Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index

T. V. Ananthapadmanabha
Voice and Speech Systems, Temple Road, Malleshwaram, Bangalore 560003, India

A. P. Prathosh[a] and A. G. Ramakrishnan
Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

(Received 18 October 2012; revised 29 October 2013; accepted 11 November 2013)

Automatic and accurate detection of the closure-burst transition events of stops and affricates serves many applications in speech processing. A temporal measure named the plosion index is proposed to detect such events, which are characterized by an abrupt increase in energy. Using the maxima of the pitch-synchronous normalized cross correlation as an additional temporal feature, a rule-based algorithm is designed that aims at selecting only those events associated with the closure-burst transitions of stops and affricates. The performance of the algorithm, characterized by receiver operating characteristic curves and temporal accuracy, is evaluated using the labeled closure-burst transitions of stops and affricates of the entire TIMIT test and training databases. The robustness of the algorithm is studied with respect to global white and babble noise as well as local noise using the TIMIT test set and on telephone quality speech using the NTIMIT test set. For these experiments, the proposed algorithm, which does not require explicit statistical training and is based on two one-dimensional temporal measures, gives a performance comparable to or better than the state-of-the-art methods. In addition, to test the scalability, the algorithm is applied on the Buckeye conversational speech corpus and databases of two Indian languages.
 [http://dx.doi.org/10.1121/1.4836055]

PACS number(s): 43.72.Ar, 43.72.Ne, 43.70.Kv [CYE] Pages: 460–471

I. INTRODUCTION

Stops form an important class of speech sounds. During the production of stops,[1] acoustic pressure is built up behind a closure at a place within the vocal tract, resulting in a silent interval or a low level acoustic signal, with or without voicing. When the pressure is released suddenly, it introduces a relatively high energy burst or transient in the acoustic signal, spanning a short interval. The production of an affricate is also similar to that of a stop consonant.[1] The instant in the acoustic signal corresponding to the sudden release is referred to as the "burst-onset"[2] or the closure-burst boundary or the closure-burst transition (CBT). The problem of automatic detection of the CBTs of stops and affricates from a continuous speech signal is recognized as important in several studies.[2–6] In the remaining part of this section, we briefly discuss the problem as relevant to (i) automatic speech recognition (ASR) and (ii) acoustic-phonetics studies. Subsequently, we review the methods proposed in the literature for detection of the CBTs.

Approaches to ASR may be classified broadly into two classes. The ones based on statistical models primarily employ hidden Markov models (HMMs) and a generic acoustic feature such as Mel-frequency-cepstral-coefficients (MFCCs) common to all the phones.[7,8] Alternative approaches are based on initially deriving the phonetic-feature-specific information from the speech signal, followed by the identification of phones.[9–12] A landmark-based ASR system is an example of the latter approach where "events" in the speech signal with rapid temporal and spectral changes, called the landmarks, are extracted in the initial stage. The subsequent step is to analyze the speech signal only around the landmarks to derive acoustic information for the purpose of classification of phones.[10] Automatic detection of the CBTs is of relevance to both types of ASRs; it has been shown that the performance of an HMM-based ASR system can be enhanced by incorporating the information of the CBTs along with the MFCCs.[4] The detection of the CBTs plays a role in identifying the burst-onset landmark, a manner class called "stops" or the distinctive feature called "interrupted" in other ASR systems.[2,9,13,14]

In acoustic-phonetics studies, detection of the CBTs has been shown to help in the identification of the appropriate analysis interval for determining the place of articulation of stops.[15] Further, voice onset time (VOT) is noted to be a significant attribute useful for the discrimination of voiced from unvoiced stops.[16] VOT also aids in accent identification, clinical applications, etc.[17,18] State-of-the-art methods proposed for automatic measurement of VOT require an *a priori* knowledge of the CBTs.[19] Thus, automatic detection of the CBTs caters to this need.

In the literature, the methods proposed to detect burst-onset landmarks, stop-bursts, manner class "stop," and stop consonants rely on the temporal and/or spectral characteristics of the speech signal around the CBTs for feature extraction and the labeled CBTs as the ground truth for validation.[2–6,13,20] We briefly review all these methods by noting the acoustic feature and the classification strategy used. For detecting the stop-bursts, Bitar[20] used the degree of abruptness in energy difference between two appropriately located frames as an acoustic measure, which was

[a]Author to whom correspondence should be addressed. Electronic mail: prathoshap@ee.iisc.ernet.in

originally proposed by Espy-Wilson,[21] in a fuzzy rule-based classifier. Liu[2] used the rate-of-rise of energy (RoR) across appropriately located frames in six specific frequency bands and a threshold-based logic to detect stop-burst landmarks. King and Taylor[13] have used short-time energy and MFCCs along with their derivatives (39 parameters) as the feature vector and trained neural networks to identify all the sound-pattern-English (SPE) features proposed by Chomsky and Halle.[22] Hou *et al.*[14] utilized a range of temporal and spectral acoustic features (energy ratios, zero-crossings, linear prediction coefficients, etc.) as inputs to classifiers such as multi-layer perceptron and Bayesian classifier to extract all the SPE features. These features were subsequently used to detect stop consonants. Lin and Wang[4] have used a two-dimensional cepstrum as the feature vector (56 dimensional) and a random forest (RF) classifier for detecting burst-onset landmarks. Niyogi *et al.*[5] used three energy measures (log of total energy, log of energy above 3 kHz, and Wiener entropy) as a feature vector in a support vector machine (SVM) classifier to detect stop consonants. Niyogi and Sondhi[3] used the same feature vector with an optimal adaptive filter consisting of 33 parameters to detect stop consonants. Salomon *et al.*[23] have used four temporal features to detect acoustic landmarks and used them in a HMM classifier to identify several manner classes including "stop." Jayan and Pandey[6] used a Gaussian mixture model (GMM) of smoothed log magnitude spectrum (256 coefficients) and the rate of change of the components of the GMM to detect stop consonants.

Generally, these methods are validated against a labeled database with marked closure-burst boundaries, such as the TIMIT database. A common criterion is that if the detection is within a certain temporal tolerance (20–40 ms) of the labeled closure-burst boundary of a stop/affricate, then the method is deemed to have detected the burst-onset landmark or a stop/affricate consonant or the manner class "stop." The performance is characterized in terms of false acceptance and rejection rates and the associated receiver operating characteristic (ROC) curve by some methods and in terms of deletion and insertion rates by others. Also, the statistics of the temporal deviation of the detected CBTs from the labeled boundary are considered for characterizing the accuracy of detection.

In this paper, we propose two new temporal features and a rule-based classifier for the detection of the CBTs and find out if it can result in a performance comparable to the best reported in the literature for similar experimental conditions. Also, we study the robustness and scalability of the proposed method. Formally, the objectives of the paper are: (i) To propose and use a one-dimensional temporal measure to detect events with abrupt increase in energy such as the CBTs of stops. (ii) To design a rule-based algorithm (without the need for statistical training) to select a subset of these events belonging to stops and affricates using a second temporal feature. (iii) To validate the algorithm on the entire TIMIT training and test databases with criteria similar to those used in the previous studies[3,4] and to characterize the performance by the ROC curves. (iv) To test the robustness of the algorithm in the presence of two types of additive noise, viz., stationary white noise and non-stationary babble noise and also on telephone quality speech. (v) To test the scalability of the algorithm on the Buckeye corpus comprising conversational speech and a database of two Indian languages.

II. PROPOSED TEMPORAL FEATURES

Research into finding new temporal measures and their application in speech processing is recognized as an important area.[23] It has been suggested that temporal measures are relatively robust and that human perception also makes use of temporal cues.[24]

In this section, we propose a temporal measure named the plosion index (PI) to detect events with abrupt change in energy. Sometimes such a change in energy (as seen around the CBTs) is also observed in events like strong voiced onsets preceded by a low-level signal. In Sec. II B, one more temporal measure, namely, the maximum normalized cross-correlation (MNCC), is proposed to discriminate a CBT from a voiced onset.

