	86.5
Stop closures	L–S, H–S	77.2	80.0	81.4
Bursts	S–H, S–L	87.7	88.7	89.1

+Following an unvoiced fricative, unvoiced stop or an affricate.

*Following a sonorant or a silence.

Based on these results, we can broadly state that H-class represents the sonorant class and L-class represents unvoiced non-sonorants, whereas voiced non-sononrants may be found either in H- or L-class.

5.3 *Onset of sonorants and non-sonorants vis-a-vis the type of transition*

The onsets of sonorants and non-sonorants are considered as landmarks [16–18]. It would be of interest to relate the onsets of sonorants and non-sonorants to the detected types of transition. We have excluded /q/ from non-sonorants as in several previous works [17, 22]. Further, within the non-

sonorants, we consider the fricatives and stop bursts separately to detect the onsets. The results are shown in table 4. For a tolerance of 30 ms, 94% of onsets of sonorants occur at L–H or S–H transitions. We have considered sonorants following unvoiced fricatives, unvoiced stops and affricates, since voiced fricatives and voiced stops may lie in H-class (see table 2). The onsets of unvoiced fricatives and affricates occur at H–L and S–L transitions 85.4% of the time within 30 ms. Stop closures are detected as onsets 80% of the time across L–S and H–S transitions. Onsets of stop bursts invariably (88.7%) follow a detected silence segment (S–H, S–L). The results are comparable even for a tolerance of 20 ms. Hence the proposed method also serves the purpose of landmark detection with a good accuracy and temporal resolution.

5.4 *Insertions*

The insertions on the whole TIMIT database are 8.7%. About a third of these insertions occur during the silence, i.e., 'others' and closures of stops. Segments like 'h#' and 'epi' occasionally contain impulse-like noise with significant amplitude resulting in some spurious S–N and N–S transitions. About 24% of the insertions occur during stops. They arise partly due to multiple bursts. A transition is also detected across a low level aspiration interval following a strong burst. While this is a desirable feature of the algorithm, since the aspiration interval is not explicitly marked, such transitions get reported as insertions. During unvoiced fricatives, especially, /f/, the amplitude of the signal varies considerably with intermittent low frequency, large amplitude pulses resulting in a high rate of insertions (about 11%).

5.5 *Comparison with the previous work*

In terms of detecting classes, this work is comparable to manner classification [2, 14] and in terms of detecting onsets, this work is closest to the landmark detection reported in the literature [17, 18].

The present work differs from the previous related works in four important aspects. (a) The temporal features used in this study are different from those proposed in the earlier studies. (b) The proposed algorithm has been tested on the entire TIMIT database, whereas the previous studies have reported results based on a limited test data (16 speakers speaking a total of 80 utterances for the development set and 16 new speakers speaking 48 utterances for the test set taken from the TIMIT database in a study by Liu [18]; 504 utterances from the test set of the TIMIT database in the study by Salomon *et al* [17]). (c) The transitions or landmarks to be detected correspond to different events. Liu [18] defined four landmarks and Salomon *et al* [17] defined three landmarks. (d) The quoted results correspond to a temporal tolerance of 30 ms [18] or 50 ms [17]. Due to these disparities, we can make only a broad qualitative comparison with the previous works. The comparison of our results with the published results of Liu [18] and Salomon *et al* [17] is summarized in table 5.

Table 5. Performance comparison of various algorithms with respect to temporal accuracy of detection.

Method	Database	Results
AGR (our method)	Whole TIMIT training and test set	93.6% and 96.7% within 20 and 30 ms
Liu [18]	48 utterances from TIMIT database	83% and 88% within 20 and 30 ms
Salomon *et al* [17]	504 utterances from test set of TIMIT database	74.8% within 50 ms

Salomon *et al* [17] tested their method on the manner classes of sonorant, fricative, stop and silence. The average accuracy using 39-dimension MFCCs or 12 parameters derived from four temporal features was reported as 70% for a tolerance of 50 ms, whereas with the combined features, it increased to 74.8%. Compared to these results, the accuracies of the proposed method are 89.8%, 84.2% and 84.2% for sonorants, unvoiced non-sonorants and silence classes, respectively, within 20 ms tolerance, when tested on the entire TIMIT database (see table 3).

