Noise robust speech rate estimation using signal-to-noise ratio dependent sub-band selection and peak detection strategy

Chiranjeevi Yarra, Supriya Nagesh, Om D. Deshmukh, and Prasanta Kumar Ghosh

Citation: The Journal of the Acoustical Society of America **146**, 1615 (2019); doi: 10.1121/1.5124473
View online: https://doi.org/10.1121/1.5124473
View Table of Contents: https://asa.scitation.org/toc/jas/146/3
Published by the Acoustical Society of America

Noise robust speech rate estimation using signal-to-noise ratio dependent sub-band selection and peak detection strategy

Chiranjeevi Yarra,[1,a)] Supriya Nagesh,[2] Om D. Deshmukh,[3] and Prasanta Kumar Ghosh[1]
[1]*Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore, 560012, India*
[2]*Departement of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA*
[3]*Xerox Research Center India, Bangalore, 560103, India*

(Received 9 January 2019; revised 8 August 2019; accepted 13 August 2019; published online 12 September 2019)

Speech (syllable) rate estimation typically involves computing a feature contour based on sub-band energies having strong local maxima/peaks at syllable nuclei, which are detected with the help of voicing decisions (VDs). While such a two-stage scheme works well in clean conditions, the estimated speech rate becomes less accurate in noisy condition particularly due to erroneous VDs and non-informative sub-bands mainly at low signal-to-noise ratios (SNR). This work proposes a technique to use VDs in the peak detection strategy in an SNR dependent manner. It also proposes a data-driven sub-band pruning technique to improve syllabic peaks of the feature contour in the presence of noise. Further, this paper generalizes both the peak detection and the sub-band pruning technique for unknown noise and/or unknown SNR conditions. Experiments are performed in clean and 20, 10, and 0 dB SNR conditions separately using Switchboard, TIMIT, and CTIMIT corpora under five additive noises: white, car, high-frequency-channel, cockpit, and babble. Experiments are also carried out in test conditions at unseen SNRs of −5 and 5 dB with four unseen additive noises: factory, sub-way, street, and exhibition. The proposed method outperforms the best of the existing techniques in clean and noisy conditions for three corpora. © *2019 Acoustical Society of America.*
https://doi.org/10.1121/1.5124473

[JHLH] Pages: 1615–1628

I. INTRODUCTION

Speech has been one of the main media of communication in human computer interface applications, in which the speech rate has been shown to be useful for automatic speech recognition (ASR)[1,2] and identification of disfluencies.[3] Cucchacarini *et al.*[4] have shown that the speech rate is correlated with the human expert's pronunciation quality ratings. Recently, systems that are similar to these applications have been used in online voice training and spoken language learning,[5] where the voice is typically noisy.[6,7] Under such noisy conditions, a reliable speech rate estimation technique could be necessary for better evaluation of the pronunciation quality. Further, for a smooth functioning of these systems, ASR needs to be robust under such noisy conditions. For ASR, speech rate based continuous frame rate normalization has been used to improve the ASR accuracy.[2,8] In contrast to normalization on the entire audio segment as in continuous frame rate normalization, variable frame rate has been used for ASR;[9–11] here, the frame rate is computed using the frames that are selected based on the energy and entropy based differences between the acoustic properties of two consecutive frames. Both of these methods have also been shown to be suitable for noisy conditions. Thus, a reliable speech rate estimation in noisy conditions could improve the ASR accuracy when techniques based on these methods are used. In addition, Borrie *et al.*[12] have shown that there is a relationship between processing dysarthric speech and speech in noise. Hence, a noise robust speech rate estimation could be useful in the assessment of dysarthric speech,[13,14] where it is known that the speaking rate is an important parameter for analytics. However, estimating the speech rate from a noisy recording still remains a challenging task.

In these applications, the speech (syllable) rate estimated directly from speech acoustics could be useful compared to other alternative approach of estimating the speech rate based on ASR,[15] for example, as proposed by Yuan *et al.*[16] The ASR based speech rate estimation is prone to recognition errors, particularly the errors are more when the speech is from noisy, dysarthic conditions and from language learners. In addition, the speech rate is expected to provide complementary information to ASR instead of being dependent on it. In these cases, the acoustic based speech rate estimation could be advantageous. Typically, the acoustic based speech rate estimation comprises two steps: (1) computing a short-time feature contour such that most of its peaks correspond to the syllable nuclei locations, (2) detecting these peaks and thereby, the syllable nuclei.

Heinrich and Schiel[17] used the contour from the short time root-mean square of a speech signal and an average based threshold mechanism to detect the peaks. Pfau and Ruske[18] estimated the vowel locations based on prominent peaks in the smoothed loudness contour. They used zero crossing rate to accurately estimate the vowel locations. Dekens *et al.*[19] proposed a low frequency modulated energy envelope and a multilevel threshold mechanism based on the

[a)]Electronic mail: chiranjeeviy@iisc.ac.in

characteristics of unvoiced regions to avoid peaks in those regions. Zhang and Glass[20] proposed a contour based on Hilbert envelope and used rhythm guided peak counting to estimate the syllable nuclei. They improved the peak counting by removing the peaks falling in unvoiced regions using voicing decisions (VDs) from the estimated pitch values. Jong *et al.*[21] used an intensity based envelope with peak counting based on VDs to estimate speech rate. Wang and Narayanan[22] proposed a feature contour called "Temporal Correlation Selected Sub-band Correlation" (TCSSBC) and a peak detection strategy (PDS) which involves smoothing and a threshold mechanism in the voiced regions. A comprehensive study comparing eight different methods for speech rate estimation has been summarized by Dekens *et al.*,[23] who found that the TCSSBC based method performs the best for speech rate estimation. We observe that most of the existing techniques use various peak detection strategies based on VDs for speech rate estimation.

The TCSSBC has been used in most of the existing speech rate estimation studies. For example, mode-shapes of the TCSSBC have been used for estimating the speech rate and detecting syllable nuclei in the support vector machine based syllabic peak classification technique.[24] The principle of TCSSBC (spectro-temporal correlation) has been extended in dictionary based speech rate estimation (DSRE) by learning activations based on a dictionary to create a feature contour.[25] Wang and Narayanan[15] further proposed a robust speech rate estimation (RSRE) method to improve the performance of TCSSBC by optimizing the parameters involved in the TCSSBC computation and the PDS with VDs. Due to the effectiveness of TCSSBC and PDS with VDs, RSRE has been considered as a baseline in most of the existing studies.[13,24–27]

However, it has been shown that the performance of RSRE drops with increasing noise.[24,25] Yarra *et al.*[24] showed that the speech rate estimation performance using RSRE drops by 50% on the CTIMIT[28] corpus compared to TIMIT.[29] Similarly, the syllable nuclei detection performance drops by 25%.[24] CTIMIT contains recordings of TIMIT in a noisy condition. We also observe that the performance of RSRE reduces when the speech is corrupted with various additive noises at different signal-to-noise ratio (SNR) conditions. Figure 1 shows the TCSSBC (blue) contour computed from speech under clean condition and additive white Gaussian noise at 20, 10, and 0 dB SNRs for an exemplary sentence (*"So would radar picket ships"*) taken from the TIMIT corpus. In the figure, the peaks detected by RSRE are indicated by the taller blue colored vertical lines and the estimated VDs (1 for predicted regions and 0 otherwise) used in RSRE are indicated by the cyan colored line. From the figure, it is observed that the number of correct syllabic peaks detected by the RSRE drops from 20 to 0 dB. It is also observed that the TCSSBC peak in the syllable "ɪ t" is missed by RSRE at 10 and 0 dB SNRs because they do not fall under the estimated VDs. These together indicate that with more additive noise, the error in the estimated VDs increases, thereby causing more syllabic peaks to be missed. Thus, errors in the estimation of VDs directly affect the RSRE performance.