A. The Plosion index

Intuitively, for a signal with a transient characterized by a significant change in local energy, the ratio of the peak amplitude in the transient to the average of absolute values over an appropriate interval excluding the immediate neighborhood of the peak amplitude may be expected to be high. In order to capture the intrinsic nature of a transient-like signal preceded by a low-level signal, as in a CBT of a stop, we define a temporal measure named the PI at an instant of interest, n_0, for a signal $s[n]$ as

$$\mathrm{PI}(n_0, m_1, m_2) = \frac{|s(n_0)|}{s_{\mathrm{avg}}(m_1, m_2)}, \tag{1}$$

where

$$s_{\mathrm{avg}}(m_1, m_2) = \frac{\sum_{i=n_0-(m_1+m_2)}^{i=n_0-(m_1+1)} |s(i)|}{m_2} \tag{2}$$

is the average of the absolute amplitudes of m_2 samples, offset from n_0, by m_1 samples. Being a ratio, the PI is a dimensionless measure, independent of the recording level. The definition of the PI may remind a reader of the measure crest factor or peak-to-average ratio existing in the literature. However, the crest factor is an index that characterizes an entire signal, where both the peak and the average values are obtained from the complete signal. In contrast, the PI is an instantaneous measure and a function of two parameters, m_1 and m_2.

In the context of the detection of the CBTs of stops/affricates from a continuous speech signal, an appropriate choice needs to be made for m_1 and m_2. Since certain low-level noise-like signal components, called the pre-frication, are usually present preceding the instant of maximum amplitude within an unvoiced stop-burst,[3] we choose m_1 as the number of samples corresponding to 6 ms (see Sec. V A for a justification for this choice). This excludes the samples of pre-frication (which are of

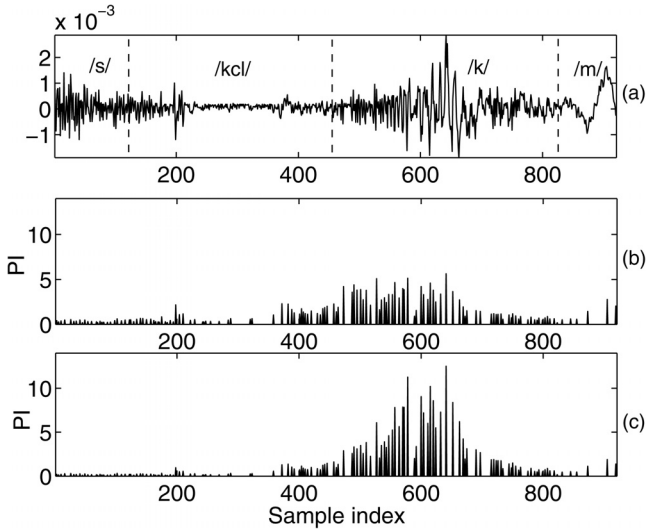

FIG. 1. Illustration of the need for offset m_1 in reducing the effect of pre-frication on the PI. (a) A segment of speech with a fricative followed by a stop, (b) the corresponding PI values computed without the offset m_1, (c) the corresponding PI values with the offset m_1.

amplitude higher than those in the stop-closure region) while computing s_{avg}. Based on the statistics of the minimum closure duration for stops,[25] m_2 is chosen as the number of samples corresponding to 16 ms. Throughout this work, m_1 and m_2 are kept fixed corresponding to these chosen values.

Figure 1 illustrates the role of m_1 in enhancing the value of the PI, through an example of a stop (/k/), shown in Fig. 1(a), occurring at a consonant cluster (/s/-/k/). Figures 1(b) and 1(c) show the corresponding PI values computed (at the peaks between successive zero-crossings) without and with the use of the offset m_1 while computing the s_{avg}, respectively. The presence of a strong pre-frication may be observed resulting in lower values of the PI in Fig. 1(b). However, the PI values almost increase twofold when the offset m_1 is used.

1. Pre-processing for the computation of the PI

The change in energy around the CBT is low for a voiced stop with a weak release preceded by a relatively strong pre-voicing component. Figure 2(a) shows an example of such a case. At the instant of release n_0, the PI

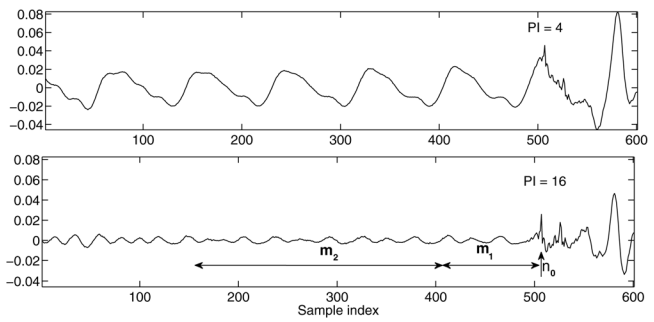

FIG. 2. Illustration of the utility of the high-pass filtering for reliable detection of the CBTs of voiced stops. (a) A segment of a voiced stop with a weak release, (b) the corresponding segment after high-pass filtering. It may be seen that there is an increase in the value of the PI by a factor of 4 after high-pass filtering.

computed on this signal is about 4. In order to attenuate the pre-voicing component preceding such a CBT and thereby enhance the amplitude contrast, the speech signal is high-pass filtered with a cut-off frequency of 400 Hz.[2] However, this does not significantly influence the abrupt change in the amplitude around the CBTs of unvoiced stops and affricates. Figure 2(b) shows the high-pass filtered signal corresponding to the same segment shown in Fig. 2(a). Now, at the instant of release, n_0, despite a decrease in the peak value, the PI increases to about 16. The intervals corresponding to m_1 and m_2 are also marked in Fig. 2(b).

Further, the peak amplitude of a transient signal is influenced by its phase characteristics. For example, consider a heavily damped sinusoid resembling a transient. Its absolute maximum amplitude depends on the initial phase angle and is the lowest for 0° and the highest for 90°. However, both the maximum amplitude and the location of the maximum in the Hilbert envelope (HE) of such a damped sinusoid are independent of the initial phase angle.[26] Hence, the PI is computed on the HE of the high-pass filtered speech signal. A Hilbert transform is computed in the time domain by convolving the speech signal with a 32-point finite impulse response of the Hilbert transformer.

Figure 3 illustrates the PI values computed at every sample for a segment of a speech signal consisting of a fricative followed by a stop followed by a vowel. The PI is high (>600) around the CBT (126 ms) and low elsewhere. It may be observed from Fig. 3 that there is an interval (marked by dashed vertical lines) around the CBT within which the PI is high. However, since the CBT is an instant, it is desirable to have only one candidate representing a transient interval. To reduce the interval measure to an instantaneous measure, we propose a merger rule, which is explained as a part of the detection algorithm in Sec. III.

2. Discriminability of the PI

In order to test the discriminability of the PI for detecting the CBTs against other events, the normalized histograms of representative PIs for (i) stops/affricates and (ii) other phones (vowels, semi-vowels, glides, nasals, and fricatives) from the entire labeled TIMIT database are computed

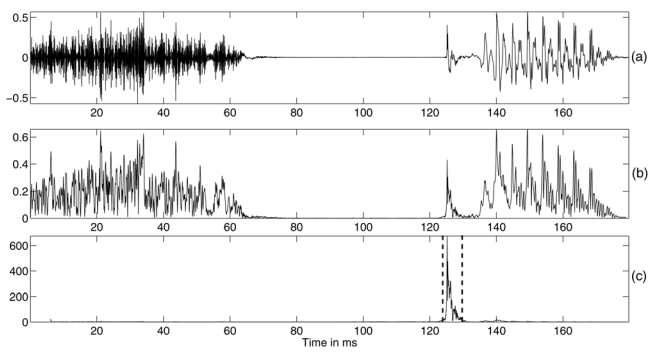

FIG. 3. Illustration of the ability of the PI to capture events with abrupt increase in energy. (a) A segment of a speech signal with a fricative followed by an unvoiced stop followed by a vowel, (b) the Hilbert envelope of the high-pass filtered speech, (c) the PI corresponding to the signal shown in (b), computed with m_1 and m_2 corresponding to the time intervals of 6 and 16 ms, respectively.

and shown in Fig. 4. The maximum value of the PI within a labeled segment is taken to be the representative PI for that phone. A total of 19 866 tokens of stops and affricates and 89 552 tokens of other phones are considered. Although a large separation of the two classes is seen, there is a considerable overlap. For example, if one chooses a threshold of 8 for the PI to separate the classes, 93% of the CBTs of stops and affricates would be detected correctly. However, 33% of other phones would be incorrectly classified as the CBTs. This is because the PI detects abrupt onset corresponding to any sound preceded by a low-level signal. It is observed that most of these arise from the strong onsets of voiced sounds which are to be discriminated from the CBTs of stops and affricates. For this purpose, we define another temporal measure, called the MNCC.