In Liu's [18] study, of the total number of landmarks, 83% and 88% were within 20 and 30 ms of the labelled boundaries, respectively. The classes considered in that study are sonorants, fricatives and bursts. These results may be compared to the temporal accuracy of detection of the present work (table 4). For a temporal tolerance of 20 and 30 ms, our temporal accuracy is 93.6% and 96.7%, respectively.

6. Robustness in the presence of noise

We evaluate our algorithm on noisy speech generated by adding different kinds of noise to the test set of TIMIT database at various signal to noise ratios (SNRs). The following noises are used for evaluating our algorithm for detection of transitions.

- Schroeder noise [24]: It is a localized white noise. As the energy level in a speech utterance varies widely with time, the clean speech is corrupted with Schroeder noise so that samplewise SNR is constant in the noisy speech. We use the model as devised in [24], where the noisy speech signal is generated by the formula $y[n] = s[n](1 + \epsilon\eta[n])$, where $s[n]$ is the speech signal, ϵ is the factor determining the noise energy, which

changes with the desired SNR, and $\eta[n]$ is the randomly chosen $+1$ or -1 with equal probability.

- White noise: It is generated from a zero-mean normal distribution, with the standard deviation being determined by the SNR desired.
- Babble noise: It is taken from the Noisex-92 database [25] and scaled appropriately to generate noisy speech with the desired SNR.

Figure 10 shows the percentage of the total number of transitions (with respect to 38,198 transitions detected in the case of clean speech) detected by our algorithm for the three types of noisy speech with SNR varying from 0 to 30 dB. It is seen that at a low SNR of 10 dB, our algorithm detects 8.6% for speech with babble noise, 39.6% in the case of white noise and 96.4% in the case of Schroeder noise.

Insertions at an SNR of 10 dB are 0.23% for white, 0.49% for babble and 9.36% for Schroeder noise. Since the number of transitions detected is low for white and babble noise, it is imperative that the % number of insertions is less. It is observed that transitions between silence and non-silence segments (S–N and N–S) are missed for white and babble noises at low SNRs since the silence regions are corrupted by noise. In the case of Schroeder noise, as the samplewise SNR is constant, silence segments are not corrupted with high noise and the corresponding transitions are preserved, detecting 33,026 (86.46%) transitions even at 0-dB SNR.

Figure 11 shows the precision of detection within a temporal tolerance of 20 ms for the three noises as a function of SNR. It is seen that among the detected transitions, even at an SNR of 5 dB, precision above 91% is achieved. It is seen that temporal accuracy does not change much with variation in SNR for Schroeder noise, since the low energy in the silence segments is preserved due to uniform local SNR and hence the S–N and N–S transitions remain intact. For white and babble noise, energy in the silence segments increases with increase in noise energy (or decrease in SNR) and hence the SI value is low even for silence segments at low SNR, which leads to missing S–N, N–S transitions.

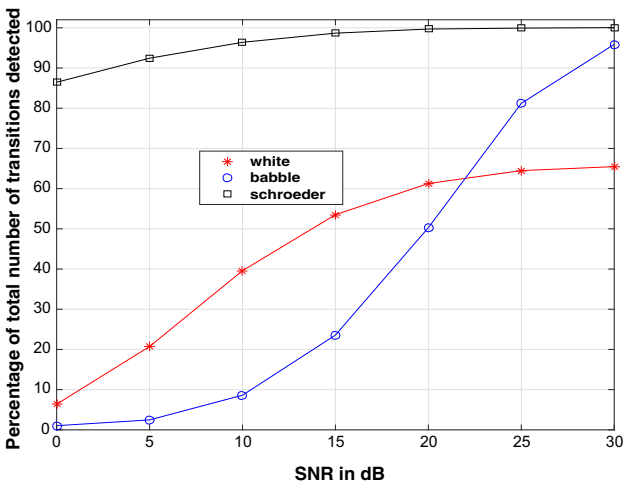


Figure 10. Percentage of total number of transitions detected on TIMIT test set as a function of input SNR.

