Detection of fricatives using *S*-transform

Hari Krishna Vydana and Anil Kumar Vuppala

Citation: The Journal of the Acoustical Society of America **140**, 3896 (2016); doi: 10.1121/1.4967517
View online: https://doi.org/10.1121/1.4967517
View Table of Contents: https://asa.scitation.org/toc/jas/140/5
Published by the Acoustical Society of America

Detection of fricatives using *S*-transform

Hari Krishna Vydana[a] and Anil Kumar Vuppala
International Institute of Information Technology, Hyderabad, India

(Received 6 May 2016; revised 15 October 2016; accepted 28 October 2016; published online 22 November 2016)

Two prime acoustic characteristics of fricatives are the concentration of spectral energy above 3 kHz and having noisy nature. Spectral domain approaches for detecting fricatives rely on capturing the information from spectral energy distribution. In this work, *S*-transform based time-frequency representation is explored for detecting fricatives from continuous speech. *S*-transform based time-frequency representation exhibits a progressive resolution which is tailored for localizing the high frequency events (i.e., onset and offset of fricative regions) with time. Spectral evidence computed from *S*-transform based time-frequency representation is observed to perform better compared to the spectral evidence computed from short time Fourier transform. The existing predictability measure based approach relies on capturing the noisy nature of fricatives. A phone level comparative analysis is carried out between *S*-transform and predictability measure based approaches and the phone distribution of the detected fricatives is observed to be complimentary. In this work, a combination of *S*-transform and predictability based approaches is put forth for detecting fricatives from continuous speech. Apart from detecting the presence of a fricative, the proposed *S*-transform based approach and combined approach exhibit better accuracy in detecting the boundaries of fricatives, i.e., extracting the durational information of fricatives.
 [http://dx.doi.org/10.1121/1.4967517]

[JFL] Pages: 3896–3907

I. INTRODUCTION

Significant research interest has been focused on detecting the place of articulation of fricatives.[1–3] However, studies directed toward detecting the fricative regions from speech are minimal.[4] The focus of this study is directed toward detecting fricative broad class from speech. During the production of a fricative vocal tract is constricted enough along its length to produce a noisy sound when air is forced through this constriction.[5] The phenomenon of frication introduces a noisy aperiodic energy concentrated above 3 kHz.[6] The class of fricatives considered during the study is comprised of sibilant fricatives ([s], [sh], [z], and [zh]) and nonsibilant fricatives ([f], [v], [th], and [dh]). Affricates are produced with a supra-glottal closure followed by a burst release, and a frication is caused by the turbulence of air associated with the burst release, by their acoustic manifestation affricates, which are closer to fricatives than bursts.[4] During the study, affricates [ch] and [jh] are also considered in the class of fricatives. The intensity of sibilants is relatively high compared to that of nonsibilants.[2] During the production of affricate "ch" the frication produced is comparable to a sibilant fricative.[4]

On a broad scale, present days automatic speech recognition (ASR) systems are two types,[7] i.e., statistical model based ASR and phonetic feature based ASR. In a statistical model based ASR, speech signal is parametrized to one of the features like mel-frequency cepstral coefficients (MFCCs), perceptive linear predictive coefficients (PLP), or Gaussian posteriorgrams. Hidden Markov models (HMMs), artificial neural networks (ANNs), and, recently, deep neural networks (DNNs)[8–10] are employed to perform regression tasks between labeled phonetic transcription and acoustic parametrization. In phonetic feature based ASR, phonetic feature specific information is initially extracted from speech, and these extracted phonetic features are used for detecting the phone identity. A landmark based speech recognition system is an example of phonetic feature based ASR. Landmarks are the regions in an utterance, where the acoustic correlates of distinctive features are more prominent.[11] A new bottom-up approach for developing robust ASR system proposed in Ref. 12 uses neural network based attribute detectors for detecting place and manner of articulations, and a merging network is employed to map the detected attributes to the corresponding phones.

In a statistical model based ASR, a uniform parametrization scheme has to be adopted for recognizing various phones. Contrasting to a statistical model based ASR in a landmark based ASR speech signal can be analyzed at various levels (segmental, sub-segmental, and suprasegmental levels) and various parameterizations, which gives a better platform for engaging the information from various theories of phonetics and neuro-science, etc.[7,13] The landmark based ASR is comprised of a broad manner classifier in its initial stage to reduce the search space.[14] Band energies, spectral peaks, and valleys in various subbands are used as acoustic parameters (APs) to compute landmarks.[13] Support vector machines (SVMs) trained with 39-dimensional MFCCs with appropriate framing and frequency shift parameters are employed for detecting fricative broad class.[14] Subband energy difference between two frames that are temporally spaced 50 ms, 26 ms apart is studied for obtaining the landmarks of fricatives.[11] A methodology for combining the acoustic phonetic knowledge with statistical learning is explored for segmenting the speech to

[a]Electronic mail: hari.vydana@research.iiit.ac.in

five broad manner classes in Ref. 15. The developed combined system has exhibited better performance compared to HMM based segmentation using 39-dimensional cepstral features. Recurrent neural networks (RNNs) trained with 39-dimensional cepstral features are explored for detecting various phonological features and these features are later used for developing a fricative broad class identification system.[16] Spectral and temporal evidences like dominant resonant frequency, epoch strength, and numerator of group delay at zero-frequency have been used as cues to detect the unvoiced fricatives from continuous speech.[4] The gain of inverse of the all-pole filter at zero-frequency (i.e., predictability measure) is used as an acoustic cue to perform sonorant-fricative classification.[17]

However, most of the above methods need la arge amount of training data and the developed system cannot be easily ported to a new language or data. Most of the above methods rely on the information from the short time spectral envelope computed using short time Fourier transform (STFT). For example, in the case of a voiced fricative, both frication event and glottal closure events exist mutually exclusively in a single glottal cycle, short time spectral envelope reflects the average spectral characteristics over the entire time frame, and the characteristics of a voiced fricative are not represented effectively due to averaging across time. Usefulness of the proposed approach to represent the events like the voiced fricatives and nonsibilant fricatives of short duration compared to conventional STFT based approach is shown in Secs. IV and VI.

The spectral energy of the fricative has a unique frequency distribution, i.e., most of the spectral energy is above 3 kHz. This property is explored in detecting fricatives through various parameters like dominant resonant frequency,[4] spectral centroid,[18] numerator of the group delay spectrum at zero-frequency,[4] and band energy ratio].[4,19,20] Apart from spectral energy distribution, fricative has a noisy nature. During the production of fricative the influence of the vocal tract is less compared to a sonorant. The relations imparted by the vocal tract in successive samples of a sonorant are more compared to a fricative, i.e., fricative is a less correlated signal compared to a sonorant. The correlations among the successive samples of a signal make the signal more predictable from the past samples. A predictability measure computed as a sum of all the linear prediction coefficients is used as an acoustic cue for sonorant fricative classification in Ref. 17. Both the acoustic cues (i.e., distribution of spectral energy and predictability measure) reflect two complementary evidences for detecting the fricatives. This paper presents a combined approach using the spectro-temporal evidence computed from the *S*-transform based spectral representation and the predictability based evidence proposed in Ref. 17 for detecting fricative broad class in speech. The method is unsupervised such that the complex process of training and the tedious process of collecting the transcription of the data are not required. This evidence can be used as complementary evidence to fricative broad class classifier in landmark based ASR. Apart from ASR, the unsupervised fricative broad class classifier is widely appreciated in audio search.