B. The MNCC

It is well known that normalized cross-correlation (NCC) quantifies the degree of similarity as a function of the lag between two finite energy signals, irrespective of their energies.[27,28] In this work, the maximum value of the NCC (MNCC) is used as the second temporal feature. By definition, the MNCC is a scalar and lies between 0 and 1.

In the literature, NCC is generally computed between the segments of a speech signal, about 20–40 ms in duration, for the purpose of pitch estimation and voiced-unvoiced decision.[27] However, in the present work, we compute NCC between the segments of speech over two successive inter-epoch intervals. This assumes that the epochal information is available. Epochs are extracted using an algorithm developed by the authors using an extended concept of the PI called the dynamic PI, which places epochs at glottal closure instants over voiced regions and at random locations over unvoiced regions.[29] The value of the MNCC, computed between two successive inter-epoch intervals, is assigned to all the samples over the first inter-epoch interval. Thus, the MNCC plotted for a speech signal appears as a staircase-like function.

For a speech signal corresponding to a voiced sound, the vocal tract impulse responses for successive pitch periods are highly correlated, resulting in a high value for the MNCC. There is no such high-correlation between two successive segments in the case of unvoiced sounds due to the random excitation, which results in a lower MNCC. Figure 5 shows a speech segment (a stop followed by a vowel, a fricative, and another vowel) with the corresponding values of the PI and the MNCC. The MNCC is low (typically less than 0.6) for the unvoiced regions and high (typically greater than 0.6) for the voiced regions.

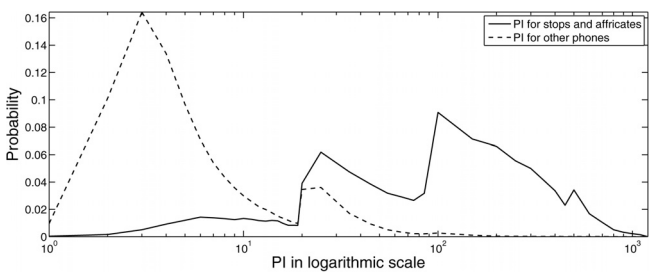

FIG. 4. Normalized histograms of the PI for stops/affricates (solid line) and other phones (dashed line) of the entire TIMIT database. The x-axis is shown in logarithmic scale for clarity. The overlap between the two groups in higher values of the PI is largely due to strong voiced onsets.

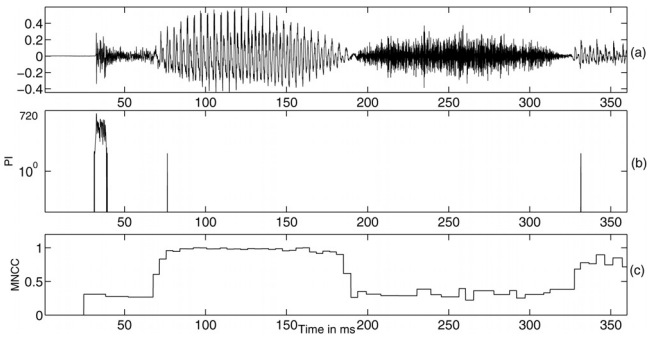

FIG. 5. Illustration of the use of the MNCC to separate the CBTs from the voiced onsets. (a) A speech segment, (b) the corresponding PI values, (c) the corresponding MNCC values, showing MNCC values greater than 0.6 for the voiced segments.

In order to test the discriminability of the MNCC for voiced-unvoiced classification, we compute the normalized histograms (Fig. 6) of the average MNCC within the labeled regions for the two classes of phones from the TIMIT database; class-A consists of a total of 29 150 tokens of unvoiced stops, affricates, and fricatives; class-B consists of a total of 73 016 tokens of vowels, semivowels, glides, and nasals. The histograms show a clear separation of the two classes with a negligible overlap area of less than 5% for both the classes at a threshold of 0.6. Thus, in this work, a threshold of 0.6 is used on the average MNCC computed over three successive inter-epoch intervals to exclude strong voiced onsets being detected as the CBTs. For example, in Fig. 5, although the value of the PI is high at the vowel onsets, they may be identified as not belonging to the CBTs since the average MNCC around those onsets is above 0.6.

Around the CBT of a voiced stop, the MNCC will have a high value due to the presence of quasi-periodicity. Hence, there is a risk of these CBTs being discarded as voiced onsets. However, a singular feature of the voiced stops is a disruption of the periodicity over one or two cycles coinciding with the release, which results in a significant "*high-low-high*" structure in the MNCC around the CBT. Figure 7(a) shows one such instance. Thus, the *high-low-high* structure in the MNCC can be used to detect the CBTs of such voiced stops. Further, the MNCC may be high even in the case of multiple bursts of a single unvoiced stop, and thus may be discarded as a voiced onset. This is because the signals

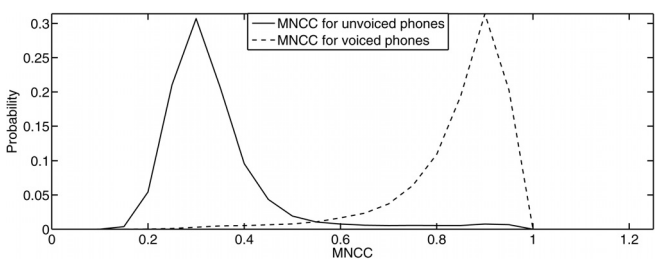

FIG. 6. Normalized histograms of the MNCC values of voiced (dashed line) and unvoiced sounds (solid line) from the entire TIMIT database. The overlap area is about 5% in either case at a threshold of 0.6.

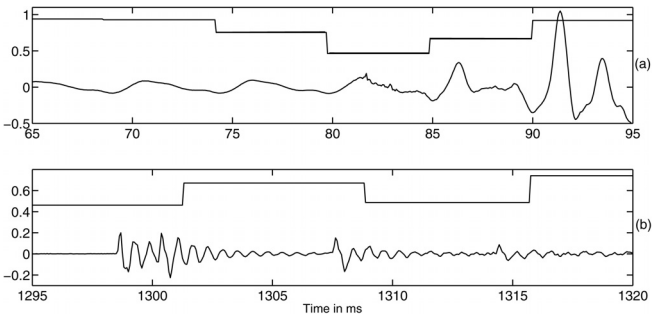

FIG. 7. (a) Illustration of the *high-low-high* structure of the MNCC for a voiced stop with a weak burst. (b) An unvoiced stop with multiple bursts resulting in MNCC > 0.6.

corresponding to the individual bursts may be correlated with one another. Figure 7(b) shows an example of an unvoiced stop with multiple bursts, along with the plot of the corresponding MNCC. This case of multiple bursts is dealt with using the "number of potential candidates," defined in Sec. III.

III. THE CBT DETECTION ALGORITHM

It may be possible to use the representative PIs and the MNCCs as the feature vector and train a classifier to detect the CBTs. Instead, we formulate certain rules to select the CBTs based on the knowledge derived by studying a number of typical cases. In other words, we "learn the rules through examples." The following are the steps in the algorithm illustrated by the flowchart in Fig. 8.

(1) The PI is computed only at the locations of the maxima of HE between every set of successive zero-crossings of the high-pass filtered signal.
(2) The instants at which the PI is greater than a threshold (T_1) are called the potential candidates.
(3) Based on the assumption that no two genuine stop (affricate) releases occur within 20 ms of each other,[4] any two successive potential candidates that are within 20 ms of each other are postulated to belong to one and the same event. In this algorithm, only the very first potential candidate within such an event is retained and is called representative burst candidate (RBC). The number of potential candidates (N_c) within that event is noted. This step is to ensure that there is only one RBC per CBT. This is referred to as the merger rule.
(4) When the average MNCC over three successive inter-epoch intervals immediately following the RBC exceeds a threshold, T_2, three possibilities arise.
   (a) RBC is a CBT of unvoiced stop with multiple bursts: This is confirmed when N_c exceeds a threshold (T_3). This is based on the observation that the number of potential candidates is significantly higher for multiple bursts than for onsets of voiced sounds.