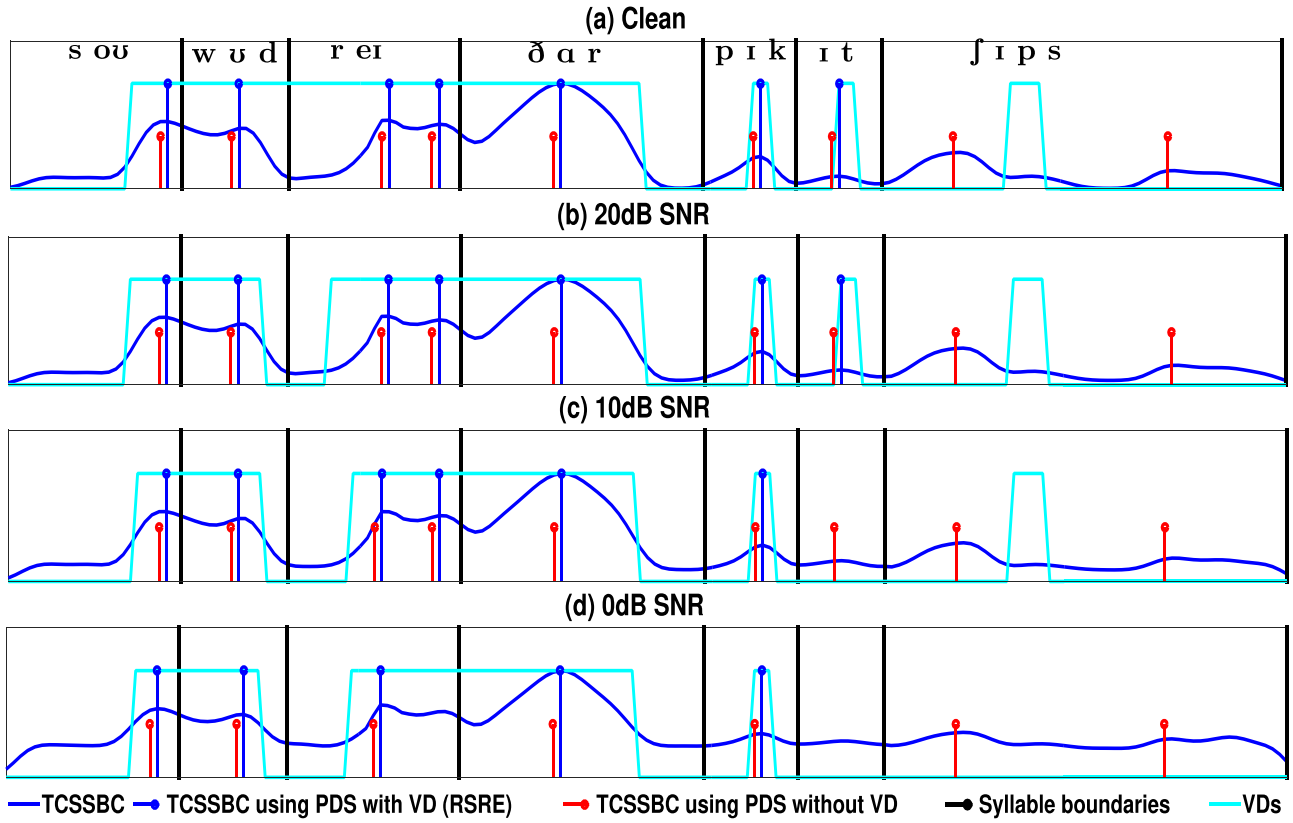

FIG. 1. (Color online) An illustrative example showing the insertion/deletion errors in TCSSBC by the PDS with and without VDs at 0, 10, and 20 dB SNR conditions.

To further investigate the effect of accuracy of the VDs in syllabic peak estimation, we estimate the TCSSBC peaks using the PDS of RSRE without using VDs, referred to as PDS without VDs. The shorter vertical red colored lines in Fig. 1 indicate the estimated peaks obtained from the PDS without VDs at clean as well as at 20, 10, and 0 dB SNRs. From the figure, it is observed that at 10 dB SNR, the peak in the syllable "ɪ t" is detected using the PDS without VDs; however, it is missed by the PDS with VDs due to incorrectly estimated VDs. This indicates that the TCSSBC peaks at low SNRs, which may be missed by the PDS with VDs, could be detected using the PDS without VDs. However, from Fig. 1, it is observed that the PDS without VDs introduces peaks in the unvoiced region of syllable "ʃ ɪ p s" in clean as well as all in three SNR conditions; this could be eliminated if VDs are considered. Thus, while the VDs are useful in syllable peak detection when they are accurate (e.g., in high SNR), they could be detrimental when they are erroneous (e.g., in low SNR).

We hypothesize that it could be beneficial to use VDs in PDS in an SNR dependent manner, since the PDS without VDs would work well in low SNRs while the PDS with VDs would provide accurate syllable peaks in high SNRs. From Fig. 1, it can be observed that both the PDSs (without and with VDs) fail to detect the syllabic peak in the syllable "ɪ t" at 0 dB SNR. This is because the TCSSBC does not have strong enough local maxima in that syllable at low SNRs. It is also observed that an extra TCSSBC peak is inserted in the syllable "r eɪ" by both the PDSs in clean as well as in 20 and 10 dB SNRs. This is because TCSSBC has two local maxima in that syllable. Both these unwanted variations could be due to the SNR dependent variations in the spectro-temporal structure in different sub-bands that are exploited to produce the TCSSBC peaks.[22]

We hypothesize that these variations could be fixed by pruning a set of non-informative sub-bands (referred to as pruned sub-bands) that distorts the expected spectro-temporal structure for obtaining the TCSSBC peaks in an SNR dependent manner. Hence, in general, the informative sub-bands could recover the missing syllabic peaks in low SNRs as well as eliminate the unwanted TCCSBC peaks at high SNRs, which, in turn, could be useful when the PDS without VDs and the PDS with VDs are used in an SNR dependent manner. We identify these pruned sub-bands in a data driven manner separately for each choice of the PDS—PDS with VDs and PDS without VDs.

In the proposed approach, a feature contour is computed with TCSSBC using a sub-set of sub-bands that is obtained by pruning the non-informative bands; here the sub-bands are computed based on the work by Huckvale.[30] The pruned sub-bands are selected according to the choice of PDS with VDs and the PDS without VDs, which is selected depending on the SNR of the signal. The pruned sub-bands for each PDS are obtained using a forward sub-band selection strategy. Experiments are performed on three corpora, namely, Switchboard,[31] TIMIT,[29] and CTIMIT.[28] In the case of Switchboard and TIMIT, the performance of the proposed method is analysed by simulating seen and unseen noisy test conditions by adding different noises at various SNRs similar to the experimental setup considered in most of the existing studies.[32–35] It should be noted that no noise is added to the utterances from CTIMIT as it has noisy recordings. For seen noisy conditions, we consider five noises, namely, babble, car (volvo), high frequency channel (hfc), F16 cockpit (f16), and white Gaussian (white) from the NOISEX-92 database,[36] and three SNRs: 20, 10, and 0 dB. For unseen noisy conditions, we consider four noises, namely, factory, sub-way, street, and exhibition from the Aurora database[37] and two SNRs: −5 and 5 dB. Compared to the best of the existing methods, the proposed method results in a higher correlation with the ground truth syllable rate for CTIMIT as well as for both the TIMIT & Switchboard corpora under clean condition. The proposed method also performs better in all seen and unseen noise and SNR conditions for both TIMIT & Switchboard.

II. DATABASE

We use ICSI Switchboard,[31] TIMIT,[29] and CTIMIT[28] corpora for all experiments in this work. Switchboard is a spontaneous speech corpus consisting of sentences spoken by 370 speakers with a wide range of speech rates, ranging from 1.26 to 9.2 syllables per second. The audio in the Switchboard corpus was collected through the telephone channel. A subset of 7300 audio segments, each of duration greater than 200 ms, is used for our experiments. In Switchboard, syllable transcriptions as well as their time aligned boundaries are available; however, phonetic transcription is not available. TIMIT is a read speech database, which has phonetically balanced 6300 sentences spoken by 630 speakers with a speech rate ranging from 1.44 to 8 syllables per second. All sentences from the TIMIT are used for our experiments. CTIMIT is similar to TIMIT except that the audio was collected through the cell phone channel under various noisy conditions. All 3370 sentences from the CTIMIT corpus, spoken by 630 speakers, are used for our experiments. The speech rate in the CTIMIT sentences ranges from 1.87 to 8 syllables per second. In TIMIT and CTIMIT, only phonetic transcriptions and their time aligned boundaries are available. Using these, we obtain syllable transcriptions and the corresponding time aligned boundaries with NIST syllabification software.[38] Following the work by Wang and Narayanan[15] for the experimentation, silent segments in the initial and final parts of each sentence of all corpora are removed. We use nine noises, namely, white, volvo, hfc, f16, babble, factory, sub-way, street, and exhibition in the experiments. The first five noises are from the NOISEX-92 database[36] and the remaining noises are from the Aurora database.[37] Babble noise has the most non-stationary characteristics among all the noises considered in this work.

III. PROPOSED APPROACH

The block diagram in Fig. 2 shows the steps involved in the proposed method. A given test audio signal goes through two stages in the proposed method: (a) feature computation and (b) peak detection strategy. The feature computation involves two steps. The first step prunes the

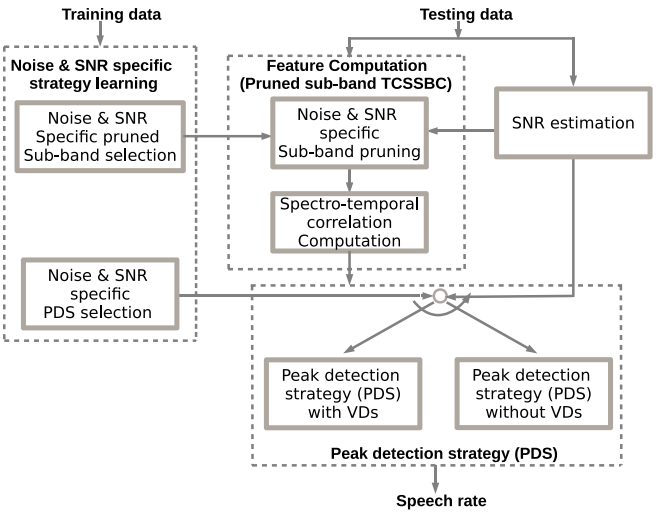

FIG. 2. (Color online) Block diagram of the proposed approach illustrating the steps involved in Noise and SNR specific strategy, Pruned sub-band TCSSBC computation, and PDS.

non-informative sub-bands, which are learnt during training, based on the noise and SNR condition of the test signal. In the second step, spectro-temporal correlation is used to compute the feature contour using the pruned sub-bands. The peak detection strategy selects either a PDS with VDs or PDS without VDs using the estimated SNR of the test signal. For selecting the PDS, we use an SNR based PDS selection criterion in which the parameters are learnt during the training stage.