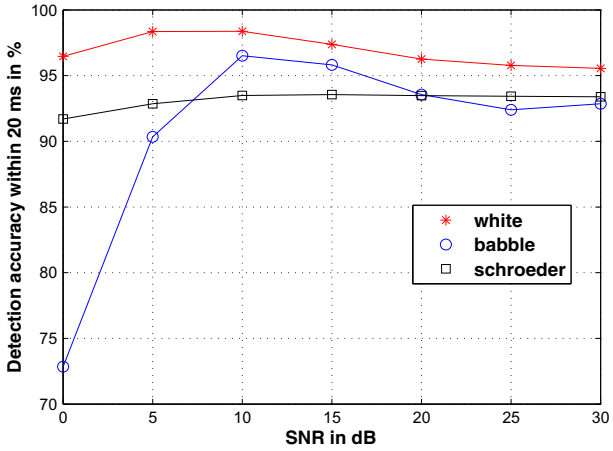


Figure 11. Precision of detected transitions (for a temporal tolerance of 20 ms) as a function of SNR.

Figure 12 shows the percentage of onsets of sonorants and fricatives detected (recall) within a tolerance of 20 ms. Since babble noise has significant low frequency energy, accuracy of detection of onsets of sonorants and fricatives suffers at low SNRs. For white and babble noises, even though we miss S–N and N–S transitions, H–L and L–H transitions are preserved at low SNRs, which result in relatively high detection rate for onsets of sonorants and fricatives.

6.1 *Results on NTIMIT database*

We also evaluate our algorithm on NTIMIT [29] database, which was created by transmitting all the TIMIT recordings through a telephone handset and over various channels of a telephone network and redigitizing them. As compared with TIMIT database, only 62.7% of the transitions are detected on NTIMIT database. Table 6 shows the relative distribution of each broad class of phones in the NTIMIT database as per the phonetic classes 'sonorant', 'non-sonorant' and 'silence'. About 62.9% of sonorants lie in H-class. About 50.5% of non-sonorants are in L-class. If we remove voiced fricatives and voiced stops from non-sonorants, then the unvoiced non-sonorants in L-class increase to 53%; 38.3% of 'silence' segments lie in S-class with 36.8% in L-class.

The relations of onsets of sonorants and non-sonorants to the detected types of transition for NTIMIT database are shown in table 7. For a tolerance of 30 ms, 66% of onsets of sonorants occur at L–H or S–H transitions.

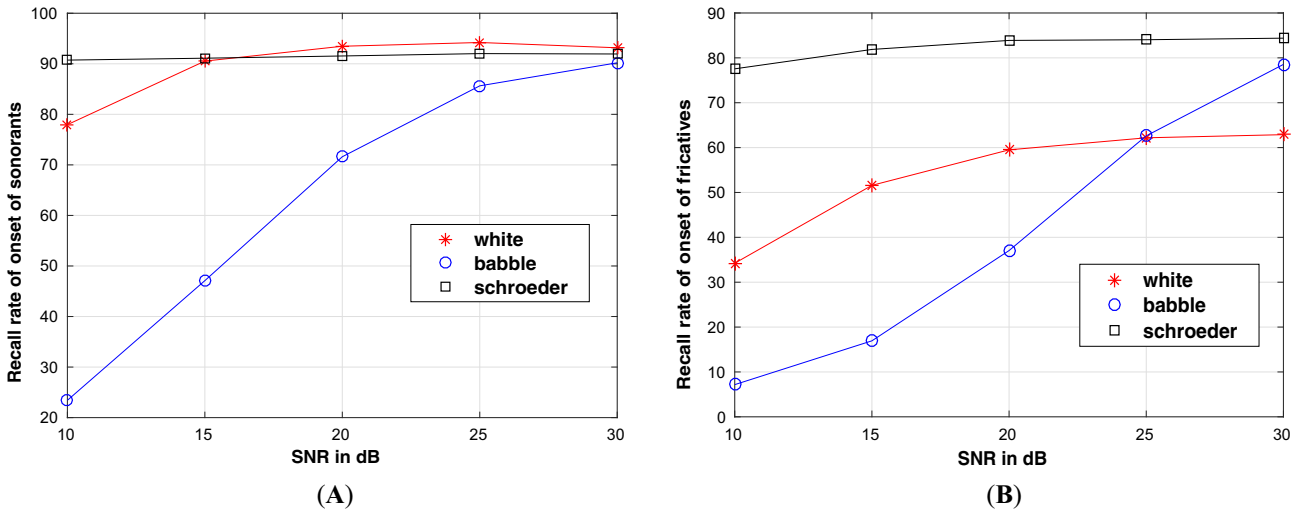


Figure 12. Percentage of onsets of (**a**) sonorants and (**b**) fricatives detected (recall) within a tolerance of 20 ms as a function of SNR in dB for white, babble and Schroeder noises.