Phonological information of speech can be obtained by observing the acoustic correlates, and the observed acoustic correlates are used as acoustic phonetic descriptors of that particular acoustic event.[13] Duration of an acoustic event is one of the acoustic phonetic descriptors that can be relied on for obtaining information about the finer class (sibilant fricatives, nonsibilant fricatives, affricates, and stops) from the broad class of fricatives. One major drawback of frame based analysis is that the durational cues, which can play a key role in discriminating various acoustic events, cannot be stringently imposed due to averaged spectral representation over a time frame. For example, the pre-frication region in stop sounds shows acoustic characteristics similar to the fricatives, and the release region of an affricate has an acoustic manifestation similar to that of a fricative. But the duration of the frication event pertaining to a sibilant fricative is much higher compared to the pre-frication region in bursts and the frication region in nonsibilant fricatives.[2] Duration of frication along with strength of frication and the existence of an associated plosion event helps to discriminate between sibilant fricatives, nonsibilant fricatives, affricates, and the stops. But for extracting the durational cues of various acoustic events like frication, a signal processing tool with good temporal resolution to localize that particular acoustic event (i.e., frication) is a pre-requisite.

In this context, *S*-transform based time-frequency representation is explored for detecting the acoustic cues pertaining to fricative landmarks. *S*-transform was initially developed for geophysics applications[21] to detect the onset and offset of high frequency signals. In this study, the hypothesis is that the capability of *S*-transform to time localize the high frequency signal gets reflected in detecting the boundaries of fricative regions in speech. Due to better time resolution for high frequency events it can better represent voiced fricatives (as shown in Sec. IV), and durational cues from the detected frication events can also be accurately obtained (as shown in Sec. VI).

The remainder of this paper is organized in the following manner. Section II describes the databases used during the study. Section III elaborates the algorithm of *S*-transform, implementation, and its implications on speech signal analysis. The effectiveness of *S*-transform in representing the frication events is presented by considering an example of voiced fricative in Sec. IV. Algorithm to detect the fricative broad class from continuous speech using *S*-transform based spectral evidences is described in Sec. V. A comparative analysis is carried out between *S*-transform and STFT based spectral evidences for detecting fricative broad class in Sec. VI. Performances of *S*-transform based approach and predictability measure based approach are compared to highlight the complementary nature of both the approaches in Sec. VI. An approach to incorporate the information from spectral distribution and measure of predictability is put forth in Sec. VII. Conclusion and future scope are presented in Sec. VIII.

II. DATABASE

Speech signals collected from a phonetician are used during the course of study. Data consists of alveolar, post-alveolar,

and retroflex voiced fricatives [z], [ʒ], [z̠] produced in isolation. Apart from that, each of the voiced fricatives is produced with three vowel contexts ([ɑ],[i],[u]). Each of the voiced fricatives is succeeded by one among the three vowels [ɑ], [i], [u] to form consonant-vowel (CV) units. The junction between voiced-unvoiced fricatives is studied using three voiced fricatives succeeded by unvoiced fricatives, i.e., [z → s], [ʒ → ʃ], [z̠ → ʂ]. All the speech signals are repeated three times to form three tokens per sound. Speech signals are sampled at 48 kHz with 16 bit quantization used for the study. During the study, signal is resampled to 16 kHz for analysis.

A. TIMIT-database

Performance of the proposed method for detecting the fricative broad class from continuous speech is evaluated using TIMIT database.[29] The labeled boundaries in the TIMIT database are used during the evaluation. Fricatives considered for the present study are [s], [z], [sh], [zh], [f], [v], [dh], [th], [ch], and [jh]. A subset of TIMIT database with 40 speakers with 20 male and 20 female, each speaker containing 10 sentences each of 2–3.5 s long are considered for study. Speech signals sampled at 16 kHz and 16 bits/sample encoding is used during the course of present study. The glottal fricatives [hh], [hv] are not considered during the analysis.

III. *S*-TRANSFORM, IMPLEMENTATION, AND IMPLICATIONS ON SPEECH

S-transform and the terminology used in this paper are adopted from Ref. 21. Recently a lot of scientific interest has been shown in analyzing *S*-transform and many other variants such as generalized *S*-transform,[22] modified *S*-transform,[23] and inverse of *S*-transform.[24] The notations of basic *S*-transform are adhered in this paper.

STFT is one of the most widely used time-frequency representations to study nonstationary signals

$$S(\tau,f)=\int_{-\infty}^{\infty} s(t)g(t-\tau)e^{-j2\pi ft}\,dt, \tag{1}$$

where $s(t)$ is the time domain signal and $g(t)$ is the window signal. Conversely, to study the time properties of a particular frequency spectrum of the entire time domain signal $S(f)$ is windowed by a frequency window $G(f)$, and the inverse Fourier transform of windowed spectrum is a time-frequency representation termed as short frequency time transform (SFTT)[25]

$$S(f,\tau)=\int_{-\infty}^{\infty} S(f)G(f-\acute{f})e^{-j2\pi\acute{f}\tau}\,d\acute{f}, \tag{2}$$

since windowing for STFT is in the time domain, the spectrum averaged over the entire time frame is used as time-frequency representation. Conversely, for SFTT the frequency representation is an average representation over the entire band. SFTT based analysis is more preferable where time boundary is crucial. Use of a window with progressive frequency resolution gives a better time-frequency representation.[26]

S-transform can be viewed as an SFTT with a progressive resolution obtained by a frequency dependent window. *S*-transform uniquely combines a frequency dependent resolution of time-frequency space.[21] Continuous *S*-transform of a function $s(t)$ is given by

$$S_T(\tau,f)=\int_{-\infty}^{\infty} s(t)\frac{|f|}{\sqrt{2\pi}}e^{-(\tau-t)^2f^2/2}e^{-j2\pi ft}\,dt, \tag{3}$$

$S_T(\tau,f_k)$ is a one-dimensional (1-D) function of time for a constant frequency f_k showing how the amplitude and phase for this frequency f_k changes with time. A choice of Gaussian window is made owing to its better localization, minimal side-lobe energy, and symmetry in both time and frequency domains.[27] The prime focus of the study is to explore the temporal resolution of *S*-transform for representing frication events.

A. Relation between *S*-transform and STFT

If the signal $s(t)$ is windowed using $g(t)$ then STFT is given by

$$S(\tau,f)=\int_{-\infty}^{\infty} s(t)g(t-\tau)e^{-j2\pi ft}\,dt. \tag{4}$$

For *S*-transform a Gaussian window with zero mean and variance σ is used

$$g(t)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\left(t^2/2\sigma^2\right)}. \tag{5}$$

Progressive frequency resolution is achieved by equating variance of the window at a frequency (f) to the inverse of that frequency, i.e.,

$$\sigma(f)=\frac{1}{|f|}. \tag{6}$$

As $g(t)$ is a Gaussian function (i.e., even function)

$$g(t-\tau)=g(\tau-t). \tag{7}$$

From relations (4), (5), and (7), *S*-transform (S_T) can be written as convolution of two functions over the variable t,

$$S_T(\tau,f)=\int_{-\infty}^{\infty} p(t,f)g(\tau-t,f)\,dt, \tag{8}$$

$$S_T(\tau,f)=p(t,f)*g(t,f), \tag{9}$$

where

$$p(t,f)=s(t)e^{-j2\pi ft}, \tag{10}$$

$$g(t,f)=\frac{|f|}{\sqrt{2\pi}}e^{-\left(t^2f^2/2\right)}. \tag{11}$$

Computing the Fourier transform of $S_T(\tau,f)$ from Eq. (9),

$$B(\beta,f) = \mathcal{F}\{S_T(\tau,f)\} = P(\beta,f)G(\beta,f). \quad (12)$$

Let $P(\beta,f)$ and $G(\beta,f)$ be the Fourier transforms of the $p(t,f)$ and $g(t,f)$,

$$P(\beta,f) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft}e^{-j2\pi\beta t}\,dt, \quad (13)$$

$$P(\beta,f) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi(f+\beta)t}dt, \quad (14)$$