A. Feature computation (Pruned sub-band TCSSBC)

The feature computation in the proposed method is based on the TCSSBC, computed using spectro-temporal correlation from 19 sub-band energy contours.[39] In order to overcome distortions in the TCSSBC in noisy conditions, we, in this work, propose a feature contour called pruned sub-band TCSSBC deduced from the TCSSBC by pruning sub-bands that are learnt during training. Below, we discuss the sub-band pruning technique following a brief description of spectro-temporal correlation and the motivation for sub-band pruning.

1. Spectro-temporal correlation

The typical steps involved in the spectro-temporal correlation computation are explained for a given set of K sub-band energy contours $y_1(n), y_2(n), \ldots, y_K(n)$, where n is the frame index, as below.

(1) The temporal correlation is computed on each sub-band energy contour with a window shift of one frame using a Gaussian window (w) of length J and variance σ^2 as follows:

$$z_i(n) = \frac{1}{J(J-1)} \sum_{j=0}^{J-2} \sum_{p=j+1}^{J-1} \{y_i(n+j)w(j)y_i(n+p)w(p)\} \quad \forall i = 1, 2, \ldots, K. \tag{1}$$

(2) At each frame (n), from all temporally correlated sub-bands, M highest energies are selected $z_{(1)}(n), z_{(2)}(n), \ldots, z_{(M)}(n)$ where $z_{(1)}(n) \geq z_{(2)}(n) \geq, \ldots, \geq z_{(M)}(n) \geq z_{(k)}(n); M+1 \leq k \leq K$. Using these M components, the sub-band correlation is computed as follows:

$$x(n) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} z_{(i)}(n) z_{(j)}(n). \tag{2}$$

The parameters (J, σ and M) in steps (1) and (2) are selected following the work by Wang and Narayanan.[15]

2. Motivation

In the proposed pruned sub-band TCSSBC feature contour, the non-informative sub-bands are eliminated before computing the spectro-temporal correlation. In this sub-section, we motivate the need for pruning such non-informative sub-bands using illustrative examples. Figure 3(a) shows the median of sub-band energies across all frames in the entire TIMIT corpus for each unvoiced phoneme. From the figure, it is observed that the energy values corresponding to the phonemes /tʃ/, /s/, and /ʃ/ are higher at the sub-band indices 15 and above compared to the remaining sub-bands. We observe that these higher energies result in unwanted TCSSBC peaks in unvoiced regions that do not correspond to the syllable nuclei, causing errors in the speech rate estimation. In order to avoid these unwanted peaks, VDs are typically used in TCSSBC based methods,[15,22,24,25] such as RSRE. Unlike using VDs, which could be inaccurate in the presence of noise, we propose to identify and eliminate non-informative sub-bands that cause unwanted peaks in the TCSSBC contour.

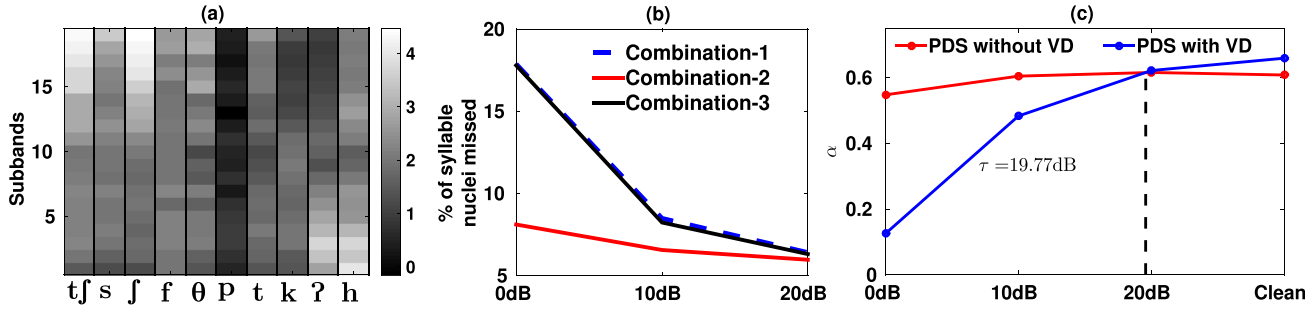

FIG. 3. (Color online) An illustrative example describing: (a) sub-band energy profiles of the unvoiced regions, (b) effect of VDs on RSRE performance under noisy conditions using three different TCSSBC and VD combinations: (1) both the TCSSBC and VDs are from the noisy signals, (2) the TCSSBC is from the noisy signal and the VDs are from the clean signal, and (3) the TCSSBC is from the clean signal and the VDs are from the noisy signal and (c) the threshold on SNR for choosing one of the PDSs (α is a metric evaluating the speech rate performance).

In order to examine the effects of errors due to inaccurate VDs and the lack of local maxima in TCSSBC around syllable nuclei, we perform a pilot experiment on speech rate estimation at three SNRS, namely, 20, 10, and 0 dB with additive white noise using the RSRE method[15] on the entire TIMIT corpus. Three different combinations of the TCSSBC and VDs are considered for the speech rate estimation. In the first combination, both the TCSSBC and the VDs are obtained from noisy signals. This setup is identical to that proposed by Wang *et al.*[15] In the second combination, the TCSSBC is obtained from the noisy signal and the VDs from the clean signal. In the third combination, the TCSSBC is obtained from the clean signal and the VDs from the noisy signal. Figure 3(b) shows the percentage of syllabic nuclei that is missed at all three SNRs for all the three combinations considered. From the figure, it is observed that the syllabic nuclei missed in the first and third combinations are more than the second combination. This indicates that the VDs from the noisy signal introduce more errors than the TCSSBC from the noisy signal. It is also observed that the first and third combinations have more errors at 0 dB than at 20 dB SNR. This implies that the errors due to inaccurate VDs increase with decreasing SNR.

We hypothesize that the speech rate estimation can be improved using the PDS without VDs by removing the dependency on VDs when they are inaccurate (i.e., at low SNR). We also hypothesize that in the absence of VDs, the unwanted TCSSBC peaks in unvoiced regions can be suppressed by pruning the sub-bands that contribute to the peaks in the unvoiced regions. In order to remove unwanted peaks in high SNR (e.g., syllable "r eɪ" in Fig. 1), we propose to perform sub-band pruning using the PDS with VDs. Similarly, for recovering the missing peaks in low SNR (e.g., syllable "ɪ t" in Fig. 1), we propose to perform sub-band pruning using the PDS without VDs. To determine the PDS used in sub-band pruning, we use the PDS obtained from a PDS selection criterion. The pruned sub-bands are learnt by maximizing either the speech rate estimation or the syllable nuclei detection performance.

3. Sub-band pruning

The pruned sub-bands that remove unwanted TCSSBC peaks could be noise and SNR specific. For example, in Fig. 3(a), it is observed that the high energy sub-bands that primarily cause TCSSBC peaks in unvoiced regions vary across the unvoiced phonemes /tʃ/ and /ʔ/. We observe that these high energy sub-bands vary across noise types as well as SNR conditions for a given unvoiced phoneme. Thus, in this work, we propose a method, called forward sub-band pruning, to identify noise and SNR specific non-informative sub-bands using both the PDSs separately. The respective sub-band sets for PDS without VDs and PDS with VDs are denoted by $S^{woVD}_{(\xi,\eta)}$ and $S^{wVD}_{(\xi,\eta)}$, where $\xi \in \{$babble, f16, hfc, volvo, white Gaussian$\}$ and $\eta \in \{0,\ 10,\ 20\,\text{dB},\ \text{clean}\}$. Further, at each SNR and noise combination, we choose a set based on the PDS obtained from the PDS selection criterion given the noise (ξ) and SNR (η) of the test signal. However, in an unknown noise and SNR condition, we identify the non-informative sub-bands by following steps in Sec. III C.

a. Forward sub-band pruning. All K sub-band energies are used in the forward sub-band pruning for each combination of noise (ξ), SNR (η) and PDS. The steps in the sub-band pruning are outlined in Algorithm 1, which takes the K sub-band energy contours as the input and returns pruned sub-bands and the corresponding maximum speech rate or syllable nuclei performance α. A full search for an optimal set of sub-bands would require $2^K - 1$ combinations, which is computationally infeasible. The forward sub-band pruning consider only $K(K+1)/2$ combinations, although it may result in a sub-optimal set of non-informative sub-bands.

ALGORITHM 1: Forward sub-band pruning – input: $Y = [Y_1, Y_2, \ldots, Y_K]$ (K sub-band energy contours), outputs: S (pruned sub-bands), α (highest speech rate or syllable nuclei performance).

(1): Initialization: $Y^s = \Phi$ (null vector). $\mathscr{P}, \mathscr{X}$ as empty vectors.
$\mathscr{I} = \{1, 2, \ldots, K\}$
(2): **for** $l = 1$ to K **do**
(3): Initialization: $\zeta = \Phi$
(4): **for** $i \in \mathscr{I}$ **do**
$x[n] \leftarrow$ compute TCSSBC using $[Y^s\ Y_i]$ following (1) and (2).
$\zeta_i \leftarrow$ Compute speech rate or syllable nuclei detection performance using TCSSBC $x[n]$
(5): **end for**
$\mathscr{P}_l \leftarrow \max_i \zeta_i$
$\mathscr{X}_l \leftarrow \arg\max_i \zeta_i$
$Y^s \leftarrow [Y^s\ Y_{\mathscr{X}_l}]$
$\mathscr{I} \leftarrow \mathscr{I} \backslash \mathscr{X}_l$
(6): **end for**
$T \leftarrow \arg\max_l \mathscr{P}_l$; $\alpha \leftarrow \max_l \mathscr{P}_l$
$S \leftarrow \mathscr{X}_{T+1:K}$; Return $\boldsymbol{S}, \boldsymbol{\alpha}$

Figure 4 shows the pruned sub-band TCSSBCs computed using the pruned sub-bands obtained from the PDS without VDs as well as the PDS with VDs (magenta & green colored contours respectively) and their respective detected peaks with the PDS without VDs & PDS with VDs for the exemplary sentence and noise used in Fig. 1. Compared to the TCSSBC in Fig. 1, Fig. 4 shows the improvements in the pruned sub-band TCSSBC that removes unwanted peaks in the syllables "ʃ ɪ p s" and "r eɪ" at all and high SNRs respectively as well as recovers the missing peak in the syllable "ɪ t" at low SNRs. However, all these improvements are not observed simultaneously in the pruned sub-band TCSSBCs using both the PDSs.