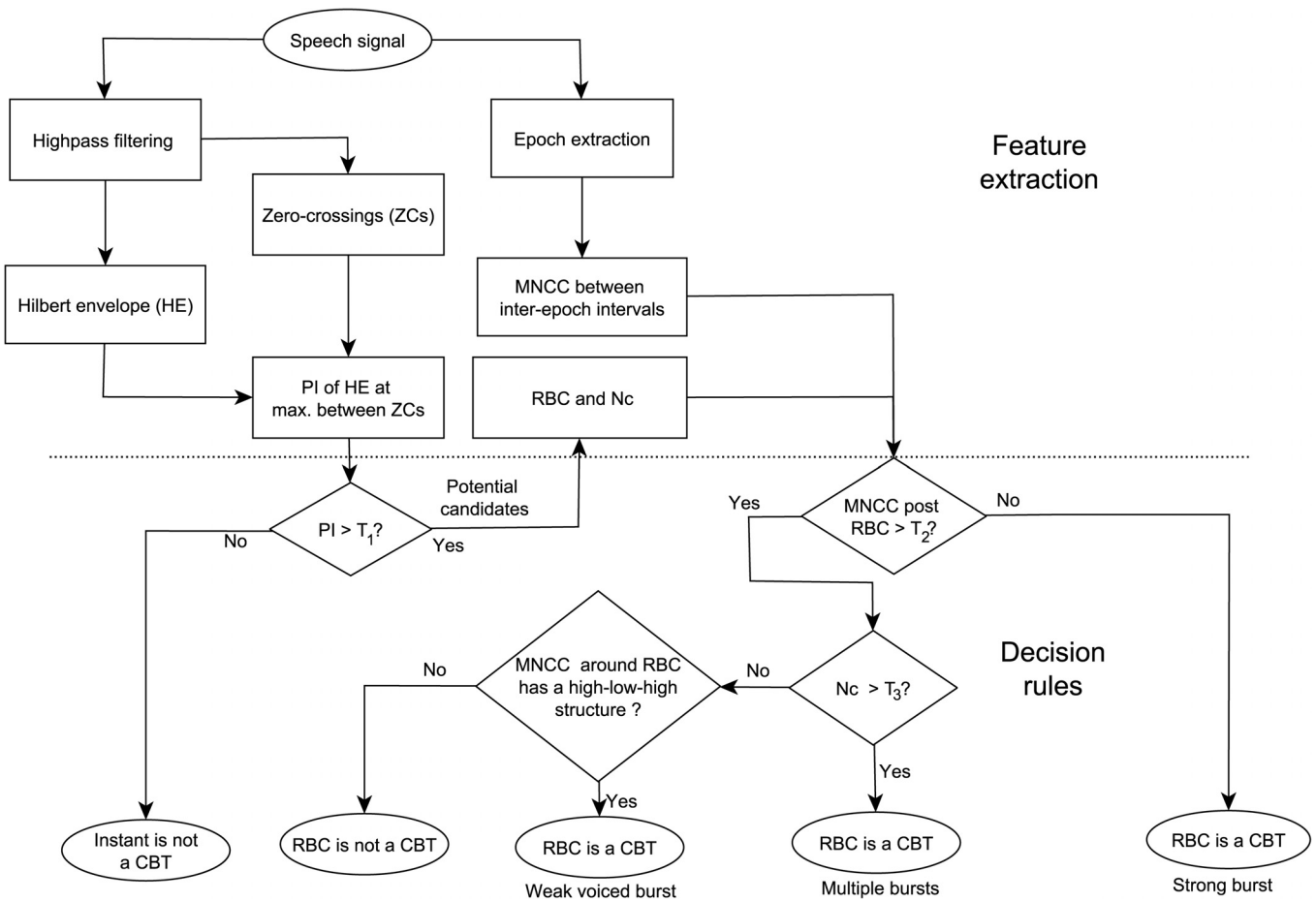

FIG. 8. Flowchart of the proposed APR algorithm for detection of the CBTs.

(b) RBC is a CBT of a voiced stop: This is ascertained by a local *high-low-high* structure in the MNCC around RBC.
(c) RBC is not a CBT (e.g., strong voiced onset): If neither of the above cases is satisfied, then RBC is removed from further consideration.

(5) When the average MNCC over three successive inter-epoch intervals immediately following RBC is less than the threshold, T_2, RBC is declared to be a CBT of a stop/affricate.

The choice for the values of the thresholds is discussed later (Sec. IV B). The output of the proposed algorithm, called the detector output, is a vector with unit impulse at the detected CBTs and zero elsewhere. The proposed algorithm, hereafter called the APR algorithm, not only detects the CBTs, but also the type of burst such as voiced with a weak release, multiple bursts, unvoiced, or voiced bursts with a relatively strong release. This is ascertained by the path traversed in the algorithm to arrive at the detector output. The maximum value of the PI within an event may be used as a measure of the strength of release.

We illustrate, in Fig. 9, a segment of a speech signal (of the utterance "*put the butcher block table in the garage*") along with the detector output obtained using the optimal thresholds. There are correct detections of the CBTs for the stops /p/, /b/, /b/, /t/, /b/, /g/, and the affricate /ch/. There is a detection for the dental fricative /dh/ around 900 ms since dental fricatives occasionally tend to be stop-like.[30] However, around 2200 ms, there is a case of /dh/ without a release, and hence there is no detection. In the region labeled as a closure, /kcl/, around 1600 ms, there is a detection that may be interpreted as incorrect.[4] However, we interpret this detection as belonging to a genuine CBT of an unlabeled /k/ in the consonant cluster (/k/-/t/) occurring at a word boundary. It is recognized that the release may or may not be present for the former stop consonant in a cluster.[31] The labeling could have been /kcl/-/k/-/tcl/-/t/. Incidentally, there is a consonant cluster /t/-/dh/ around 900 ms. However, the burst of /t/ is unreleased in this case and there is no detection.[32]

IV. EVALUATION PROCEDURE AND EXPERIMENTAL DETAILS

Since the goal of the APR algorithm is to detect the CBTs of the stops and affricates, these phones are said to belong to the target class. However, phones such as glottal-stops,[33] flaps,[34] and dental fricatives[30] also may manifest the CBTs. Since the manifestation of the CBT is not consistent for these phones, detector outputs, if any, occurring during these labeled segments are excluded while calculating the performance measures. Previous studies[4] have also followed a similar criterion for these phones. All other phones are included in the rejection class.

A. Performance measures

A labeled database is an absolute necessity for the validation of the CBT detection algorithm. We have adopted the standard performance measures described in the literature[3,4] that are defined below.

(1) Correct detection: A detection is considered to be correct if it lies within ±20 ms of the labeled closure-burst boundary. The tolerance of 20 ms is to account for any possible inaccurate boundary markings present in the databases.[3]
(2) Missed detection: This occurs when there is no detection within ±20 ms of the labeled CBT of a phone from the target class.
(3) False detection: A detection is considered false, if it occurs within the labeled region of a phone from the rejection class.
(4) False acceptance rate (FAR): The number of false detections divided by the total number of phones from the rejection class.
(5) False rejection rate (FRR): The number of missed detections divided by the total number of phones from the target class.
(6) Temporal deviation of detection: The statistics of the deviations of the locations of the detected CBTs from the labeled boundaries, computed only for the correct detections.[4]

B. Choice of thresholds and the ROC curves

As one varies the thresholds for detection, there is a trade-off between FAR and FRR. Based on the risk factors and the application, one may like to make different choices for FAR and FRR and accordingly select the thresholds. Hence a knowledge of the nature of the trade-off between FAR and FRR is required. This is provided by the ROC for any detection problem, where FAR is plotted against FRR. When there is no specific preference for either FAR or FRR, then the performance is specified by the equal error rate (EER), which corresponds to that point in the ROC curve where FAR = FRR. Hence, we characterize the performance of the algorithm by means of the ROC curves and derive the EER from the same. The ROC may be obtained by varying the thresholds for the PI and the MNCC. Since the

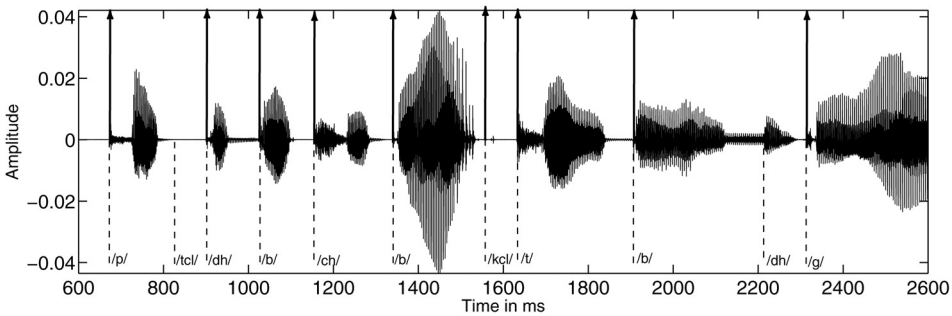

FIG. 9. Illustration of the detected CBTs for a segment of a speech signal of the utterance "*put the butcher block table in the garage*" taken from the TIMIT test set. The detected instants are shown by vertical lines along with the corresponding TIMIT transcriptions at those locations.

distributions of the MNCC values for voiced and unvoiced classes show a clear separation (an overlap area of less than 5% for either classes at a threshold of 0.6), we fix T_2 at 0.6 and vary only the threshold T_1 meant for the PI to generate the ROC curves. T_3 is fixed at 7, based on our empirical observations.