$$P(\beta,f) = S(\beta+f), \quad (15)$$

$$G(\beta,f) = \int_{-\infty}^{\infty} \frac{|f|}{\sqrt{2\pi}}e^{-\left(t^2f^2/2\right)}e^{-j2\pi\beta t}\,dt, \quad (16)$$

where

$$\mathcal{F}\left\{e^{-at^2}\right\} = \sqrt{\frac{\pi}{a}}e^{-\pi^2f^2/a}, \quad (17)$$

$$G(\beta,f) = \frac{\cancel{|f|}}{\cancel{\sqrt{2\pi}}}\sqrt{\frac{\cancel{2\pi}}{\cancel{f^2}}}e^{-2\pi^2\beta^2/f^2}, \quad (18)$$

$$G(\beta,f) = e^{-2\pi^2\beta^2/f^2}. \quad (19)$$

From Eq. (12),

$$B(\beta,f) = P(\beta,f) \times G(\beta,f), \quad (20)$$

$$B(\beta,f) = S(\beta+f)e^{-(2\pi^2\beta^2/f^2)}. \quad (21)$$

S-transform is the inverse Fourier transform of Eq. (21) (for $f \neq 0$),

$$S_T(\tau,f) = \mathcal{F}^{-1}\{B(\beta,f)\} = \int_{-\infty}^{\infty} S(\beta+f)e^{-(2\pi^2\beta^2/f^2)}e^{2\pi\tau\beta}\,d\beta. \quad (22)$$

B. Implementing *S*-transform

- Compute the discrete Fourier transform (DFT) of the N-point time series.
- Compute the DFT of N-point Gaussian function to select the frequency range.
- Shift the spectrum of time series such that the frequency of the spectrum to be selected matches with the zero-frequency of the frequency selecting Gaussian function and multiply both the signals.
- Compute inverse discrete Fourier transform (IDFT) of the product obtained in the above step to obtain S-transform representation of the time series.

Owing to the use of the frequency dependent progressive window with its variance inversely related to frequency, S-transform gives better time resolution for high frequency signals and frequency resolution for low frequency signals.

C. Implications of *S*-transform on the speech signal

- Consider a 2 s utterance with 16 kHz sampling frequency, from Eq. (22) it is evident that a 32 000 point DFT is to be computed, i.e., the frequency scale is divided into 16 000 points. A Gaussian of length 32 000 for selecting one frequency point is needed. A 32 000 point IDFT for every frequency point is to be computed and there are 16 000 such points. It can be observed that, the length of the input signal dictates the frequency resolution of the S-transform.
- It is evident that the abovementioned approach for computing S-transform turns out to be computationally very expensive even for a 2 s utterance. In this paper, a block processing approach is used to attain a balance between the computational load and resolution of S-transform. Suppose that for obtaining S-transform with a frequency resolution of 10 Hz, i.e., one frequency point for every 10 Hz, to cover a frequency range of 0–8 kHz 800 points are to be used. For obtaining this frequency resolution, a block of 1600 samples are needed to compute S-transform, i.e., 100 ms block.
- S-transform for every 100 ms block is computed and the computed S-transform is appended with the S-transform of remaining blocks in an utterance to get the S-transform of an utterance.

A restricted study is carried out to study the effectiveness of S-transform for representing speech signals. During the study, the typical example of a voiced fricative is considered and this study is presented in Sec. IV.

IV. APPLYING *S*-TRANSFORM FOR SPEECH SIGNALS

The effectiveness of S-transform for detecting the boundary of fricative regions is studied in this section and the database mentioned in Sec. II is used for this study. A voiced fricative has a unique production mechanism in which voicing and frication are present in a single glottal cycle. Although both the events voicing and frication are present in one glottal cycle, both are mutually exclusive, i.e., voicing is maintained only in the glottal closing phase and frication is present only in the glottal opening phase. Frication energy (high frequency energy) will be periodic with glottal period, peak in the frication energy occurs at glottal opening phase and valley in frication energy are at the glottal closure instants. A periodic frication energy in correspondence with the glottal period indicates the presence of a voiced fricative. Speech signal of a voiced fricative [z] is presented along with conventional wideband spectrogram and S-transform in Fig. 1.

From Fig. 1 the nature of the voiced fricative can be clearly observed in S-transform compared to STFT wideband band spectrogram. As shown in Fig. 1(b) sharp onset and offset of frication is visible in S-transform based spectrogram compared to the conventional wideband spectrogram. As shown in Fig. 1(c) wideband spectrogram may show similar results when viewed as plots, but due to the lack of sharp onset and offset points for high frequency energy the frication energy gets spread to glottal closure cycle. The sharp boundaries of the frication are the major cues for indicating periodic presence of frication in a voiced fricative. Signal representing the periodic presence of frication energy is computed from conventional wideband, due to the lack of sharp boundaries and the presence of wideband artifacts computed evidence will eventually lose its periodic nature.

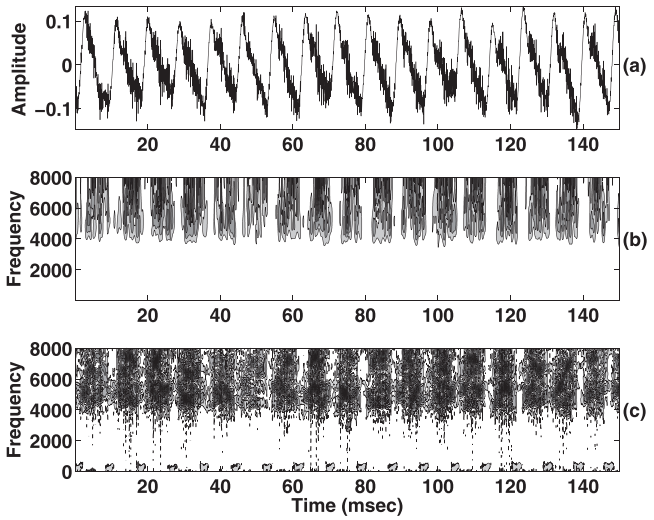

FIG. 1. Time-frequency representation of alveolar voiced fricative [z] using conventional wideband spectrogram and *S*-transform. (a) Speech signal, (b) *S*-transform based spectrogram, (c) conventional spectrogram with 2 ms frame size and one sample frame shift.

Figure 2 illustrates the production mechanism and its correlation with the acoustic cue developed to detect voiced fricative. Figure 2(a) shows the existence of impulsing due to glottal closure instants and frication mutually exclusively in a glottal cycle. From Fig. 2(b) it is evident that frication is confined to only the glottal opening cycle and there is no frication in the glottal closure cycle. Figure 2(c) is the 1-D temporal curve obtained by computing the spectral energy above 5 kHz and the Fig. 2(d) shows the short time energy contour obtained by computing energy of Fig. 2(c) with a window size of 5 ms and a window shift of one sample. Peak in the frication index pointing to the glottal opening cycle indicates high frication energy in the glottal opening phase, valley pointing to glottal closure instant indicates the lack of frication. Presence of the voiced fricative in the vicinity of a vowel is presented in Fig. 3, although the vowel also exhibits

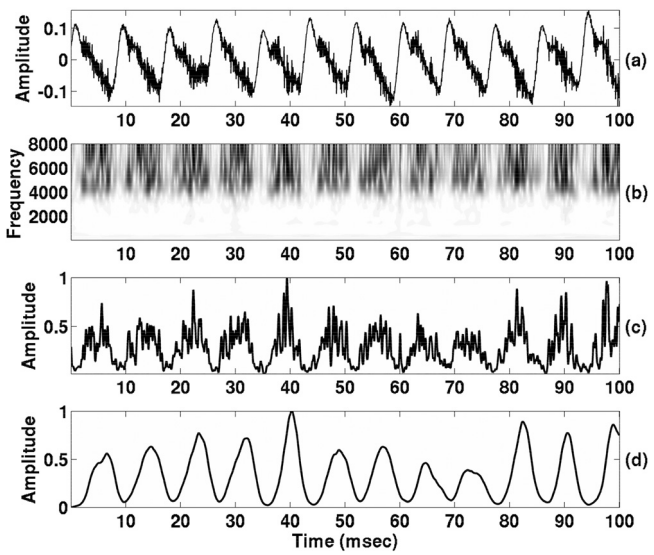

FIG. 2. Characteristics of alveolar voiced fricative [z]. (a) Speech signal, (b) *S*-transform based spectrogram, (c) 1-D temporal curve obtained by computing the spectral energy above 5 kHz, and (d) short time energy of contour shown in (c).