From Fig. 4, it is observed that the pruned sub-band TCSSBC from the PDS without VDs is effective in removing unwanted peaks in the syllable "ʃ ɪ p s" at all SNRs. However, at high SNRs (clean and 20 dB SNR), it fails to remove the extra peak in "r eɪ" syllable, as observed in Fig. 1. Such an unwanted peak is eliminated using the pruned sub-band TCSSBC from the PDS with VDs, for which the pruned sub-bands are learnt considering only the voiced regions using the PDS with VDs. Similarly, it is interesting that the pruned sub-band TCSSBC from the PDS with VDs fails to preserve the syllabic peak in "ʃ ɪ p s" syllable at 0 and 10 dB SNRs compared to the pruned sub-band TCSSBC from the PDS without VDs. This is because the pruned sub-band TCSSBC from the

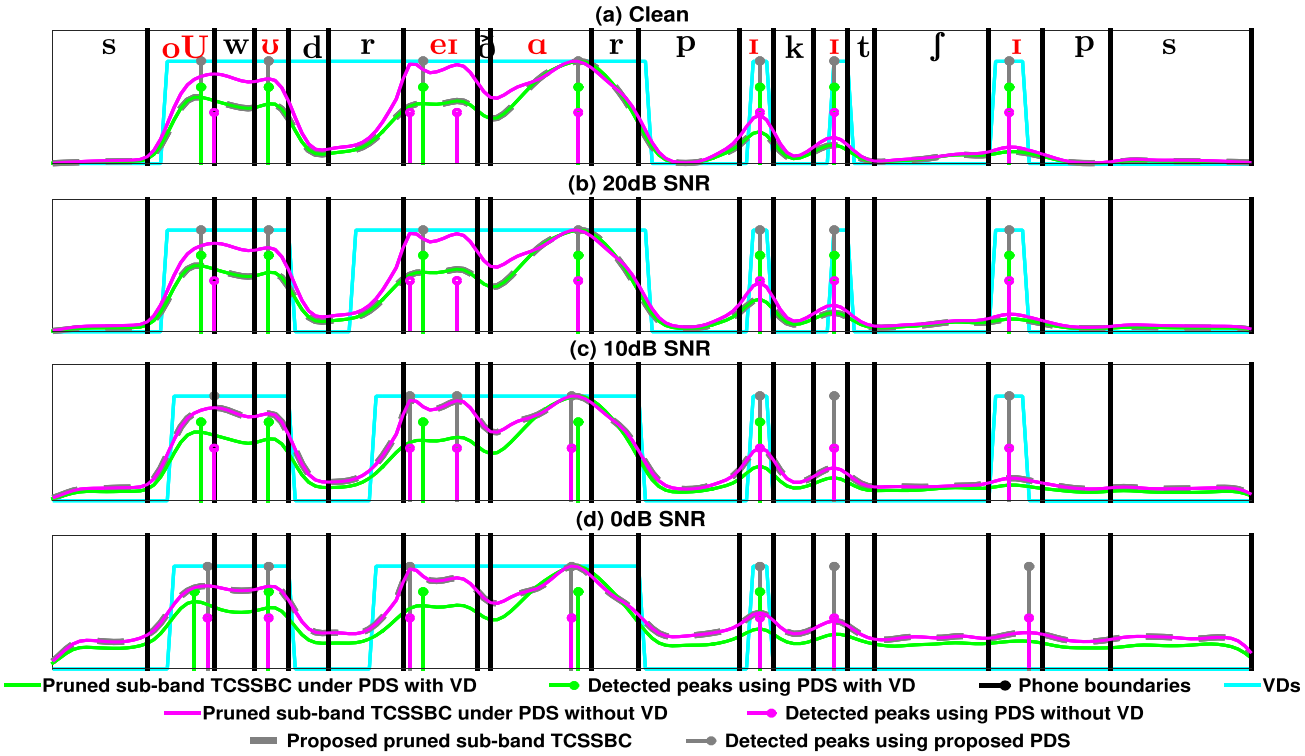

FIG. 4. (Color online) Illustrative example from Fig. 1 describing the benefit of the proposed PDS dependent sub-band pruning strategy and the PDS selection criterion (red color phonemes indicate the syllable nuclei), which remove unwanted peaks in the syllables "ʃ ɪ p s" and "r eɪ" and recover a missing peak in the syllable "ɪ t" compared to TCSSBC.

PDS with VDs does not have strong local maxima in low SNRs. These together suggest that the speech rate performance could be improved by the appropriate selection of one of the pruned sub-band TCSSBCs, which, in turn, depends on the PDS obtained from the PDS selection criterion.

B. Peak detection strategy

We examine the benefit of sub-band pruning using different PDSs on the speech rate estimation performance using a pilot experiment on the entire TIMIT corpus. This is illustrated for white Gaussian noise in Fig. 3(c), which shows $\alpha_{(\xi,\eta)}^{wVD}$ and $\alpha_{(\xi,\eta)}^{woVD}$ (speech rate performance) obtained on the TIMIT corpus in three SNRs and clean conditions (η = 0, 10, 20, clean) using blue and red colored contours respectively. In this case, ξ is the white Gaussian noise. Note that $\alpha_{(\xi,\eta)}^{wVD}$ is obtained by using the PDS with VDs on the pruned sub-band TCSSBC from the PDS with VDs. Similarly, $\alpha_{(\xi,\eta)}^{woVD}$ is obtained by using the PDS without VDs on the pruned sub-band TCSSBC from the PDS without VDs. It is clear that $\alpha_{(\xi,\eta)}^{wVD}$ is higher than $\alpha_{(\xi,\eta)}^{woVD}$ at 20 dB and clean conditions. This indicates that the PDS with VDs is useful at high SNRs. On the other hand, at low SNRs, $\alpha_{(\xi,\eta)}^{woVD}$ is higher than $\alpha_{(\xi,\eta)}^{wVD}$ indicating the benefit of the PDS without VDs in 0 and 10 dB. Hence, we combine their complementary advantages to propose a strategy for using the PDS with VDs and the PDS without VDs in an SNR dependent manner.

a. PDS selection criterion. Given a test signal with known noise (ξ) and SNR (η), the PDS without VDs and the PDS with VDs are selected depending on the value of SNR as follows:

$$\widehat{PDS} = \begin{cases} \text{PDS with VDs} & \textit{for } \eta > \tau_\xi \\ \text{PDS without VDs} & \textit{for } \eta \leq \tau_\xi, \end{cases} \tag{3}$$

where τ_ξ is a noise-specific SNR threshold that is determined based on the training set. In order to determine the threshold, we first compute $\alpha_{(\xi,\eta)}^{wVD}$ and $\alpha_{(\xi,\eta)}^{woVD}$ at three SNRs (20, 10, 0 dB) and clean conditions, as shown by the blue and red dots in Fig. 3(c) for example ξ being white Gaussian noise. We then determine the threshold by finding the SNR at which these two α vs SNR curves (obtained by linear interpolation) intersect each other. We observe that both the curves intersect at an SNR above 0 dB in both TIMIT and Switchboard corpora for all the noises considered in this work. In the white Gaussian noise case [as shown in Fig. 3(c)], the threshold turns out to be $\tau_\xi = 19.77$ dB.

Figure 4, along with Fig. 1, demonstrates the effectiveness of the SNR dependent PDS selection in the proposed method under known noise (ξ = white Gaussian) and SNR conditions compared to using the PDS with VDs (i.e., RSRE[15]) or the PDS without VDs on TCSSBC consistently in all SNRs. In Fig. 4, the grey colored dashed contour and vertical lines indicate the selected pruned sub-band TCSSBC and the peaks detected by $\widehat{PDS}$ in an SNR dependent manner. From the figure, it is observed that the proposed method detects the peaks correctly in every syllable (one peak in each syllable nuclei) in all SNRs except an extra peak in the syllable "r eɪ" at 10 dB SNR. These improvements in the proposed method could be due to the SNR dependent sub-band pruning and PDS, where it is expected to preserve the missing peaks as well as eliminate unnecessary peaks in the TCSSBC contour.