C. Databases and experimental setup

We validate the proposed algorithm on three different labeled databases, which differ significantly in terms of speakers, dialects, recording conditions, speaking styles (read vs conversational), and languages. These diverse conditions contribute to a wide variability in the acoustic characteristics of the speech signal. Also, we consider different types of degradations on one of the databases (TIMIT). This section describes all the experiments conducted.

1. The TIMIT database—Clean speech

To validate the APR algorithm on read speech, we use the TIMIT[35] database, which is labeled at the phone level. It consists of a total of 6300 utterances spoken by 630 speakers belonging to several dialects of North America. The database is divided into the training and test sets of 8 dialects each, comprising 4620 and 1680 utterances, respectively. The APR algorithm has been validated on the entire TIMIT training and test databases independently. In TIMIT, the closure-burst boundaries are marked explicitly for all stops and affricates, which are taken as the ground truth for validation.

2. The TIMIT database with white and babble noise—Global SNR

To study the noise robustness of the APR algorithm, we test it on the entire TIMIT test set with two types of additive noise, stationary white noise, and realistic, non-stationary babble noise. White noise is generated using a zero mean Gaussian distribution whose variance is set in accordance with the desired global SNR. Samples of babble noise are taken from the Noisex-92 database[36] and appropriately scaled to obtain the desired global SNR. Although TIMIT utterances used in the test set have a mean SNR of 39.5 dB,[3] in our calculations we have assumed the speech to be clean. Thus, the actual SNRs are slightly lower than the SNRs of 30, 20, and 10 dB reported in this study.

3. The TIMIT database with Schroeder noise—Local SNR

The global SNR is predominantly determined by the strong voiced segments. Therefore, the local SNR around the CBTs would be much lower and not directly predictable. In order to study the performance of the APR algorithm at specific local SNRs around the CBTs, we have adopted the Schroeder noise model and the procedure given by Niyogi and Sondhi[3] for generating the noisy speech of a desired local SNR. According to this model, the noisy speech signal $y(n)$ is generated at every sample n using the formula $y(n) = s(n)[1 + \varepsilon\eta(n)]$, where $s(n)$ is the clean speech signal, $\eta(n)$ is the binary valued (-1 and 1) noise sample, and ε is the parameter determined by the specified local SNR. Three cases of local SNRs, namely, 20, 10, and 0 dB are used in this study. Only the TIMIT test set is considered in order to compare the results of the APR algorithm with the published results.

4. The NTIMIT database—Telephone quality

To study the performance against channel degradation, we employ the NTIMIT test database,[37] which is the telephone quality version of the TIMIT database. The utterances in NTIMIT differ from those in TIMIT in two important respects, namely, a reduction of bandwidth from 0–8000 Hz to 300–3400 Hz and a degradation in SNR from 39.5 to 26.8 dB.[3]

5. The Buckeye corpus—Conversational speech

To test the scalability of the algorithm on conversational speech, we consider the Buckeye corpus[38] consisting of several hours of recordings of spontaneous American English speech of 40 speakers from central Ohio. Informal conversations were elicited by an interviewer in a seminar room with the speaker allowed to move freely. The corpus is phonetically labeled using a two-stage labeling process involving forced alignment and manual correction. The corpus is available in the public domain.[39]

In this corpus, the entire interval from the closure to the onset of the next sound (e.g., vowel onset), including the burst, has been assigned the label of the stop/affricate consonant. Hence, we modify the definition of the correct detection: A detection is defined to be correct if it lies anywhere within the entire region labeled as stop/affricate. Since there are no separate labels for closure and burst intervals, temporal deviations of detection cannot be measured. A randomly selected subset of the speech data from all the 40 speakers has been considered. Since the duration of each speech file is very long (on the order of ten minutes) and consists of several utterances with intermittent long pauses, any detection following a labeled long silence is ignored. The number of stops and affricates in the selected subset is 1972 and the number of phones from the rejection class is 11 307.

6. The MILE database—Dravidian languages

To further test the scalability of the algorithm, we consider the MILE database comprising about 2000 utterances of phonetically rich sentences of two Dravidian languages, Kannada and Tamil, spoken by male speakers (one for each language) annotated manually at the phone level. These were recorded in a studio environment for the purpose of the development of a text-to-speech synthesis system[40] in the MILE lab, Indian Institute of Science. Here, the CBTs are not explicitly labeled. Hence, the performance evaluation is the same as that used for the Buckeye corpus. The target class includes all the stops and affricates of the corresponding languages. The number of tokens in the target and rejection classes is 2352 and 11 700 for Kannada and 2359 and 13 635 for Tamil databases, respectively.

V. EXPERIMENTAL RESULTS

In this section, we present the results of the experiments in the same order as described in Sec. IV. The results are

TABLE I. Summary of all the experiments. APR algorithm is compared with three state-of-the-art algorithms on the TIMIT database without and with various kinds of additive noise.

Dataset	Details	Liu (deletion%)	N&S (EER)	L&W (EER)	APR (EER)
TIMIT test	∼7 k stops, ∼50 k others; 160 speakers	19	15 (subset used)	7.3	7.7
TIMIT training	∼21 k stops, 130 k others; 470 speakers	—	—	—	7.9
TIMIT noise	White; global SNR = 30,20,10 dB	—	20,46,67	—	9.5,15,28.5
TIMIT noise	Babble; global SNR = 30,20,10 dB	—	—	—	9,13.5,26.5
TIMIT noise	White Schroeder; local SNR = 20,10,0 dB	—	21–22	—	7.8,8.1,10.8
NTIMIT	Telephone quality	22	35	—	18.5
Buckeye corpus	Conversational speech; 40 speakers	—	—	—	19
MILE corpus	Kannada and Tamil; 2 speakers	—	—	—	16,12

compared with some state-of-the-art algorithms and summarized in Table I. An analysis of errors is also presented with reference to the TIMIT database.

A. The TIMIT database—Clean speech

1. The ROC curves and EERs

Figure 10 depicts the ROC curves for the TIMIT training and test databases. It is noteworthy that there is very little difference in the ROC for the test and the training databases with EERs (EER-APR) of about 7.7% for the test and 7.9% for the training databases, respectively. Incidentally, EER is achieved around a threshold for the PI of 8 which corresponds to about 9 dB. In the literature, an energy difference of 9 dB has been used for the detection of stop bursts.[2,41] In our study, it has been noted that if the PI alone is used for the CBT detection without the MNCC and the associated rules, EER increases to 12%.

In general, the CBTs of unvoiced stops are detected better than those of voiced stops. This may be because the burst release is weaker in the case of voiced stops. Specifically, the detection accuracy is the highest for /p/ (around 96%) and the lowest for /g/ (around 86%). For affricates, it is about 87%.