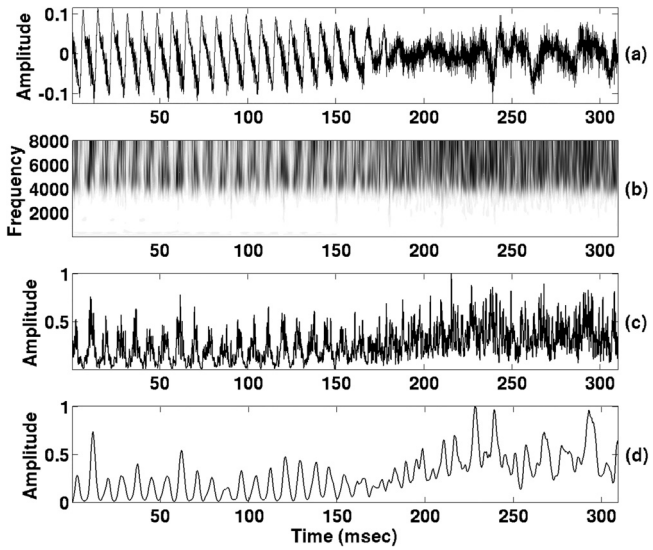

FIG. 3. Characteristics of alveolar voiced fricative [z] with a vowel context [i]. (a) Speech signal, (b) *S*-transform based spectrogram, (c) 1-D temporal curve obtained by computing the spectral energy above 5 kHz, and (d) short time energy of contour shown in (c).

periodicity similar to a voiced fricative but the strength of the frication energy for a vowel is much less.

The boundary of voiced-unvoiced fricative is presented in Fig. 4. The efficiency of *S*-transform to localize the changes in signal with time can be observed. For an unvoiced fricative, frication energy is aperiodic and is almost a noisy surface with high average energy compared to the voiced fricative. The origin of the aperiodic nature in the fricative energy gives the boundary of voiced fricative when succeeded by an unvoiced fricative. Figure 4 illustrates characteristics of voiced-unvoiced junction. From Fig. 4(b) it can be observed that frequency distribution of both voiced and unvoiced frication is almost similar, frication is continuous in an unvoiced fricative. From Fig. 4(d), i.e., frication energy is high in unvoiced fricative compared to voiced fricative. The effectiveness of

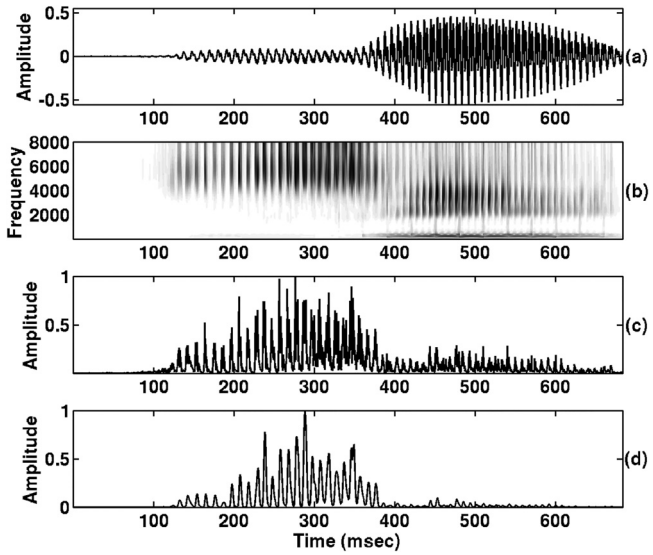

FIG. 4. Characteristics of the boundary between the alveolar voiced-unvoiced fricative [z → s]. (a) Speech signal, (b) *S*-transform based spectrogram, (c) 1-D temporal curve obtained by computing the spectral energy above 5 kHz, and (d) short time energy of contour shown in (c).

S-transform for detecting the frication regions can be clearly visualized from Figs. 3 and 4.

V. PROPOSED *S*-TRANSFORM BASED APPROACH FOR DETECTING FRICATIVES

The phenomena of frication introduces a noisy aperiodic energy concentrated around above 3 kHz. Although some fricatives have some lower frequency energy but spectral energy of fricatives is mostly concentrated in higher frequency regions, i.e., above 3 kHz. Spectral energy above 1.5 kHz is considered as energy due to the frication in a fricative. The ratio of the sum of energies of all the frequency bands above 1.5 kHz to sum of energies at all frequency bands below 1.5 kHz at every sample gives a 1-D time domain signal called frication ratio (F_r). The frication ratio is computed as

$$F_r = \frac{\sum_{k=N_f}^{N} |S_T(k,f)|}{\sum_{k=1}^{N_f} |S_T(k,f)|}, \tag{23}$$

where $|S_T(k,f)|$ is the magnitude response of the *S*-transform time-frequency representation. N_f denotes the index of *S*-transform coefficient corresponding to 1.5 kHz. N is the *S*-transform coefficient corresponding to $f_s/2$ kHz, where f_s is the sampling rate of the speech signal. The obtained frication ratio (F_r) computed by the ratio is mean-smoothed with a window of 20 ms to make the temporal evolutions of frication ratio smooth and the evidence is peak normalized and this is termed "frication index" in this paper. A threshold based decision is enforced on the frication index contour to detect the fricative and nonfricative regions. A durational constraint of 10 ms is also used to remove the spurious evidences, i.e., if duration of the detected evidence has less than 10 ms then such evidences are not considered as fricative. The criteria for choosing the values of threshold are presented in Sec. V A.

A. Arriving at the threshold

In this work, a threshold based decision is adapted to discriminate fricative and nonfricative regions. The region of speech where the frication index is greater than the threshold θ is the hypothesized fricative region. The region of speech where the frication index is less than the threshold θ, is the detected nonfricative region. In this work, a duration based decision is used to discriminate true evidences (fricatives) and spurious evidences (nonfricative regions). Initially a threshold based decision is used on the frication index and then the hypothesized fricative regions greater than 10 ms are considered as the true fricative regions.

The threshold value in this method is empirically obtained using dataset II A. Thresholds for the proposed method are obtained using the number of phones correctly classified. The fricative region computed from the algorithm is termed as detected fricative region or predicted fricative region in this work. For a fricative, if the percentage overlap between the predicted fricative region and the ground truth is greater than the overlap criterion, then the particular fricative phone is termed as correctly classified fricative phone. The overlapping criterion is employed to study the effectiveness of the algorithm in detecting the boundary of fricative regions. The nonfricative region, which is falsely detected as fricative by the algorithm, is termed as falsely detected fricative region. A nonfricative phone with falsely detected fricative region >10 ms duration is termed as falsely detected fricative. In the literature, the percentage of correctly detected fricative phones and that of falsely detected fricative phones have been reported as true-detection rate and false-alarm rate, respectively.