C. Sub-band pruning and PDS selection under unknown noise and SNR

For a test signal, in general, the noise and SNR may not be known *a priori*. We, in this work, propose a method for identifying the pruned sub-bands and computing the SNR threshold when either the noise or the SNR or both could be unknown. For this purpose, we follow the steps in the flow chart depicted in Fig. 5 for known and unknown noise and SNR combinations: (1) noise and SNR are known (denoted by N & S), (2) noise is known and SNR is unknown (N&S′), (3) noise is unknown and SNR is known (N′&S), and (4) noise and SNR are unknown (N′&S′). In the first step, we check whether ξ or η or both of them belong to their known values used in training. In the second step, the PDS is selected ($\widehat{PDS}$) using Eq. (3). The respective pruned sub-bands for each $\widehat{PDS} \in \{\text{PDS without VDs, PDS with VDs}\}$ are selected when the SNR or noise or both are unknown as follows:

$$
\begin{aligned}
&\Gamma^{woVD}_{(N,S)} = S^{woVD}_{(\xi,\eta)}; \quad \Gamma^{wVD}_{(N,S)} = S^{wVD}_{(\xi,\eta)} \\
&\Gamma^{woVD}_{(N,S')} = \bigcap_{\substack{\eta\ s.t. \\ \eta \le \tau_\xi}} S^{woVD}_{(\xi,\eta)}; \quad \Gamma^{wVD}_{(N,S')} = \bigcap_{\substack{\eta\ s.t. \\ \eta \le \tau_\xi}} S^{wVD}_{(\xi,\eta)} \\
&\Gamma^{woVD}_{(N',S)} = \bigcap_{\substack{\xi\ s.t. \\ \eta \le \tau_\xi}} S^{woVD}_{(\xi,\eta)}; \quad \Gamma^{wVD}_{(N',S)} = \bigcap_{\substack{\xi\ s.t. \\ \eta \le \tau_\xi}} S^{wVD}_{(\xi,\eta)} \\
&\Gamma^{woVD}_{(N',S')} = \bigcap_{\substack{\eta,\xi\ s.t. \\ \eta \le \tau_\xi}} S^{woVD}_{(\xi,\eta)}; \quad \Gamma^{wVD}_{(N',S')} = \bigcap_{\substack{\eta,\xi\ s.t. \\ \eta \le \tau_\xi}} S^{wVD}_{(\xi,\eta)}.
\end{aligned}
\tag{4}
$$

When the noise or SNR or both are unknown, the intersection of several pruned sub-band sets is chosen, i.e., common sub-bands among the pruned sub-band sets to preserve informative sub-bands. However, the parameters η and τ_ξ in Eq. (3) change according to the known and unknown conditions of noise and SNR. In the case of unknown noise, the τ_ξ is replaced with $\widehat{\tau}$, which is obtained by averaging all the τ_ξ belonging to the noises considered in this work; however, when the noise is known, τ_ξ is used directly. Similarly, in the case of unknown SNR, estimated SNR ($\widehat{\eta}$) value is used in place of η. The SNR is estimated on a test utterance following the work by Gerkmann *et al.*[40,41]

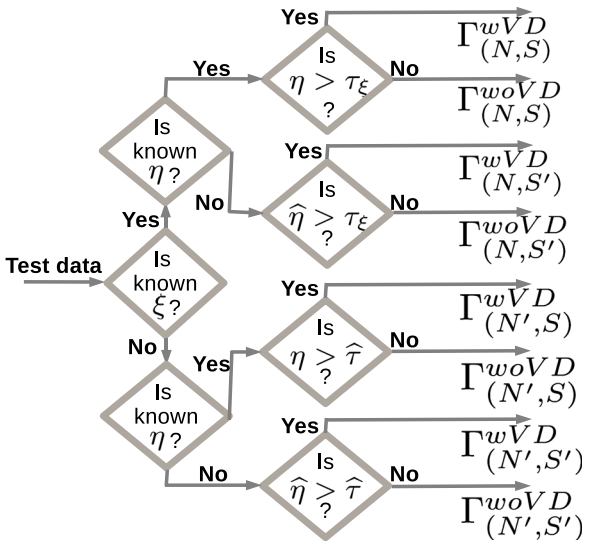

FIG. 5. (Color online) Flow chart for the selection of pruned sub-bands under known and unknown noise & SNR combinations: (1) noise and SNR are known (denoted by N&S), (2) noise is known and SNR is unknown (N&S′), (3) noise is unknown and SNR is known (N′&S), and (4) noise and SNR are unknown (N′&S′).

D. Experimental setup

We consider Pearson correlation coefficient[42] (ρ) between the estimated syllable rate and the ground truth syllable rate across all test sentences as an objective measure for evaluating the speech rate estimation. We consider the RSRE[15] and DSRE[25] techniques as baselines in our experiments. Following the TCSSBC computation in RSRE, we consider a set of $K = 19$ sub-band energy contours computed using the method given by Huckvale[30] with non uniform sub-bands as described by Holmes.[39] Experiments are performed on data from CTIMIT, TIMIT, and Switchboard. For parameter learning, we consider the TIMIT and Switchboard data under clean and noisy conditions with additive noises namely, white Gaussian, volvo, hfc, f16, and babble at 0, 10, and 20 dB SNRs. The data in the clean and each noisy conditions is divided into three parts randomly: 10% as the training set for learning the pruned sub-bands, 40% as the development set for learning threshold (τ_ξ) and pruned sub-bands $S^{woVD}_{(\xi,\eta)}$ and $S^{wVD}_{(\xi,\eta)}$, and the remaining 50% as the test set. For TIMIT and Switchboard, we learn the parameters separately for each corpora, which are used for the respective test sets. Although, during training, the threshold and pruned sub-bands are learnt in a noise and SNR specific manner, the known and unknown noise (N and N') and SNR (S and S') conditions in the test sets are simulated by using the threshold and pruned sub-bands as illustrated in Fig. 5. This is done to show the effectiveness of the proposed method under all known and unknown conditions of noise and SNR seen during parameter learning. In order to know the performance of the proposed method under noise and SNR conditions unseen in training, we consider CTIMIT as well as TIMIT and Switchboard data under noisy conditions with four additive noises namely, factory, sub-way, street, and exhibition at −5 and 5 dB SNRs. In addition to ρ, we consider F-score measure for learning the parameters: threshold and pruned sub-bands. F-score is computed following the work by Landsiedel *et al.*[27] and Yarra *et al.*[24] We hypothesize that learning the parameters using F-score could have an advantage over those using ρ. This is because the ρ computation depends only on the estimated and actual number of syllables in the test sentences. On the other hand, F-score depends on the accuracy of the estimated syllable nuclei locations with respect to the actual ones.

E. Hyper-parameter optimization

1. SNR threshold

Table I shows the estimated τ_ξ values using ρ and F-score for TIMIT and using ρ for Switchboard. In the case of Switchboard, τ_ξ is learnt using only ρ since the phone boundaries (needed for syllable nuclei information for F-score computation) are not available. From the table, it is

TABLE I. Noise specific SNR threshold values for TIMIT and Switchboard corpora for all five noise conditions, using ρ and F-score measures on TIMIT and using ρ measure on Switchboard.

		white	volvo	hfc	f16	babble
TIMIT	using ρ	19.91	19.36	21.26	19.30	13.86
	using F-score	19.77	18.21	19.71	17.55	16.27
Switchboard	using ρ	12.18	9.58	12.33	9.88	9.25

observed that among all noises, τ_ξ is the least in babble noise for both measures in both corpora. This could be because the babble noise has a highly time varying (non-stationary) spectrum similar to speech.[43] The non-stationary babble noise spectrum could alter the spectro-temporal correlation more than other noises to varying degree depending on the SNR. This, in turn, could result in a lower crossover point between the performance vs SNR curves for the PDS without VDs and the PDS with VDs schemes [Fig. 3(c)]. Similarly, the spontaneous speech in the Switchboard corpus may be less structured,[44,45] resulting in a different spectro-temporal correlation compared to the read speech in TIMIT; this, in turn, could result in a lower τ_ξ value for Switchboard compared to TIMIT for all noises considered in this work.

2. Pruned sub-bands

Figure 6 shows the pruned sub-bands ($S^{woVD}_{(\xi,\eta)}$ and $S^{wVD}_{(\xi,\eta)}$) obtained using Algorithm 1 for five noises in three SNRs and clean conditions considering two performance measures—ρ and F-score—for TIMIT and using ρ for Switchboard. It should be noted that the pruned sub-bands are used for speech rate estimation only for some combinations of noise, SNR and PDS. This is because both the PDSs are not used for every combination of noise and SNR; rather, it depends on the PDS selection criterion. In Fig. 6, the ✓ marks indicate such used combinations. In the case of unknown noise and SNR, the pruned sub-bands $\Gamma^{woVD}_{(N',S')}$ and $\Gamma^{wVD}_{(N\prime,S\prime)}$, obtained using Eq. (4), are also shown in the figure.

From the selected set of pruned sub-bands for TIMIT [Figs. 6(a) and 6(b)] corresponding to the ✓ marks, it is observed that the sub-bands 16–19 are selected as pruned sub-bands in all five noises at all three SNRs for both the measures except for hfc, f16, and babble at 0 dB SNR under ρ measure as well as for babble at 0 dB SNR under F-score measure. On the other hand, sub-bands 1–3 are detected as informative bands in all five noises at all three SNRs for both the measures except at 0 and 10 dB under babble noise for ρ measure as well as volvo and bable noises at 0 dB SNR and f16 noise at 20 dB SNR for F-score measure. This could be because the first formant typically falls in the frequency range of the sub-bands 1–3. Also, the sub-bands 16–19 have frequency ranges beyond the typical locations of third formants.[46] However, such a consistency in pruned sub-bands is not seen in the case of Switchboard [Fig. 6(c)] probably due to the spontaneous nature of speech, unlike the read speech in TIMIT, which is more structured;[44,45] hence, it has less spectral variability than the spontaneous speech in Switchboard.