Table 6. Distribution of each broad class of phones in the NTIMIT database among the five classes.

Phone class	H	L	S	HL	LH
Sonorant	62.89	15.29	5.46	12.21	4.16
Non-sonorant	19.77	50.45	12.12	13.54	4.11
Silence	18.01	36.86	38.31	4.96	1.86
Voiced non-sonorant	25.78	42.32	14.77	12.54	4.59
Unvoiced non-sonorant	17.68	53.29	11.20	13.88	3.95

Table 7. Percentage of onsets of broad phonetic classes detected, as a function of temporal tolerance in the NTIMIT database.

Onset of	Type	20	30	40
Sonorants+	L–H, S–H	63.49	66.38	67.43
Unvoiced fricatives/ affricates*	H–L, S–L	36.42	39.17	41.23
Stop closures	L–S, H–S	26.52	28.37	29.06
Bursts	S–H, S–L	30.48	32.32	33.10

+Following an unvoiced fricative, unvoiced stop or an affricate.

*Following a sonorant or a silence.

It is observed that the results on NTIMIT database are poorer than those for the TIMIT database due to the bandpass filtering and channel noise in the NTIMIT database.

7. Conclusion

For the DFs and PFs to be complementary to the statistical approach, we believe that an acoustic-phonetics knowledge-based approach needs to be pursued. In our understanding, the highlight of such an approach is that it does not require a huge amount of training data and a small development set is considered sufficient. In this paper, we have proposed a knowledge-based approach to the problem of detecting transitions in both clean and noisy speech signal. Further, several studies have pointed out the robustness of temporal features in speech perception [17, 26]. In the proposed method, using only four simple measures, we have been able to demonstrate that landmarks like the onsets of sonorants (L–H, S–H), unvoiced sonorants (H–L, S–L), closures of stops and stop bursts can be detected with a high accuracy ($> 85\%$) and with a good temporal resolution (20 ms). These results are as good or better than those from state-of-the-art methods, which make use of high-dimensional acoustic features and sophisticated classifiers. Although a number of techniques exist for segmentation, alternate approaches are to be explored, since they may complement one another and offer robustness. There are no specific, fixed thresholds in our method. The thresholds dynamically adapt to the local statistics of the peaks and valleys within each analysis frame and hence are able to be generalized well and detect the transitions between different classes of phones.

7.1 *Future work*

The algorithm is based on the knowledge of the relative distribution of the amplitudes of the different broad classes of phones in specific frequency bands, and thus can be extended to other applications. During the course of this investigation, we have made some observations, which are noted here for future work. (a) We could inquire how OFE/OLE measures perform instead of the abrupt energy change measures used in the literature for the detection of landmarks [18], manner classes [17], bursts [27] and vowel

onset points [28]. (b) Our preliminary investigation shows that OFE and OLE measures computed on a speech signal, instead of bandpass signal, are useful to identify certain transitions within vocalic segments. Also, OFE and OLE may be computed on subband signals. Such an analysis on a high frequency band could detect frication within the unvoiced signal. (c) The number of extrema in a speech signal relative to the number of extrema in the corresponding bandpass signal is a useful parameter for distinguishing between voiced and unvoiced segments. (d) We have observed that bursts most often lie at the end of a silence or L-class. This narrows down the search interval for detecting the bursts. These preliminary observations need to be formalized and tested in a future work.