In arriving at a threshold decision (θ), an overlapping criterion of 50% is considered. Performance of the proposed algorithm at various thresholds is presented in Table I. Column 1 gives the class of fricatives, i.e., sibilant, nonsibilant, affricates, and falsely detected fricatives. Column 2 is the total number of phones in the database pertaining to each class. Columns 3–12 are the number of correctly detected fricatives at various thresholds. From Table I the point to be speculated is that the thresholds are very low and of the order 0.001–0.1. This is major because the work is aimed at detecting the boundary of a fricative region, which is a region between two valleys of frication index. Although the threshold appears low from the number and class of falsely detected fricatives, it can be observed that this threshold is good enough to reject most of the nonfricative phones. In this work, a threshold of 0.01 is chosen for detecting the fricative regions. From Table I, although the threshold values <0.01 (present threshold) might look good in the scenario of number of correctly detected fricatives, but for such thresholds a part of the vowel region gets falsely detected as fricatives, to avoid such circumstances a threshold to 0.01 is used in this work.

TABLE I. Performance of the proposed algorithm at various threshold values.

		Threshold value (θ)																		
Fricative class	Total number of phones	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1
Sibilant	1110	1109	1108	1108	1106	1102	1101	1100	1099	1098	1096	1075	1054	1031	1004	980	952	927	899	604
Nonsibilant	651	492	441	395	367	337	307	290	277	259	239	145	104	77	58	45	38	32	28	18
Affricates	175	175	175	175	174	174	173	172	169	169	169	165	162	157	155	152	148	144	140	85
Correctly detected fricatives	1936	1776	1724	1678	1647	1613	1581	1562	1545	1526	1504	1385	1320	1265	1217	1177	1138	1103	1067	707
Falsely detected fricatives	12 603	2279	1658	1336	1146	1005	920	831	768	721	675	460	346	282	251	209	181	162	140	95

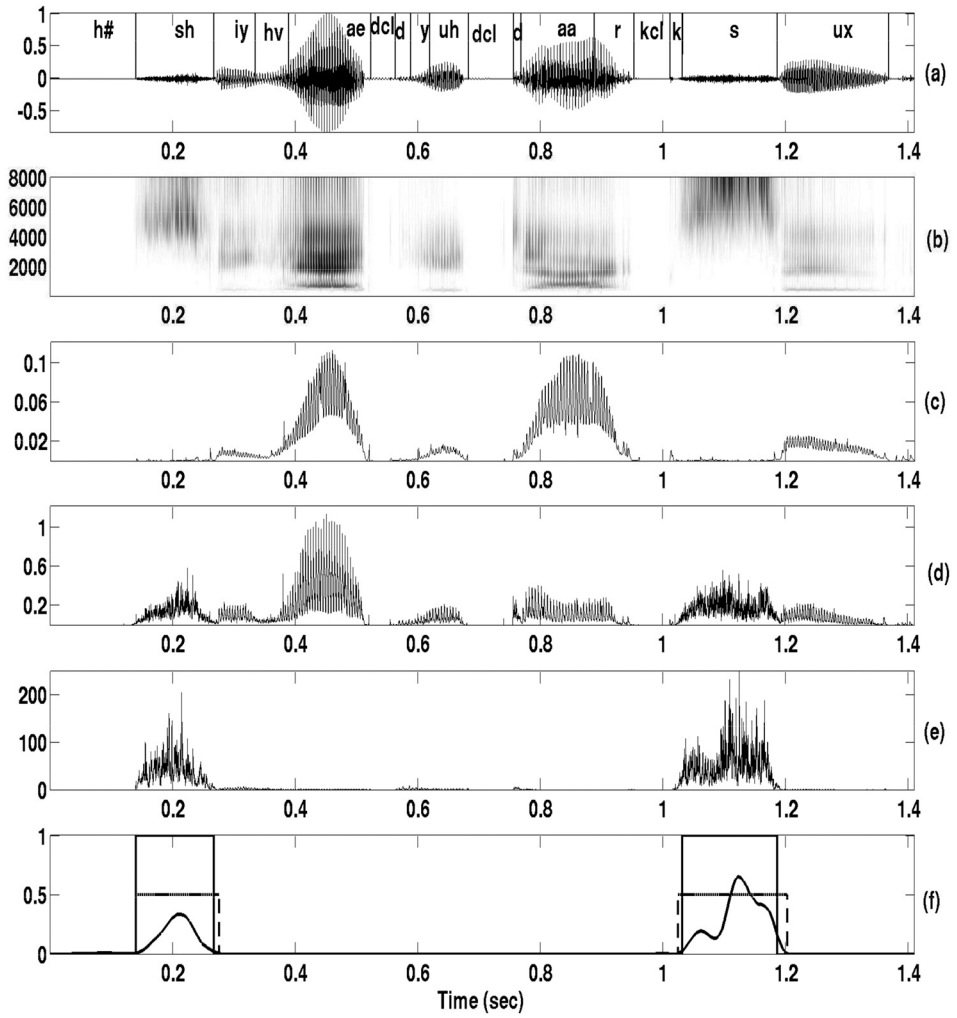

FIG. 5. (a) Speech signal of phonetically labeled TIMIT utterance "She had your dark suit in greasy wash water," (b) *S*-transform, (c) 1-D temporal curve obtained by computing the spectral energy below 1.5 kHz, (d) 1-D temporal curve obtained by computing the spectral energy above 1.5 kHz, (e) frication ratio (F_r), (f) friction index and fricative regions predicted by the algorithm is shown in dotted line and the ground truths are indicated using the solid line.

Figure 5 shows the evidences used in this algorithm. Phonetically labeled TIMIT utterance "She had your dark suit in greasy wash water" is shown in Figs. 5(a) and 5(b) is its *S*-transform. Figure 5(c) is a 1-D temporal curve obtained by computing the spectral energy below 1.5 kHz. Figure 5(d) is a 1-D temporal curve obtained by computing the spectral energy above 1.5 kHz and frication ratio (F_r) computed by the ratio of Figs. 5(d) and 5(c) is shown in Fig. 5(e). Fricative index obtained is shown in Fig. 5(f). In Fig. 5(f), the fricative regions predicted by the algorithm are shown in dotted lines, and the ground truths obtained from the transcriptions are indicated by the solid line. The ground truth and predicted fricative regions are plotted with different amplitudes to indicate the accuracy of the approach in detecting the boundary.

VI. EVALUATION OF PROPOSED *S*-TRANSFORM BASED APPROACH FOR DETECTING FRICATIVES

Performance of the proposed method is evaluated using TIMIT database.[29] The information about the boundary and identity (fricative and nonfricative) of phones is obtained from the transcriptions provided in the database. During the present study [s], [z], [sh], [zh], [f], [v], [dh], [th], [ch], and [jh] are considered as the sounds under fricative broad-class, and they are detected from continuous speech. The effectiveness of the proposed *S*-transform based approach to detect the boundary of fricative and nonfricative regions is studied by computing the percentage correctly detected fricatives (TAR, true-acceptance rate) at various overlapping criteria in Table II. Column 1 of Table II specifies the class of fricatives, i.e., sibilant, nonsibilant, and affricates. Columns 2–6 indicate the percentage of fricatives getting detected at overlapping criteria ranging from 50% to 90%.

One interesting observation from the above results is that the degradation in the performance of sibilant fricatives with increase in overlapping criteria is much less from the range of 50%–80%. This better efficiency of detecting the fricative boundary can be attributed to the efficiency in *S*-transform in localizing high frequency events in speech like frication onset and offset points. But a slight decrease in the performance can be observed if the overlapping criteria are further increased and this decrease can be viewed as the

TABLE II. Performance of the proposed *S*-transform based approach at various overlapping criteria (TAR, true-acceptance rate).

Class of sounds	TAR at various overlapping criteria				
	50%	60%	70%	80%	90%
Sibilant fricatives	98.73	97.83	96.48	93.33	84.68
Nonsibilant fricatives	36.71	31.18	25.34	18.58	11.36
Affricates	96.57	95.42	94.28	92	84.57

cumulative effect of two factors. One of them is due to averaging effects induced in fricative index due to the use of a time window in computing the energy of the frication ratio. The other is due to marking errors in phonetic labeling.