F. Results

Table II shows the ρ values computed on the test sets of TIMIT and Switchboard using the two baselines as well as

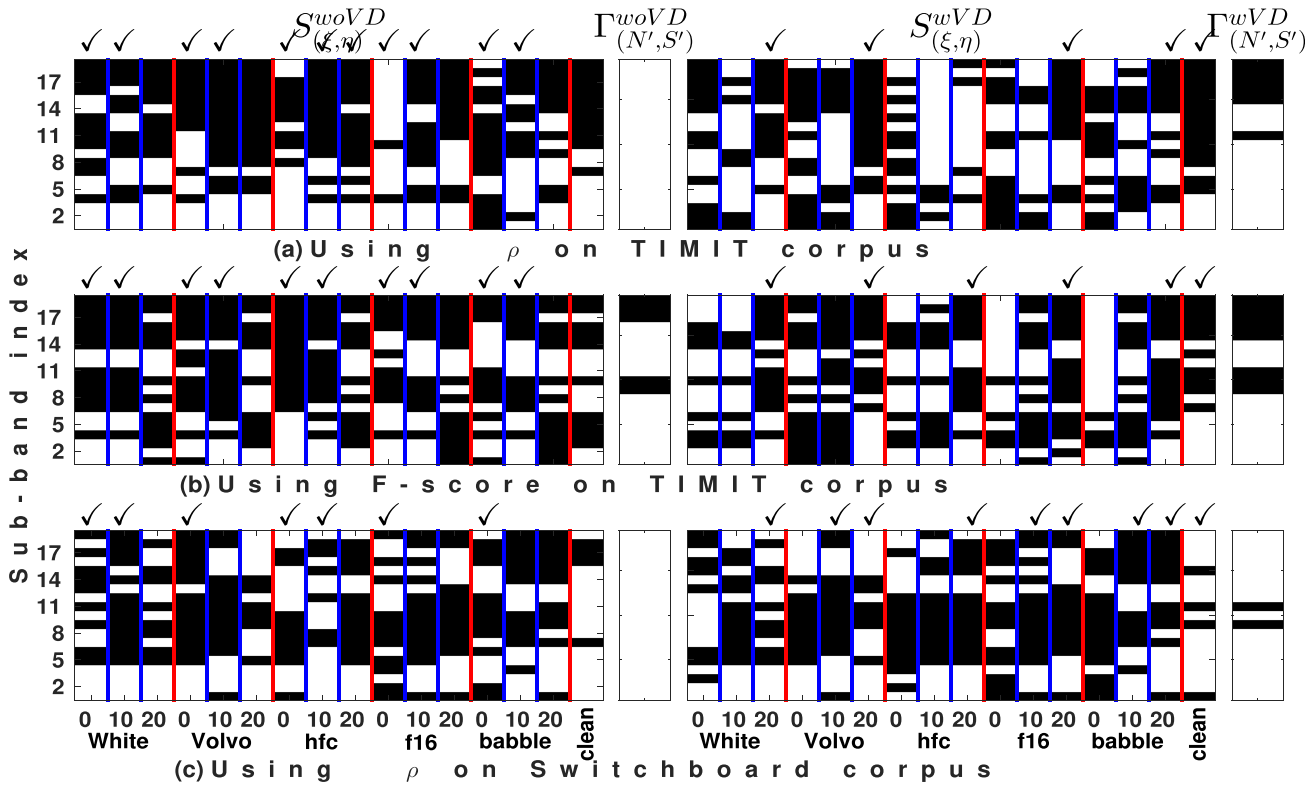

FIG. 6. (Color online) Pruned sub-bands, ($S^{woVD}_{(\xi,\eta)}$ and $S^{wVD}_{(\xi,\eta)}$), (black colored regions) for (a) TIMIT using performance measure as ρ, (b) TIMIT using F-score, and (c) Switchboard using ρ under both known and unknown ξ & η. $\Gamma^{woVD}_{(N',S')}$ and $\Gamma^{wVD}_{(N',S')}$ indicate the pruned sub-bands obtained in the case of unknown noise and SNR using Eq. (4). The ✓ marks marks correspond to the PDS selected for every combination of noise and SNR separately in (a), (b), and (c).

TABLE II. Correlation coefficient (ρ) obtained on the test sets of TIMIT and Switchboard under clean condition considering the parameters learnt separately using ρ and F-score measures.

	TIMIT				Switchboard		
			Proposed				Proposed
	RSRE	DSRE	using ρ	using F-score	RSRE	DSRE	using ρ
Clean	0.6639	0.6704	0.6756	**0.6803**	0.6627	0.6486	**0.6629**

the proposed method with SNR threshold and pruned sub-bands obtained using both ρ and F-score measures. The proposed scheme is executed assuming unknown noise and SNR (i.e., N′&S′ combination). The bold entries in the table indicate the highest ρ values among different schemes for each corpus. From the table, it is observed that the ρ values obtained using the proposed method are higher than those with both the baselines on TIMIT, while the ρ values are comparable on Switchboard. It is also observed that for TIMIT, the ρ value is higher when sub-band pruning is performed using F-score compared to that using ρ, indicating the robustness of the F-score for sub-band pruning.

a. Performance under seen noise conditions. Table III shows the ρ values obtained using the two baselines as well as the proposed method for test cases with additive noise for five seen noises at three seen SNRs. In the table, the bold entry for every corpus in each row indicates the highest ρ value for each noise and SNR combination. From the table, it is observed that the estimated ρ values by the proposed method are more than those by the RSRE and DSRE methods for both TIMIT and Switchboard corpora at all noises and SNR combinations. In the case of TIMIT corpus, the ρ values obtained based on F-score measure are higher than those based on ρ for five noises at three SNRs, except for the white noise at 10 dB SNR and f16 noise at 0 dB SNR. This indicates that the parameters obtained using F-score measure are more robust than those using ρ. This observation is consistent with the observations in the clean condition (Table II). When averaged across all noises, the proposed method (with parameters optimized using both ρ and F-score) performs significantly ($p < 0.01$ with t-test) better than the best of the baseline schemes (in most cases RSRE performs better than DSRE) at all three SNRs on TIMIT. However, on Switchboard where ρ measure is used for selecting pruned sub-bands, the proposed method performs significantly ($p < 0.01$ with t-test) better only at 0 dB SNR, while at the remaining SNRs, there is no significant difference between the performance of the best baseline and the proposed method. This suggests that the proposed method achieves a better speech rate estimation accuracy over both the baselines through the selection of PDS in an SNR specific manner.

From Table III, it is observed that the increase in ρ values from the proposed method over those from the RSRE varies across SNRs. These improvements are summarized using bar-plots in Fig. 7, representing the percentage of improvement in the difference between ρ values ($\rho^{Proposed} - \rho^{RSRE}$) in the 6th and 3rd columns and the 9th and 7th

TABLE III. Correlation coefficient (ρ) obtained on the test sets of TIMIT and Switchboard with five noises and three SNR conditions considering the parameters learnt separately using ρ and F-score measures.

		TIMIT				Switchboard		
				Proposed				Proposed
		RSRE	DSRE	using ρ	using F-score	RSRE	DSRE	using ρ
White	0 dB	0.1166	0.1058	0.5381	**0.5609**	–0.0795	–0.0911	**0.4776**
	10 dB	0.4880	0.4900	**0.6359**	0.6239	0.5310	0.5233	**0.5326**
	20 dB	0.6364	0.6439	0.6435	**0.6540**	0.6456	0.6409	**0.6476**
Volvo	0 dB	0.3009	0.2859	0.6131	**0.6140**	0.4104	0.3358	**0.5757**
	10 dB	0.5842	0.5774	0.6183	**0.6229**	0.6239	0.6110	**0.6249**
	20 dB	0.6571	0.6545	0.6551	**0.6665**	0.6486	0.6408	**0.6496**
hfc	0 dB	0.1307	0.1155	0.4102	**0.4130**	0.0011	–0.0187	**0.4028**
	10 dB	0.4880	0.4852	0.6107	**0.6138**	0.5655	0.5584	**0.5677**
	20 dB	0.6360	0.6391	0.6417	**0.6596**	0.6403	0.6446	**0.6450**
f16	0 dB	0.1616	0.1489	**0.4100**	0.3997	0.0282	0.0006	**0.4179**
	10 dB	0.4997	0.4951	0.6059	**0.6179**	0.5471	0.5402	**0.5489**
	20 dB	0.6413	0.6421	0.6433	**0.6566**	0.6467	0.6366	**0.6475**
Babble	0 dB	0.1962	0.1810	0.2504	**0.2732**	0.1030	0.0603	**0.2218**
	10 dB	0.5056	0.5036	0.5225	**0.5306**	0.5503	0.5444	**0.5524**
	20 dB	0.6451	0.6440	0.6453	**0.6595**	0.6502	0.6348	**0.6532**
Average	0 dB	0.1812	0.1674	0.4465	**0.4521**	0.0926	0.0574	**0.4191**
		(0.0735)	(0.0649)	(0.1387)	(0.1362)	(0.1892)	(0.1474)	(0.1295)
	10 dB	0.5131	0.5103	0.6018	**0.6019**	0.5635	0.5555	**0.5653**
(SD)		(0.0404)	(0.0341)	(0.0456)	(0.0400)	(0.0358)	(0.0300)	(0.0355)
	20 dB	0.6431	0.6447	0.6457	**0.6593**	0.6462	0.6395	**0.6488**
		(0.0086)	(0.0052)	(0.0053)	(0.0046)	(0.0038)	(0.0035)	(0.0030)
Overall Average		0.4458	0.4408	0.5647	**0.5711**	0.4341	0.4175	**0.5443**