References

[1] Fant G 2003 *Speech sounds and features*. Cambridge, MA: The MIT Press (Chapter 2)

[2] Hasegawa J M, Baker J, Borys S, Chen K, Coogan E, Greenberg S, Juneja A, Kirchhoff K, Livescu K, Mohan S, Muller J, Sonmez K and Wang T 2005 *Landmark-based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop*

[3] SaiJayram A K V, Ramasubramanian V and Sreenivas T V 2002 Robust parameters for automatic segmentation of speech. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. I-513–I-516

[4] van Hemert J P 1991 Automatic segmentation of speech. *IEEE Trans. Signal Process.* 39: 1008–1012

[5] Muralishankar R, Srikanth R and Ramakrishnan A G 2003 Subspace and hypothesis based effective segmentation of co-articulated basic-units for concatenative speech synthesis. In: *Proceedings of IEEE TENCON*, October 15–17, Bangalore, vol. 1, pp. 388–392

[6] Obrecht R A 1986 Automatic segmentation of continuous speech signals. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2275–2278

[7] Svendsen T and Soong F K 1987 On the automatic segmentation of speech signals. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 77–80

[8] Sarkar A and Sreenivas T V 2005 Automatic speech segmentation using average level crossing rate information. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 397–400

[9] Ananthakrishnan G, Ranjani H G and Ramakrishnan A G 2006 Language independent automated segmentation of speech using Bach scale filter-banks. In: *Proceedings of the IV International Conference on Intelligent Sensing and Information Processing*, pp. 115–120

[10] Jakobson R, Fant G and Halle M 1952 *Preliminaries to speech analysis: the distinctive features and their correlates*. Cambridge, MA: The MIT Press

[11] Chomsky N and Halle M 1968 *The sound pattern of English*. Cambridge, MA: The MIT Pres

[12] King S and Taylor P 2000 Detection of phonological features in continuous speech using neural networks. *Comput. Speech Lang.* 14: 333–353

[13] Frankel J, Wester M and King S 2007 Articulatory feature recognition using dynamic Bayesian networks. *Comput. Speech Lang.* 21: 620–640

[14] Juneja A and Espy-Wilson C Y 2002 Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning. In: *Proceedings of the IEEE International Conference on Neural Information Processing*, pp. 726–730

[15] Juneja A and Espy-Wilson C Y 2008 A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition. *J. Acoust. Soc. Am.* 123: 1154–1168

[16] Stevens K N 2002 Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* 111: 1872–1891

[17] Salomon A, Espy-Wilson C Y and Deshmukh O 2004 Detection of speech landmarks: use of temporal information. *J. Acoust. Soc. Am.* 115: 1296–1305

[18] Liu S A 1996 Landmark detection for distinctive feature-based speech recognition. *J. Acoust. Soc. Am.* 100: 3417–3430

[19] Lippmann R P 1997 Speech recognition by machines and humans. *Speech Commun.* 22: 1–15

[20] Mesgarani N, Cheung C, Johnson K and Chang E F 2014 Phonetic feature encoding in human superior temporal gyrus. *Science* 343: 1006–1010

[21] Reddy D R 1966 Phoneme grouping for speech recognition. *J. Acoust. Soc. Am.* 41: 1295–1300

[22] Ananthapadmanabha T V, Prathosh A P and Ramakrishnan A G 2014 Detection of the closure–burst transitions of stops and affricates in continuous speech using the plosion index. *J. Acoust. Soc. Am.* 135: 460–471

[23] Garofolo J S, Lamel L F, Fisher W M, Fiscus J G, Pallett D S and Dahlgrena N L 1993 *DARPA TIMIT acoustic-phonetic continuous speech corpus*. NISTIR Publication No. 4930. Washington, DC: U.S. Department of Commerce

[24] Niyogi P and Sondhi M M 2002 Detecting stop consonants in continuous speech. *J. Acoust. Soc. Am.* 111: 1063–1076

[25] Noisex-92 [Online] Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html

[26] Rosen S 1992 Temporal information in speech: acoustic, auditory, and linguistic aspects. *Philos. Trans. R. Soc. London B: Biol. Sci.* 336: 367–373

[27] Niyogi P and Ramesh P 2003 The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets. *Speech Commun.* 41: 349–367

[28] Prasanna S R M, Reddy B V S and Krishnamoorthy P 2009 Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. *IEEE Trans. Audio Speech Lang. Process.* 17: 556–565

[29] Jankowski C, Kalyanswamy A, Basson S and Spitz J 1990 NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database. In: *Proceedings of the 1990 International Conference on Acoustics, Speech, and Signal Processing, ICASSP-90*, pp. 109–112