2. Temporal deviation

To quantify the accuracy of the detected locations of the CBTs with reference to the labeled boundaries, we use the temporal deviation of detection. The deviation δ_i associated with each correct detection is defined as $\delta_i = t_i - t_i^*$, where t_i is the detected location and t_i^* is the labeled closure-burst boundary in TIMIT. Figure 11 shows the probability density function (normalized histogram) of δ and the cumulative distribution function of the absolute value of δ for the entire TIMIT test and training databases (combined together). The percentages of the detected CBTs are 64%, 84%, 97%, and 100% for deviations of 5, 10, 15, and 20 ms, respectively. The mean deviation is 1.8 ms for unvoiced stops and 3.3 ms for the voiced stops. The standard deviation is 6 ms for unvoiced stops and 5.1 ms for voiced stops. The distribution of δ is skewed to the right because the hand-labeled boundary in TIMIT often precedes the actual location of the release as also noted by Lin and Wang.[4] A transcriptor may mark the closure-burst boundary at the beginning of the pre-frication interval. This may explain the skewness observed and justify the choice of m_1 corresponding to an interval of 6 ms.

3. Comparison with the previous work

We compare the results of the APR algorithm with those of three state-of-the-art algorithms: RoR-based (denoted by "Liu"),[2] adaptive filtering approach (denoted by "N&S"),[3] and RF-based (denoted by "L&W").[4] Strictly speaking, the results are not comparable because of the different sizes of the datasets considered, different criteria for the temporal tolerance for detection, and differences in the target sets considered. The number of tokens considered for testing in our study is the highest among all the studies reported in the literature.

Figure 10 also shows the ROC curve of the N&S algorithm (manually read from their study[3] and re-plotted here)

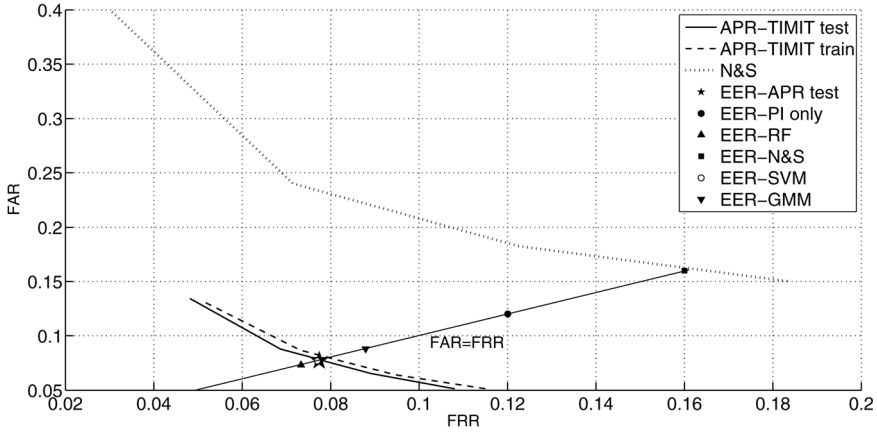

FIG. 10. The ROC curves of the APR algorithm (solid line: TIMIT test database, dashed line: TIMIT training database) with the EERs compared with some state-of-the-art methods. FAR: false acceptance rate; FRR: false rejection rate. The ROC curve (dotted line, for a subset of the TIMIT test database) and EER for the N&S algorithm are taken from the paper by Niyogi and Sondhi (Ref. 3). EERs for RF, SVM, and GMM are taken from the work of Lin and Wang (Ref. 4).

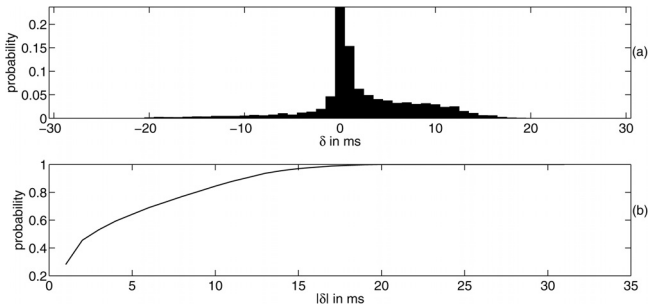

FIG. 11. Histograms of the temporal deviation, δ, for the TIMIT test and training databases combined. (a) PDF and (b) CDF.

and the EERs achieved by the different algorithms. The trade-off between FAR and FRR is less severe in the case of the APR algorithm than the N&S algorithm. As an example, to achieve an FRR of 5%, the APR algorithm results in an FAR of 13% against 32% for the N&S algorithm. The EER of the N&S algorithm is 16%, with affricates included in the rejection class. Lin and Wang[4] report an EER of 7.3% using a RF classifier, 7.7% using a SVM classifier, and 8.8% using a 16-component GMM on the TIMIT test set. These EERs are also indicated in Fig. 10. The EER of the APR algorithm for the TIMIT test set (7.7%) equals that of SVM and is marginally (0.4%) less than that of RF. However, in the study by L&W, a temporal tolerance of more than 30 ms is used for defining a correct detection. If the temporal tolerance is increased to 40 ms from 20 ms in the APR algorithm, the EER decreases to 7.2% from 7.7% on the TIMIT test set, which is better than that with both the RF and SVM classifiers used in L&W.[4] Lin and Wang have noted that the computational load of SVM makes it impossible to be used as an efficient burst detector. On the other hand, our proposed algorithm uses only two temporal measures and a simple rule based classifier. Liu has not reported the EER, but reports 19% deletion (FRR) for stop-bursts with a temporal tolerance criterion for detection being 30 ms. Another study by Niyogi *et al.* reports an EER of about 13% using SVM classifier on a single dialect of the TIMIT test database.[5] The EER for this case is not shown in Fig. 10.

In the L&W algorithm, the percentage of detections are 64%, 86%, 99.2%, and 99.6% for temporal deviations of 5, 10, 20, and 30 ms, respectively. The corresponding results for the APR algorithm are 64%, 84%, 100%, and 100%, respectively. The mean and standard deviation of δ are 4.7 and 5.7 ms, respectively, for the L&W algorithm compared to 2.7 and 5.8 ms, respectively, for the APR algorithm on the test database. Given the aforementioned facts, the performance of the proposed algorithm appears significant, with the results being comparable to the best in the literature. The features (temporal and spectral) used in these studies being different, their merits could possibly be advantageously combined.

B. The TIMIT database with white and babble noise—Global SNR

To the best of our knowledge, there are very few studies in the literature reporting on CBT detection performance in

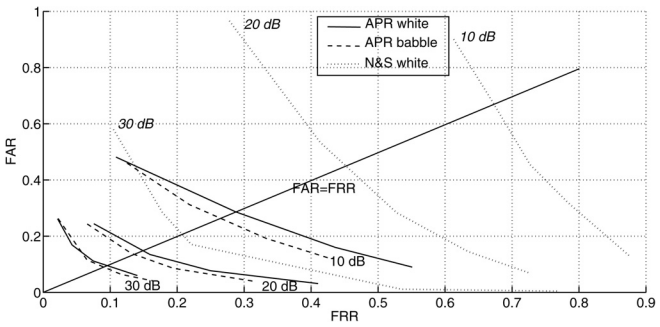

FIG. 12. The ROC curves of the APR algorithm for the TIMIT test database with additive white (solid line) and babble noise (dashed line) under various global SNRs. Also shown are the ROC curves of the N&S algorithm (Ref. 3) (dotted line) for white noise for the same SNRs.

the presence of noise. The ROC curves of the APR and N&S algorithms on noisy speech are shown in Fig. 12 for three different SNRs. The APR algorithm achieves EERs of 9.5, 15, and 28.5% at 30, 20, and 10 dB global SNRs, respectively, for white noise as compared to 20, 46, and 67%, respectively, reported by Niyogi and Sondhi.[3] It is observed that the EER (15%) of the APR algorithm at 20 dB global SNR is about the same as that achieved by the N&S algorithm on clean speech. Further, the degradation with decreasing SNR is rapid in the case of the N&S algorithm. Although Liu has reported the results for landmark detection in the presence of noise, those results are not on the TIMIT database and the performance for the detection of the CBTs has not been explicitly mentioned.

Figure 12 also illustrates the ROC curves of the APR algorithm for babble noise for the same SNR values. To the best of our knowledge, there is no previous study on the CBT detection with babble noise. It is interesting to note that the performance in the presence of speech-like babble noise is about the same as that with white noise. The degradation in the presence of noise may be caused by the presence of noise components during the closure interval and the smudging of the transient nature of the burst, which reduces the PI.