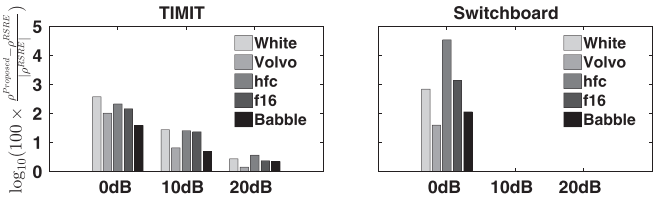

FIG. 7. Improvement in the performance of the proposed method with respect to RSRE for five noises and three SNRs. The improvement is shown using percentage change in the ρ values with respect to $|\rho^{RSRE}|$.

columns in Table III with respect to ρ^{RSRE}. It is found that the absolute improvements in ρ using the proposed method are more in low SNR compared to high SNR for each noise in case of both the corpora. We observe that the estimated SNR values for test signals at 0 dB SNR are below the average SNR threshold ($\widehat{\tau}$) for all five noises in both corpora. Thus, the proposed method selects the PDS without VDs for peak detection at 0 dB SNR. This indicates that the proposed method accurately detects several syllabic peaks that are missed by RSRE due to inaccurate VDs at 0 dB SNR. It is also observed that the absolute improvement in ρ value is the highest for white Gaussian noise and the lowest for the babble noise at 0 dB SNR. This could be because white noise has equal energy per unit bandwidth.[47] Hence, it has a consistent structure across all frames compared to the non-stationary babble noise, resulting in more effective pruned sub-bands in white noise compared to those in babble noise.

From Table III, it is observed that the improvement in speech rate estimation by the proposed method is more for TIMIT compared to that for Switchboard when averaged across all noises and SNRs. One of the reasons for such performance difference could be the robustness of F-score in learning pruned sub-bands, which is used for TIMIT but not for the Switchboard corpus. It could also be due to less consistency in the pruned sub-bands because of the spontaneous nature of speech in Switchboard, which results in less suppression of peaks in the unvoiced regions in Switchboard compared to the TIMIT corpus. More incorrect selection of PDS and pruned sub-bands in Switchboard due to wrongly estimated SNR could be another reason for the performance difference between the TIMIT and Switchboard corpora. For example, Table IV shows the percentage of confusion between the (estimated) PDS obtained using the PDS selection criterion under unknown noise and SNR conditions and the (expected) PDS obtained assuming that the noise and SNR of the test sentence are known. The percentages in the table also denote the confusion between the estimated and expected pruned sub-band sets. From the table, it is observed that the confusions are more in Switchboard (0.03 and 49.04) than those in the TIMIT (0.69 and 0.00) corpus, which, in turn, could cause poor performance of the proposed method in the Switchboard compared to that in the TIMIT corpus. We observe that the majority of confusion in the Switchboard corpus is due to an incorrect PDS selection for test signals at 10 dB SNR. This is because the average SNR threshold $\widehat{\tau}$ for Switchboard is 10.64 dB. While the PDS without VDs would be the right choice for a test signal at 10 dB SNR (as $10 < 10.64$), a positive SNR estimation error more than 0.64 dB would result in the selection of PDS with VDs, which, in turn, could drop the performance of the proposed method.

TABLE IV. Percentage of confusion among estimated (rows) and expected (columns) PDSs due to errors in SNR estimation. The estimated PDSs are obtained using the PDS selection criterion under unknown noise and SNR conditions. The expected PDSs are obtained assuming that the noise and SNR of the test sentence are known.

	TIMIT		Switchboard	
	PDS without VDs	PDS with VDs	PDS without VDs	PDS with VDs
PDS without VDs	100.00	0.69	50.96	0.03
PDS with VDs	0.00	99.31	49.04	99.97

Further, from Table IV, it is clear that the proposed method mostly uses PDS with VDs (similar to RSRE) in the cases of both TIMIT and Switchboard at clean and 20 dB SNR conditions under all five noises. Thus, higher ρ values with the proposed method on both the corpora in Table II at clean and in Table III at 20 dB SNR condition compared to those with RSRE suggest the effectiveness of sub-band pruning in the proposed method. Similarly, on TIMIT, the proposed method uses PDS without VDs at 0 and 10 dB SNR conditions. Hence, the higher ρ values with the proposed method in Table III at 0 and 10 dB SNRs compared to those with RSRE are due to both sub-band pruning and VDs removal. However, from Fig. 6, it is observed that in the case of PDS without VDs, the pruned sub-bands are non-empty only when those are learnt using F-score measure. Thus, improvements in the cases using ρ measure are only due to the removal VDs, suggesting the effectiveness of VDs removal in the proposed method. Further, higher ρ values using F-score measure compared to those using ρ measure suggest the effectiveness of sub-band pruning in the 0 and 10 dB SNRs.

b. Performance under unseen noise conditions. Table V shows the ρ values obtained on the test sets of TIMIT and Switchboard under four unseen noises at -5 and 5 dB SNR conditions considering the parameters learnt separately using ρ and F-score measures. From the table, it is observed that the ρ values obtained with the proposed method are higher than those with both the baselines in all cases except in exhibition noise under 5 dB SNR condition on the Switchboard corpus. This indicates the benefit of the proposed method even under unseen conditions. Further, we observe that the PDS without VDs is selected for all four noises at both unseen SNRs on TIMIT. Hence, under this condition, the improvements in ρ values over RSRE for the TIMIT corpus indicate the benefit of the VDs removal in the proposed method, when the parameters learnt using ρ are used. The ρ values obtained using the parameters learnt with F-score are higher than those using the parameters learnt with ρ. This indicates the benefit of only sub-band pruning because the pruned sub-bands (as observed in Fig. 6) are empty in the case of learning with ρ.

Table VI shows the ρ values obtained on CTIMIT using the two baselines and the proposed method separately

TABLE V. Correlation coefficient (ρ) obtained on the test sets of TIMIT and Switchboard with four unseen noises and two unseen SNR conditions, considering the parameters learnt using ρ and F-score measures separately.

		TIMIT				Switchboard		
				Proposed				Proposed
		RSRE	DSRE	using ρ	using F-score	RSRE	DSRE	using ρ
Factory	−5 dB	−0.0225	−0.0335	0.3635	**0.3723**	0.0406	0.0256	**0.3620**
	5 dB	0.3870	0.3723	0.5218	**0.5564**	0.4636	0.4370	**0.5125**
Sub-way	−5 dB	0.0354	0.0191	0.3232	**0.3348**	0.0225	0.0069	**0.2678**
	5 dB	0.4115	0.4003	0.4972	**0.5072**	0.4432	0.4392	**0.4731**
Street	−5 dB	0.1700	0.1567	0.3415	**0.3546**	0.1710	0.1156	**0.3991**
	5 dB	0.4909	0.4435	0.5132	**0.5278**	0.4960	0.4736	**0.5037**
Exhibition	−5 dB	0.0253	0.0139	0.3473	**0.3625**	0.0313	0.0059	**0.2431**
	5 dB	0.3739	0.3651	0.5283	**0.5416**	**0.4756**	0.4563	0.4681
Average	−5 dB	0.0520	0.0391	0.3439	**0.3561**	0.0663	0.0385	**0.3180**
		(0.0826)	(0.0820)	(0.0166)	(0.159)	(0.0701)	(0.0522)	(0.0745)
(SD)	5 dB	0.4158	0.3953	0.5151	**0.5333**	0.4696	0.4515	**0.4893**
		(0.0524)	(0.0355)	(0.0135)	(0.0209)	(0.0221)	(0.0170)	(0.0221)
Overall Average		0.2339	0.2172	0.4295	**0.4447**	0.2680	0.2450	**0.4037**

considering the parameters (pruned sub-bands and τ) learnt from TIMIT using F-score, TIMIT using ρ, and Switchboard using ρ. The table also shows the ρ values averaged across all seen noises and all three seen SNRs and clean condition on TIMIT and Switchboard under both matched and mismatched train-test conditions. This is repeated by averaging across all unseen noises and two unseen SNRs. In the table, the rows indicate the test corpus considered and the columns indicate the training conditions from which the parameters are obtained. Blue color entries in the table mark the ρ values obtained under mismatched condition. Higher ρ values with the proposed method in all conditions over those with the two baseline schemes indicate the merit of the proposed method. We observe improvements in the ρ values on Switchboard with the parameters from TIMIT using F-score over those using ρ and those with RSRE. Improvements in these two cases respectively suggest the advantages of sub-band pruning only and sub-band pruning, plus VDs removal even under mismatched train-test conditions. Further, it is interesting to observe that under both seen and unseen conditions, ρ values are improved on Switchboard when the parameters learnt from TIMIT data are used compared to the case when Switchboard data are used. On the other hand, on TIMIT, the ρ values are lower under mismatched condition compared to those under matched condition. Overall, the results in the table indicate that the proposed method estimates speech rate more consistently within and across corpora when parameters are learnt using F-score.

Unlike additive noise conditions, CTIMIT contains noisy speech that is recorded in realistic scenarios and we observe that the estimated SNRs for the recordings in CTIMIT range from 7.36 to 35.17 dB with mean and standard deviation (SD) of 17.48 and 4.48 dB, respectively. From Tables II and VI, it is observed that the performance on the CTIMIT (noisy version of TIMIT) degrades compared to that on TIMIT for both the baselines and the proposed method. However, the degradation is less in the proposed method. In particular, the ρ values obtained using the proposed method, separately with all three sets of parameters, are higher than those using both the baseline schemes. This indicates that the proposed method is robust even under adverse real noisy scenarios. Among all three sets of parameters, the highest ρ values are obtained with the parameters using F-score. The improvements in the ρ values with the parameters from TIMIT using F-score compared to those from Switchboard using ρ are mostly due to sub-band pruning only.