The relevance of the *S*-transformed space in representing the frication events a phone level analysis is presented in Table III. In Table III, proposed *S*-transform based approach for detecting fricative regions is compared with STFT based approach and predictability based approach. Columns 1 and 2 of Table III are the identity and total number of the fricative phones in the database. Columns 3–7, 8–12, and 13–17 are the number of correctly detected fricatives in *S*-transform, STFT, and predictability based approaches at various overlapping criteria. Row 13 of Table III is the total number of correctly detected fricative phones. Row 14 of Table III is the total number of falsely detected fricatives.

A. STFT based approach for fricative detection

Short time spectral envelope computed using conventional STFT with a frame size of 20 ms and frame shift of 5 ms is used for detecting the fricative regions. Fricative index described in the Sec. V computed from the short time spectral envelope is used for detecting the fricative regions. A threshold of 0.01 is chosen empirically, the regions of speech whose fricative index is >0.01 is considered as a detected fricative region. A comparison between performance of the proposed *S*-transform based approach and the STFT based approach for fricative detection is presented in Table III. In this work, the performance of STFT based approach using the wideband spectral envelope (i.e., with a frame size of 5 ms and a frame shift of one sample) is computed and the performance is poorer than the conventional STFT based approaches.

B. Comparing the performances of STFT and *S*-transform based approaches

From columns 3 and 8 of Table III, it can be observed that the number of correctly detected fricative phones in the *S*-transform based approach is slightly higher than the STFT based approaches for a 50% overlapping criteria. From columns 3–7 and 8–12 of Table III, it can be observed that with an increase in the overlapping criteria, the number of correctly detected fricatives drastically decreases in STFT based approach, which is not the case in the *S*-transform based approach. The better performance of *S*-transform based evidence for detecting fricative regions can be attributed to the capability of *S*-transform to localize the high frequency events.

Although the number of falsely detected fricatives in the STFT based approach appears to be less, but this less number in misclassified phones can be attributed to the inability of the STFT based spectral envelope to represent weak frication in the burst regions due to its averaged spectral representation over the time frame.

The interesting observation that can be noted in case of voiced fricatives ([z], [zh], [v], and [dh]), i.e., from rows 4, 6, 8, and 10 of Table III is that the performance of the *S*-transform based approach is consistently better than the STFT based approach. The better performance of *S*-transform can be attributed to its capability of localizing the high frequency events. In *S*-transform representation, a spectral envelope is obtained at every sample, but in STFT based representation a spectral envelope is obtained for every time frame. In an *S*-transform representation the spectral envelope is an averaged spectral representation over all the frequency components while a STFT based spectral envelope is an averaged spectral representation over a time frame. The representation in *S*-transform is chosen with progressive resolution such that high frequency events like frication onset and offset are better time localized. In the case of voiced fricative both voicing and frication present in a glottal cycle mutually exclusively in time, which demands a representation that can maintain the mutually exclusive nature of voiced fricative. *S*-transform can better represent the characteristics of a voiced fricative maintaining both frication and voicing in a single glottal cycle but conventional STFT gives an averaged spectral envelope for a time frame of 20 ms (more than one glottal cycle). Although

TABLE III. Comparing the performance of *S*-transform, STFT, and predictability based approaches at various overlapping criteria.

		S-transform based approach					STFT based approach					Predictability based approach				
		Overlapping criteria					Overlapping criteria					Overlapping criteria				
Phone	Total	50	60	70	80	90	50	60	70	80	90	50	60	70	80	90
s	606	606	605	601	593	551	606	603	597	574	477	606	605	604	584	392
z	301	291	287	284	270	238	278	271	248	205	149	262	251	222	156	71
sh	191	189	185	178	166	144	190	188	183	171	133	191	191	190	182	129
zh	12	10	9	8	7	7	9	8	8	6	5	10	9	8	4	2
f	189	117	101	87	61	33	123	105	82	52	12	181	172	164	132	65
v	175	30	23	15	9	8	19	14	8	5	3	27	16	11	8	3
th	54	33	26	24	20	12	26	21	15	12	7	48	46	37	48	27
dh	233	59	53	39	31	21	21	16	10	5	3	86	78	65	69	63
ch	71	69	68	67	66	62	69	67	65	61	46	71	71	70	69	63
jh	104	100	99	98	95	86	93	89	82	70	43	96	96	92	88	69
Correctly detected fricatives	1936	1504	1456	1401	1318	1162	1434	1382	1298	1161	878	1578	1535	1463	1340	884
Falsely detected fricatives	12 438			675					526					911		

STFT has the inability to reflect the characteristics of a voiced fricative in a spectral domain but the number of voiced fricatives getting correctly detected in the STFT based approach is not as low as expected at 50% overlapping criterion, i.e., voicing in a voiced fricative is sustained only to a part of the voiced fricative and later part of the voiced fricative behaves similarly to the unvoiced fricative.[4,5] But for computing the boundary of a voiced fricative *S*-transform exhibits better capability, which can be seen in rows 4, 6, 8, and 10 in Table III at high overlapping criteria, i.e., (60%–90%). Performance of both the approaches is poor in detecting the fricative [f], which can be observed from row 7 of Table III. The fricative [f] has a weak high frequency energy, and the approaches that rely on capturing the frequency distribution information (*S*-transform and STFT based approaches) for detecting fricatives perform poorly.

C. Comparing the performances of predictability and *S*-transform based approaches

1. Linear prediction coefficients (LPC)

The approach presented in Ref. 17 is termed as predictability based approach or LPC based approach for fricative broad class detection. Performance of *S*-transform and predictability based approaches is compared in Table III. Fricatives are noisy in nature, i.e., Fricative sounds are less correlated signals compared to sonorant sounds. Less predictable nature of a fricative is captured in predictability based approach where *S*-transform and STFT rely on capturing the frequency distribution properties of fricatives for detecting the fricatives. Although the performance of *S*-transform and predictability based approaches for detecting fricative broad class are comparable, both the approaches exploit two different properties of a fricative sound (i.e., spectral distribution and noisy nature). The complementary nature of both the approaches can be clearly observed in case of fricatives [z] and [f], i.e., rows 7 and 4 of Table III. In case of the fricative phone [z] both voicing and frication are present in a single glottal cycle as shown in Fig. 4. Voiced fricative can also be viewed as a periodic frication modulated by a glottal source. The predictability measure computed as the sum all linear predictive coefficients is used in the predictability based approach[17] for detecting fricatives. Linear predictive coefficients are obtained by minimizing the mean squared error between the signal predicted by the linear combinations of past samples and the original signal. Although frication is less predictable the voicing envelope of a voiced fricative is predictable so the predictability in voiced fricative makes the predictability based approach perform poorly in detecting voiced fricatives. Similar observations can be made in case of other voiced fricatives ([zh] and [v]). In case of fricative [f] high frequency energy is very weak so the *S*-transform based approach, which relies on capturing the spectral energy distribution, performs poorly but the predictability based approach, which relies on the predictability of the signal, performs better than the *S*-transform based approach. From Table III, it can be observed that the performance of sibilant fricatives is higher than nonsibilant fricatives, which is in accordance with the earlier studies.[4,17] In the nonsibilant class the performance of voiced sounds is even more poor and the similar trend is observed in the case of sibilant fricatives.

TABLE IV. A phone level analysis of falsely detected fricatives in *S*-transform based approach (FAR, false-alarm rate).