C. The TIMIT database with Schroeder noise—Local SNR

Figure 13 shows the ROC curves obtained for this experiment along with those of the N&S algorithm. EERs of around 7.8, 8.1, and 10.8% are obtained at 20, 10, and 0 dB

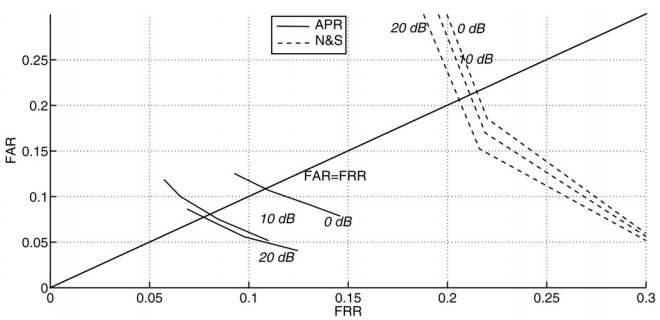

FIG. 13. The ROC curves for the TIMIT test database with the additive Schroeder noise for various local SNRs for the APR (solid line) and the N&S algorithms (Ref. 3) (dashed line).

SNRs, respectively, with the APR algorithm as compared to about 21%–22% for the N&S algorithm for all the three SNRs. For 0 dB SNR, EER obtained with the APR algorithm is almost one half of that obtained by the N&S algorithm.

For the APR algorithm, the performance at 20 dB local SNR is almost the same as that on clean speech. This advantage arises because the amplitude of local noise samples during stop closures is relatively small and, hence, the PI is not degraded significantly. This shows that the APR algorithm effectively captures the transient nature of the CBTs and the robustness depends on how well the transient nature is preserved.

D. The NTIMIT database—Telephone quality

The ROC curve for the complete NTIMIT test database of the APR algorithm is shown in Fig. 14. An EER of 18.2% has been achieved. The degradation of performance, compared to the TIMIT database (EER 7.7%), arises because of the limited channel bandwidth and lower SNR. However, the EER value (18.2%) is comparable to (15%) that on TIMIT for 20 dB global SNR with additive noise. The ROC curve for the N&S algorithm is also shown in Fig. 14, where the NTIMIT test set was used both for training and testing (with 1346 tokens from the target class), for which an EER of about 31% has been reported. However, the performance was poorer (35% EER) when the adaptive filter was trained using the TIMIT training set.[3] Liu[11] also reports the results for a subset of NTIMIT (251 tokens from the target class). A deletion rate of 22%, insertion rate of 5% with 12% substitution, and 17% neutral landmarks has been reported. The better performance of the APR algorithm may be due to an appropriate choice of the knowledge-based temporal measures used.

E. The Buckeye corpus—Conversational speech

Figure 15 shows the ROC curve of the APR algorithm for the experiment on the Buckeye corpus. An EER of 19% has been achieved, which is about 12% more than that obtained for read speech of the TIMIT database. It is interesting to note that the threshold for the PI for this EER is about the same as that for the TIMIT database. The FRR for unvoiced stops is less (13%) than that for voiced stops (27%). The results are generally observed to be better for female speakers.

There have been very few studies on stop detection in conversational speech. A previous study has considered the

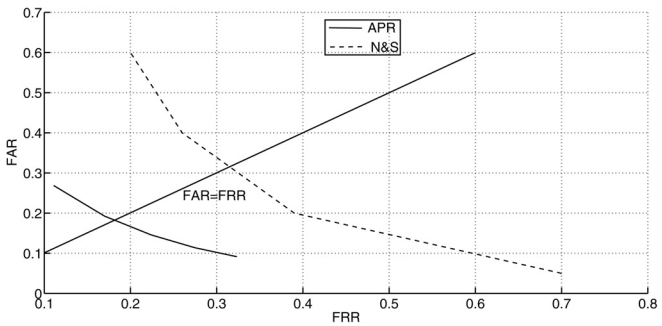

FIG. 14. The ROC curves of the APR (solid line) and N&S algorithms (Ref. 3) (dashed line) for the NTIMIT test database.

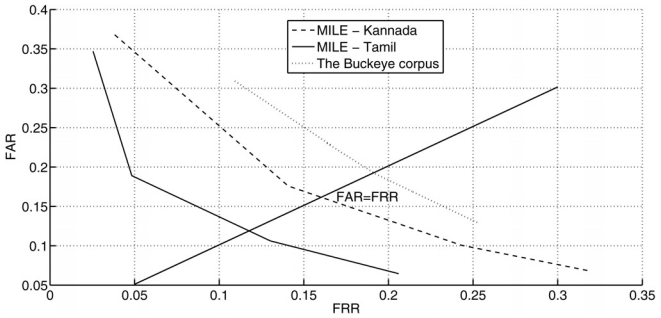

FIG. 15. The ROC curves generated by the APR algorithm for the Buckeye corpus (dotted line) and the MILE databases (dashed line: Kannada database, solid line: Tamil database).

detection of stop releases in conversational speech for the Switchboard corpus.[42] However, the results of that study cannot be compared with the present work because (i) a hierarchical scheme is used for landmark detection, where stops form a subclass under the class [-sonorant]; i.e., detection accuracy for stop-bursts is given assuming that the class [-sonorant] is known and (ii) frame-wise accuracy is given in that study. It has been observed in another study[13] that despite high frame-wise accuracy (∼90%) for phonological features, the overall phone accuracy can be very low (∼60%). Thus, more detailed studies are warranted on the CBTs of conversational speech.

F. The MILE database—Dravidian languages

The ROC curves for the two MILE databases are also shown in Fig. 15. An EER of about 16% is achieved for Kannada, while an EER of 12% is achieved for Tamil at a threshold for the PI which is about the same as that arrived at for the TIMIT and the Buckeye databases. The detection accuracy for voiced stops is lower than that for unvoiced stops. Especially, it is least for /g/ (as in the case of TIMIT) at around 50% for both these languages. If one excludes /g/ while calculating the performance measures, an EER of 11% is achieved for both languages. This is a small-scale study to test the validity of the algorithm on other languages. However, a large-scale study is warranted. Nevertheless, the results obtained are better than those reported in a recent study across six languages where the average detection rate for stops is around 74% for language-specific classifiers and 64% for cross-lingual and multilingual classifiers.[43]

Table I summarizes all the experiments, their results, and the comparison with the previous work. It may be seen that, independent of statistical training and with only two temporal measures, the APR algorithm (i) is as effective as the best in the literature for the entire TIMIT database, (ii) is better than the state-of-the-art techniques for all other experiments considered, namely, global white and babble noise, local noise, and telephone speech, and (iii) is scalable to conversational speech and two languages other than English.

G. Analysis of errors

In this section, we analyze the causes for errors obtained in the experiments conducted on TIMIT since it is the only

database with closure-burst boundary labeling. The CBTs are missed by the APR algorithm in the following cases: (i) Occasionally, some stops are produced without a prominent release, resulting in a value for the PI less than the threshold.[44] An extreme case of this is when there is no release at all.[45] (ii) Some unvoiced stop consonants (often /t/) manifest temporally like a strong fricative without a well-defined closure-burst signal structure. These cases result in a low PI. (iii) Affricates sometimes manifest signal properties more likely to be similar to those of the fricatives than the stops.

Falsely detected CBTs occur in the following cases: (i) Onset of vowels and glides with irregular periodicity, vocal fry, etc. (ii) Nasal-vowel transition with a sudden release resulting in a high-frequency component resembling a voiced-burst release.[13] (iii) Stop-fricative boundaries that have been labeled in the TIMIT as /αcl/-/β/, where α is a stop and β is a fricative. A genuine weak burst of the stop may indeed be present at the boundary, in which case the algorithm has actually detected it.[46] However, this issue needs further investigation. (iv) A transient-like signal structure occurring within a fricative segment, especially during /f/.[47] (v) Impulse-like noise within the silence segments marked as "h#," "pau," "epi," and stop closures, which are not related to stop-bursts.[3]

VI. CONCLUSION

A. Summary

The problem of detecting CBT instants from a continuous speech signal is addressed in this paper using two simple temporal measures, without the need for statistical training and complex classification machines. The PI proposed appears to be an appropriate acoustic correlate for the detection of the transient nature of the bursts. The usefulness of the maximum normalized cross correlation is demonstrated for reducing the spurious candidates at voiced onsets and for detecting weak bursts of voiced stops. Since the algorithm makes use of two scalar temporal measures and a simple rule-based classifier, it is expected to be computationally efficient. The algorithm has been extensively validated on databases recorded under diverse recording conditions, operating environments, dialects, languages and styles of speech (read and conversational). The robustness of the algorithm has been studied on stationary and non-stationary noise as well as on speech with channel degradation. The results are found to be comparable or better than the state-of-the-art methods for similar experimental conditions. Based on the present work, we infer that by an appropriate choice of acoustic correlates specific for a phonetic feature and a simple set of rules (a knowledge-based approach), an algorithm can perform as well as sophisticated statistical classifiers using high-dimensional feature vectors. Hence, it appears that it is worth pursuing a knowledge-based approach for discovering such correlates for all the phonetic features.