Class	Phones	Total number of phones	Number of falsely detected fricatives
Bursts	k	390	160
	t	319	224
	d	300	113
	g	167	30
	b	183	11
	p	213	25
Others		9568	112
Nonfricative phones		12 438	675 (5.4%, FAR)

A phone level analysis is carried out to study the phone distribution of falsely detected fricatives and the observations are tabulated in Table IV. Columns 1 and 2 of Table IV are the class and identity of falsely detected fricatives. Column 3 of Table IV is the total number of phones in the database. Column 4 is the number of phones falsely detected as fricatives and the percentage (%) of falsely detected fricatives. From Table IV, among the total number of nonfricative phones (12 438) only (5.4%) of phones got detected as fricatives. The majority of falsely detected fricatives belong to the class of stop sounds and sounds associated with aspiration. Around 83% of the falsely detected fricatives are stops, 10% are aspirated sounds.

VII. COMBINING *S*-TRANSFORM AND PREDICTABILITY BASED APPROACH FOR DETECTING FRICATIVES

As described in Sec. VI, both the approaches, i.e., *S*-transform and predictability based approaches, exploit two complementary properties of fricatives. Although the overall accuracy of both the approaches is comparable, the phone distribution of correctly detected fricatives in these approaches is quite complementary, which can be observed from Table III. The complementary nature of both the approaches is exploited for fricative detection by combining both the approaches.

The evidence from both the approaches (*S*-transform and predictability based approaches) is combined in the manner shown:

$$\gamma[n] = \alpha_{S_T}[n] + \alpha_{\mathrm{LPC}}[n]. \tag{24}$$

Here, $\alpha_{S_T}[n]$ and $\alpha_{\mathrm{LPC}}[n]$ are the predicted fricative regions from *S*-transform and predictability based approaches. $\alpha_{S_T}[n]$ and $\alpha_{\mathrm{LPC}}[n]$ are the binary contours with the value "1" in the predicted fricative regions and "0" in the nonfricative regions. The regions of speech where the combined evidence greater than or equal to 1 is considered as the predicted fricative regions in the combined approach. A fricative region detected in any one of the approaches is considered as a predicted fricative region in the combined approach. Predicted

TABLE V. Performance of combined approach at various overlapping criteria.

Phone	Total	Overlapping criteria				
		50	60	70	80	90
s	606	606	606	605	605	597
z	301	293	289	286	284	272
sh	191	191	191	190	189	175
zh	12	10	10	10	10	9
f	189	186	179	172	159	104
v	175	38	31	24	15	10
th	54	51	49	45	43	27
dh	233	101	91	77	63	44
ch	71	71	71	70	69	68
jh	104	101	101	101	100	92
Sibilant fricatives	1110	1100	1096	1091	1088	1053
Nonsibilant fricatives	651	376	350	318	280	185
Affricates	175	172	172	171	169	160

fricative regions in the combined approach ($\alpha_{\text{comb}}[n]$) is given by Eq. (25),

$$\alpha_{\text{comb}}[n] = \begin{cases} 1, & \text{if } \gamma[n] \geq 1 \\ 0, & \text{elsewhere.} \end{cases} \tag{25}$$

Column 1 of Table V is the identity of the fricative phone. Column 2 is the total number of a particular fricative phone in the dataset. Columns 3–7 is the number of fricatives detected using the combined approach at different overlapping criteria. Rows 12–14 is the number of sibilant, nonsibilant, and affricate phones detected using the combined approach. Rows 16–18 are the percentage of sibilant, nonsibilant, and affricate phones detected using the combined approach.

Performance of the combined approach is expected to have the cumulative effect of performances in both the approaches. From Table V, an overall improvement in detecting all the fricatives can be observed. In the case of fricative [z], i.e., column 3 of Table V, performance of the combined approach is higher than the predictability based approach and comparable to the *S*-transform based approach. In the case of fricative [f], i.e., column 6 of Table V, performance of the combined approach is higher than the *S*-transform based approach and comparable to the predictability based approach. As the predicted fricative regions from both the approaches are combined at the decision level the best performing scenarios from both the approaches get reflected in the combined approach.

A comparative analysis of the percentage of correctly detected fricatives in the *S*-transform based approach, predictability, and combined approaches is presented in Table VI. Column 1 of Table VI is the class of fricatives, i.e., sibilants, nonsibilants, and affricates. Column 2 is the total number of fricatives in the dataset. Columns 3–7, 8–12, and 13–17 are the percentages of correctly detected fricatives in the *S*-transform based approach, predictability based approach, and combined approach at different overlapping criteria varying from 50%–90%. From row 1 of Table VI, it can be observed that, in the case of sibilant fricatives performance of *S*-transform based approach is consistently more than the predictability based approach; the inefficiency of detecting the voiced fricatives [z] in the predictability based approach majorly gets reflected in the better performance of the *S*-transform based approach. It can be observed from row 2 of Table VI, that in the case of nonsibilant fricatives due to their weak frequency distribution the performance of *S*-transform based approach is poorer compared to the predictability measure based approach. Performance of the *S*-transform based approach at larger overlapping criteria is significantly higher than the predictability based approach. Performance of the combined approach is consistently higher than both *S*-transform and predictability based approaches. The cumulative effect of higher detection accuracies for sibilant fricatives in the *S*-transform based approach and higher detection accuracies of nonsibilant fricatives in the predictability based approach can be observed in the combined approach. A significant improvement in percentage of correctly detected fricatives at higher overlapping criteria can be observed in combined approach.

The performance of the combined approach for detecting fricative manner class is compared with the state-of-the-art manner class detector frameworks presented in Refs. 12 and 28. A Gaussian mixture model/Hidden Markov model (GMM/HMM) based manner class detector and a DNN/HMM based manner class detector are implemented using the KALDI recipe and the results are tabulated in Table VII. Form the results of Table VII, it can be noted that although the performance of both the detectors is comparable at 50% overlapping criteria, but at higher overlapping criteria a drastic decrease in the performance can be observed. The percentage of falsely detected fricatives is 6% and 3.6% for GMM/HMM and DNN/HMM based detectors, respectively, which is less than that of the proposed combined approach (12%). The lower number of falsely detected fricatives in both the detector frameworks can be attributed to the following reason: In the proposed combined approach majority of stop sounds got detected as fricatives, but in a detector framework stop are modeled as a separate class so the

TABLE VI. Comparing the performance of proposed *S*-transform based approach, predictability based approach, and the combined approach (TAR, true-acceptance rate).

Class of phone	Total	*S*-transform based approach (TAR)					Predictability based approach (TAR)					Combined approach (TAR)				
		Overlapping criteria					Overlapping criteria					Overlapping criteria				
		50	60	70	80	90	50	60	70	80	90	50	60	70	80	90
Sibilant fricatives (%)	100	98.73	97.83	96.48	93.33	84.68	96.31	95.14	92.25	83.42	53.51	99.09	98.73	98.28	98.01	94.86
Nonsibilant fricatives (%)	100	36.71	31.18	25.34	18.58	11.36	52.53	47.93	42.55	39.48	24.27	57.75	53.76	48.84	43.01	28.41
Affricates (%)	100	96.57	95.42	94.28	92	84.57	95.43	95.43	92.57	89.71	95.43	98.28	98.28	97.71	96.57	91.42

TABLE VII. Comparing the performance of proposed combined approach with the state of the art manner class detectors (TAR, true-acceptance rate).

Class of phone	Total	GMM/HMM detector (TAR)					DNN/HMM detector (TAR)					Combined approach (TAR)				
		Overlapping criteria					Overlapping criteria					Overlapping criteria				
		50	60	70	80	90	50	60	70	80	90	50	60	70	80	90
Sibilant fricatives (%)	100	97.92	97.65	96.47	90.24	68.35	96.65	94.85	89.24	71.88	39.33	99.09	98.73	98.28	98.01	94.86
Nonsibilant fricatives (%)	100	52.69	51.61	47.31	37.02	19.97	66.36	57.3	42.7	21.35	10.45	57.75	53.76	48.84	43.01	28.41
Affricates (%)	100	68.39	66.67	59.2	42.53	30.46	83.33	68.97	52.3	34.48	17.82	98.28	98.28	97.71	96.57	91.42

percentage of stops getting detected as fricatives is reduced greatly. An alternate acoustic cue to detect the bursts onset is to be explored to further reduce the percentage of falsely detected fricatives.