B. Future research directions

We list below some important findings which need further investigation. (i) The parameters m_1 and m_2 have been kept constant for all the experiments conducted in this study. However, small-scale experiments on conversational speech have shown that the performance can be improved by optimizing the values of these parameters. (ii) The definition of the PI can be extended to auditory subbands of the speech signal. This may further help in improving the performance of the CBT detection in the presence of noise, detection of place of articulation, and other landmarks. (iii) An extended definition of the PI, dynamic plosion index, can be explored for estimating the closure duration and VOT of stops. (iv) Instead of taking a binary decision based on a threshold (hard decision), a confidence measure can be defined to quantify the degree of certainty in decision making. For instance, if the threshold is fixed at 8, a genuine burst with PI around 7.95 would be missed. However, this can still be declared as a stop burst with a confidence measure close to but less than unity (7.95/8), which can be used later with other features for making a decision about the phone.

[1]K. N. Stevens, *Acoustic Phonetics* (MIT Press, Cambridge, MA, 1998), Chap. 7–9, pp. 323–350, 405–415, 512.
[2]S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," J. Acoust. Soc. Am. **100**, 3417–3430 (1996).
[3]P. Niyogi and M. M. Sondhi, "Detecting stop consonants in continuous speech," J. Acoust. Soc. Am. **111**, 1063–1072 (2002).
[4]C.-Y. Lin and H.-C. Wang, "Burst onset landmark detection and its application to speech recognition," IEEE Trans. Audio, Speech, Lang. Process. **19**, 1253–1264 (2011).
[5]P. Niyogi, C. Burges, and P. Ramesh, "Distinctive feature detection using support vector machines," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (1999).
[6]A. R. Jayan and P. C. Pandey, "Detection of stop landmarks using Gaussian mixture model of speech spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (2009), pp. 4681–4684.
[7]F. Jelinek, *Statistical Methods for Speech Recognition* (MIT Press, Cambridge, MA, 1997), Chaps. 1–15, pp. 1–280.
[8]L. Rabiner and B. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ, 1993), Chaps. 6–8, pp. 321–482.
[9]N. Bitar and C. Espy-Wilson, "A knowledge-based signal representation for speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (1996), pp. 29–32.
[10]A. Juneja and C. Espy-Wilson, "A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition," J. Acoust. Soc. Am. **123**, 1154–1168 (2008).
[11]S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 1995.
[12]V. Zue, "The use of speech knowledge in speech recognition," Proc. IEEE **73**, 1602–1615 (1985).
[13]S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," Comput. Speech Lang. **14**, 333–353 (2000).
[14]J. Hou, L. Rabiner, and S. Dusan, "Automatic speech attribute transcription (ASAT)—The front end processor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (2006).
[15]K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368 (1978).
[16]L. Lisker and A. Abramson, "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**, 384–422 (1964).
[17]T. Cho and P. Ladefoged, "Variation and universals in VOT: Evidence from 18 languages," J. Phonetics **27**, 207–229 (1999).
[18]P. Auzou, C. Ozsancak, R. Morris, M. Jan, F. Eustache, and D. Hannequin, "Voice onset time in aphasia, apraxia of speech and dysarthria: A review," Clin. Linguist. Phonetics **14**, 131–150 (2000).
[19]M. Sonderegger and J. Keshet, "Automatic measurement of voice onset time using discriminative structured prediction," J. Acoust. Soc. Am. **132**, 3965–3979 (2012).
[20]N. Bitar, "Acoustic analysis and modeling of speech based on phonetic features," Ph.D. thesis, Boston University, Boston, MA, 1997.
[21]C. Espy-Wilson, "Acoustic measures for linguistic features distinguishing the semivowels /w,j,r,l/ in American English," J. Acoust. Soc. Am. **92**, 736–757 (1992).

[22]N. Chomsky and M. Halle, *The Sound Pattern of English* (MIT Press, Cambridge, MA, 1968), Chaps. 1–7, pp. 1–490.
[23]A. Salomon, C. Espy-Wilson, and O. Deshmukh, "Detection of speech landmarks: Use of temporal information," J. Acoust. Soc. Am. **115**, 1296–1305 (2004).
[24]C. W. Turner, P. E. Souza, and L. N. Forget, "Use of temporal envelope cues in speech recognition by normal and hearing impaired listeners," J. Acoust. Soc. Am. **97**, 2568–2576 (1995).
[25]P. K. Ghosh and S. S. Narayanan, "Closure duration analysis of incomplete stop consonants due to stop-stop interaction," J. Acoust. Soc. Am. **126**, EL1–EL7 (2009).
[26]T. V. Ananthapadmanabha and B. Yegnanarayana,"Epoch extraction of voiced speech," IEEE Trans. Acoust., Speech, Signal Process. **23**(6), 562–570 (1975).
[27]D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, edited by W. Kleijn and K. Paliwal (Elsevier, New York, 1995), pp. 495–518.
[28]N. Dhananjaya, B. Yegnanarayana, and P. Bhaskararao, "Acoustic analysis of trill sounds," J. Acoust. Soc. Am. **131**, 3141–3152 (2012).
[29]A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," IEEE Trans. Audio, Speech, Lang. Process. **21**(12), 2471–2480 (2013).
[30]S. Zhao, "The stop-like modification of /ð/: A case study in the analysis and handling of speech variation," Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA, 2007.
[31]J. B. Henderson and B. H. Repp, "Is a stop consonant released when followed by another stop consonant?" Phonetica **39**, 71–82 (1982).
[32]Another example of a consonant cluster with two distinct releases occurs in the word "expectation" in test-dr1-faks0-si943 of the TIMIT test database, between 2.85 and 2.99 s. A burst corresponding to /k/ can be seen in that utterance around 2.918 s as confirmed by temporal and spectral characteristics.
[33]P. Ladefoged and K. Johnson, *A Course in Phonetics*, 6th ed. (Wadsworth, Boston, MA, 2011), Chap. 3, p. 62.
[34]D. Crystal, *A Dictionary of Linguistics and Phonetics*, 6th ed. (Blackwell, Malden, MA, 2008), pp. 191–192.
[35]J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgrena, *DARPATIMIT, Acoustic-Phonetic Continuous Speech Corpus*, NISTIR Publication No. 4930 (US Department of Commerce, Washington, DC, 1993).
[36]"Noisex-92," URL http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex. html, (Last viewed 9/28/2013).
[37]C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing* (1990), pp. 109–112.
[38]M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," Speech Commun. **45**, 89–95 (2005).
[39]"Buckeye corpus," URL http://buckeyecorpus.osu.edu/ (Last viewed 9/28/2013).
[40]"Thirukkural and Vak TTS system," URL http://mile.ee.iisc.ernet.in/tts (Last viewed 9/28/2013).
[41]P. Mermelstein, "Automatic segmentation of speech into syllabic units," J. Acoust. Soc. Am. **58**, 880–883 (1975).
[42]M. Hasegawa-J, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sonmez, and T. Wang, *Landmark-Based Speech Recognition: Report of the 2004 Johns Hopkins Summer Workshop* (2005).
[43]S. M. Siniscalchi, T. Svendsen, D. C. Lyu, and C. Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target specific training data," IEEE Trans. Audio, Speech, Lang. Process. **20**, 875–887 (2012).
[44]For example, the /t/ beginning at 3.76 s in the TIMIT test sentence dr2-mdbb0-si1825.
[45]Examples for this case may be seen in the TIMIT test sentence test-dr1-mjsw0-si1010 for the phone /b/ starting at 1.13 s and test-dr2-mmdb1-sx95 for the phone /g/ starting at 0.25 s and 1.06 s.
[46]For example, /tcl/, /s/ at 3.43 s in test-dr1-faks0-si943.
[47]For example, /f/ in test-dr1-faks0-si943, at 1.505 s.