A phone level analysis is carried out to study the phone distribution of falsely detected fricatives in the combined approach and the observations are tabulated in Table VIII. Columns 1 and 2 of Table VIII are the class and identity, respectively, of falsely detected fricatives. Column 3 of Table VIII is the total number of phones in the database. Column 4 is the number of phones falsely detected as fricatives.

In the combined approach, the fricative region associated with bursts gets detected more accurately, leading to an increase in the number of falsely detected fricative phones. From Table VIII, among the total number of nonfricative phones (12 438) 9.9% of phones got falsely detected as fricatives. Majority of falsely detected fricatives belong to the class of stop sounds and sounds associated with aspiration. Around 81% of falsely detected fricatives are stops, 11% are aspirated sounds. The turbulence generated during burst release in stop consonants gets detected as fricatives, and abruptness at the boundary and presence of an associated plosion event can be used as a evidence to discriminate fricatives and stops.

VIII. CONCLUSION

In this study, *S*-transform based time-frequency representation is explored for fricative landmark detection. The implications of *S*-transform on speech signal are studied and *S*-transform is customized for speech signal analysis. It is observed from the study that spectral domain cues are more prominent when they are used in *S*-transformed space. The *S*-transform based approach, which relies on capturing the spectral energy distribution, is further combined with an existing predictability based approach that relies on the predictability in the signal. The performance of the combined approach is superior to the performances of both the approaches implemented individually. *S*-transform is more efficient when the exact time location of the particular landmark is crucial. The *S*-transform based landmark approach can be further extended to detection of landmarks such as bursts, voicing onset locations. The *S*-transform basis is to be further analyzed and modified to generate new spaces with interesting properties. Abruptness at the boundary of the fricatives and the presence of an associated plosion event can be used to discriminate between the fricatives and bursts.

TABLE VIII. A phone level analysis of falsely detected fricatives in combined approach (FAR, false-alarm rate).

Class	Phones	Total number of phones	Number of falsely detected fricatives
Bursts	k	390	275
	t	319	309
	d	300	185
	g	167	71
	b	183	44
	p	213	119
Others		9568	231
Nonfricative phones		12 438	1234 (9.9%, FAR)

ACKNOWLEDGMENTS

The authors would like to thank Professor Peri Bhaskararao, International Institute of Information Technology (IIIT) Hyderabad, for his valuable discussions about the phonetic aspects of fricatives. The database described in Sec. IV has been collected from Professor Peri Bhaskararao, IIIT Hyderabad. The authors would like to thank MeitY (Ministry of Electronics and Information Technology) for supporting the research under the Visvesvaraya PhD fellowship scheme.

[1]A. M. A. Ali and J. V. der Spiegel, "Acoustic-phonetic features for the automatic classification of fricatives," J. Acoust. Soc. Am. **109**(5), 2217–2235 (2001).
[2]A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of english fricatives," J. Acoust. Soc. Am. **108**(3), 1252–1263 (2000).
[3]C. Chan and K. Ng, "Separation of fricatives from aspirated plosives by means of temporal spectral variation," IEEE Trans. Acoust., Speech Signal Process. **33**(5), 1130–1137 (1985).
[4]N. Dhananjaya, "Signal processing for excitation-based analysis of acoustic events in speech," Ph.D. dissertation, Department of Computer Science and Engineering, IIT Madras, 2011, pp. 129–184.
[5]S. Narayanan and A. Alwan, "Noise source models for fricative consonants," IEEE Trans. Speech Audio Process. **8**(3), 328–344 (2000).
[6]C. H. Shadle, "The acoustics of fricative consonants," Ph.D. dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1985, pp. 17–18.
[7]T. Ananthapadmanabha, A. Prathosh, and A. Ramakrishnan, "Detection of the closure-burst transitions of stops and affricates in continuous speech using the plosion index," J. Acoust. Soc. Am. **135**(1), 460–471 (2014).
[8]G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag. **29**(6), 82–97 (2012).
[9]L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Canada (2013), pp. 8599–8603.
[10]A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Canada (2013), pp. 6645–6649.
[11]S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," J. Acoust. Soc. Am. **100**(5), 3417–3430 (1996).

[12]S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "A bottom-up modular search approach to large vocabulary continuous speech recognition," IEEE Trans. Audio, Speech, Lang. Process. **21**(4), 786–797 (2013).
[13]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**(2), 1154–1168 (2008).
[14]A. Jansen and P. Niyogi, "Modeling the temporal dynamics of distinctive feature landmark detectors for speech recognition," J. Acoust. Soc. Am. **124**(3), 1739–1758 (2008).
[15]A. Juneja and C. Espy-Wilson, "Segmentation of continuous speech using acoustic-phonetic parameters and statistical learning," in *Proceedings of the 9th International Conference on Neural Information Processing*, Qatar (2002), Vol. 2, pp. 726–730.
[16]S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," Comput. Speech Lang. **14**(4), 333–353 (2000).
[17]T. Ananthapadmanabha, A. Ramakrishnan, and P. Balachandran, "An interesting property of LPCs for sonorant vs fricative discrimination," arXiv:1411.1267 (2014).
[18]A. Frid and Y. Lavner, "Acoustic-phonetic analysis of fricatives for classification using SVM based algorithm," in *Proceedings of the 26th Convention of Electrical and Electronics Engineers in Israel*, Israel (2010), 000751–000755.
[19]K. Driaunys, V. Rudvzionis, and P. Zvinys, "Analysis of vocal phonemes and fricative consonant discrimination based on phonetic acoustics features," Inform. Technol. Control **34**(3), 257–261 (2015).
[20]A. M. A. Ali, J. Van der Spiegel, P. Mueller, G. Haentjens, and J. Berman, "An acoustic-phonetic feature-based system for automatic phoneme recognition in continuous speech," in *Proceedings of IEEE International Symposium on Circuits and Systems*, Florida (1999), Vol. 3, pp. 118–121.
[21]R. G. Stockwell, L. Mansinha, and R. Lowe, "Localization of the complex spectrum: The *s* transform," IEEE Trans. Signal Process. **44**(4), 998–1001 (1996).
[22]C. R. Pinnegar and L. Mansinha, "The *s*-transform with windows of arbitrary and varying shape," Geophysics ***68***(1), 381–385 (2003).
[23]C. R. Pinnegar and L. Mansinha, "The bi-Gaussian *s*-transform," SIAM J. Sci. Comput. **24**(5), 1678–1692 (2003).
[24]C. Simon, S. Ventosa, M. Schimmel, A. Heldring, J. J. Danobeitia, J. Gallart, and A. Manuel, "The *s*-transform and its inverses: Side effects of discretizing and filtering," IEEE Trans. Signal Process. **55**(10), 4928–4937 (2007).
[25]L. Cohen, *Time-Frequency Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1995), Vol. 299, pp. 94–95.
[26]I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," IEEE Trans. Inform. Theory **36**(5), 961–1005 (1990).
[27]A. Janssen, "Optimality property of the Gaussian window spectrogram," IEEE Trans. Signal Process. **39**(1), 202–204 (1991).
[28]I.-F. Chen, S. M. Siniscalchi, and C.-H. Lee, "Attribute based lattice rescoring in spontaneous speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Italy (2014), pp. 3325–3329.
[29]J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus," Technical Report NISTIR 4930 National Institute of Standards and Technology, Gaithersburg, MD (1993).