G. Discussions

1. Noise and SNR specific analysis

The results in Tables II and III are computed when the noise type and SNR of the signal are unknown (denoted by N′&S′). We also investigate the performance of the proposed method in the following known and unknown noise and SNR combinations—(1) N&S, (2) N&S′, and (3) N′&S. The selection of pruned sub-bands and PDS is done according to Fig. 5. Table VII shows the average of ρ values across all five noises at each SNR for the TIMIT and Switchboard corpora. In the case of TIMIT, we consider the ρ values obtained from the parameters learnt using F-score. From the table, it is observed that the average ρ values under N&S, N&S′, N′&S, and

TABLE VI. Performance of the proposed method on CTIMIT as well as TIMIT and Switchboard under matched and mismatched train-test cases in seen and unseen noise conditions.

				Proposed		
				TIMIT		Switchboard
Noise condition	Test corpora	RSRE	DSRE	using ρ	using-Fscore	using ρ
Unseen	TIMIT	0.2339	0.2172	0.4295	**0.4447**	0.4237
	Switchboard	0.2680	0.2450	0.4244	**0.4391**	0.4037
Seen	TIMIT	0.4574	0.4529	0.5924	**0.5984**	0.5784
	Switchboard	0.4582	0.4296	0.5788	**0.5867**	0.5740
CTIMIT		0.2764	0.2892	0.4017	**0.4178**	0.3696

TABLE VII. Average of the ρ values across all five noises with the proposed method under the following known and unknown noise and SNR combinations—(N&S), (N&S′), (N′&S), and (N′&S′). In the case of TIMIT, we consider the parameters learnt using F-score only. Average values of the ρ across all five noises with the RSRE are the same under all four conditions, and those are 0.1812, 0.5131, and 0.6431 and 0.0926, 0.5635, and 0.6462 at 0, 10, and 20 dB SNRs, respectively, on TIMIT and Switchboard corpora.

	TIMIT			Switchboard		
	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB
N&S	0.4744	0.6025	0.6580	0.4302	0.5566	0.6299
N&S′	0.4545	0.6031	0.6582	0.4197	0.5457	0.6299
N′&S	0.4568	0.6018	0.6592	0.4133	0.5759	0.6486
N′&S′	0.4521	0.6018	0.6593	0.4192	0.5653	0.6488

N′&S′ do not vary much among themselves in most of the cases on both corpora. This indicates that a knowledge of the noise and SNR specific information does not improve the performance significantly, which, in turn, suggests that the proposed method is robust to unknown noisy conditions under all five noises considered in this work. Similarly, it also suggests that the proposed method is robust to the estimation errors reported in Table IV in selecting the PDS.

We further investigate the variation in performance of the proposed approach with respect to the variations in the threshold (τ) used in PDS selection. Figure 8 shows the ρ vs τ plots under all five known noise conditions as well as under unknown noise condition. In the case of the known noise condition, the ρ value is computed using the estimated speech rates belonging to all three SNR and clean conditions. However, in the case of an unknown noise, we use the estimated speech rates belonging to all five noises at all three SNR and clean conditions for computing ρ. In the figure, we indicate a range (shown using color-filled rectangles in Fig. 8) of τ values under each known and unknown noise condition that corresponds to ρ values that lie above 99% of the highest ρ value. Also, in the figure, the color-filled circles indicate the τ_ξ reported in Table I using F-score for all five noises and the respective $\widehat{\tau}$ in the unknown noise case. From the figure, it is observed that the τ_ξ and $\widehat{\tau}$ values fall within their respective ranges for unknown noise as well as all five noises except for white and hfc. This indicates that the estimated τ_ξ and $\widehat{\tau}$ values are robust to unknown noise as well as most of the seen noises considered in this work. Further, considering the ρ value obtained using 40% of the data as a development set, we find that the ρ values are within $\pm 5\%$ when the size of the development data is varied from 5% to 40% in steps of 5%

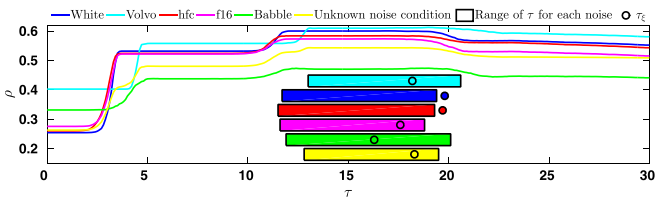

FIG. 8. (Color online) Effect of variation in threshold (τ) on the performance of the proposed method for the TIMIT corpus under all five known noise conditions as well as unknown noise condition.

under all five noise conditions as well as under unknown noise condition.

2. Phoneme-specific analysis

Sub-band pruning in the proposed method changes the strength of local maxima in the pruned sub-band TCSSBC from that in the TCSSBC at both consonant and vowel (syllable nuclei) phonemes. Typically, at each syllable, the peaks detected by the proposed method depend on the vowel peak strength with respect to its neighboring phoneme peak strength as well as on the PDS. When the PDS with VDs is used, typically at 20 dB and clean conditions, it is expected that both RSRE and the proposed method have the similar performances. However, from Tables II and III, it is observed that the proposed method performs better than RSRE. Figures 9(a) and 9(b) show exemplary syllable segments ("ʃɪns" and "zɪk") to illustrate the reason for improvements in the proposed method. From the figures, it is observed that the proposed method neither misses the syllable peak nor inserts extra peaks in the syllables; however, the RSRE misses the syllable peaks. This could be because, in both segments ("ʃɪns" and "zɪk"), sub-band pruning results in a reduction of the height of the valley before the phoneme "ɪ" in addition to reduction of the peak strengths in the respective phonemes "ʃ" and "z" compared to those in TCSSBC. The changes in valley heights and peak strengths in the pruned sub-band TCSSBC from TCSSBC would affect peak detection using the PDS with VDs;[15] this leads to the correct syllable peak detection by the proposed method.

We observe that the improvements in correct syllable peak detection by the proposed method are found to be the highest in the syllables containing the phoneme "ɪ," which belongs to the short vowel category. This is observed in clean and all three SNR conditions. On the other hand, the least improvement is found in the syllables containing the phoneme "ɑ," which belongs to the long vowel category. It is also interesting to observe that the six highest improvements are found in the syllables containing the vowels "ə," "ɝ," "ɪ," "i," "ʊ," and "u" out of which "ə," "ɝ," "ɪ," and "ʊ" are short vowels. On average, we observed that the improvement in correct syllable peak detection by the proposed method is higher in the syllables containing short vowels than those containing long vowels at all SNRs. This could be because short vowels are largely affected due to inaccurate VDs mostly at low SNRs; hence, RSRE misses most of the syllable peaks belonging to the syllables containing those vowels. On the other hand, in general, the proposed method introduces extra peaks within the syllables containing long vowels. Figures 9(c) and 9(d) show the two exemplary syllable segments "teɪlz" and "saɪd," containing extra peaks detected by the proposed method. From the figures, it is observed that the extra peaks are detected due to the valley at the boundary of the phoneme "eɪ" in the syllable "teɪlz" and that in the middle of the phoneme "aɪ" in the syllable "saɪd."

In this work, analysis on the proposed method is primarily performed in a controlled manner under additive noisy conditions. The improvements under unseen additive noisy

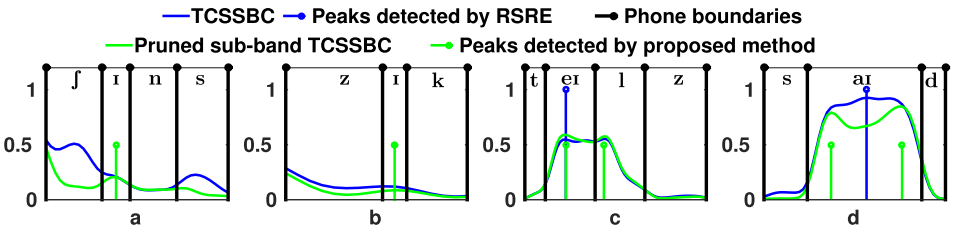

FIG. 9. (Color online) Illustrative exemplary syllable segments for describing variability in the peaks detected by the proposed method and RSRE. (a) ʃɪns, (b) zɪk, (c) teɪlz, and (d) saɪd.

conditions and mismatched train-test conditions indicate the usefulness of the proposed method in handling a range of additive noises.[34] Further, the improvements on CTIMIT suggest the benefit of the proposed method in realistic situations when the target data is phonetically and/or linguistically similar to TIMIT. However, further analysis is required to know the effectiveness of the proposed method under more realistic conditions, particularly in spontaneous speech recordings.

IV. CONCLUSIONS

We propose a noise robust speech rate estimation technique comprising a sub-band pruning strategy to compute a feature contour as well as a syllabic peak detection strategy in an SNR dependent manner. The pruning strategy selects a set of sub-bands from the sub-band energies used in the traditional TCSSBC contour. This is done to minimize noise and SNR dependent unwanted variations in TCSSBC. The selection strategy decides the use of VDs in peak detection to minimize the errors due to low accuracy in VDs estimation, particularly at low SNRs. Experiments with three corpora, namely, Switchboard, TIMIT, and CTIMIT reveal that speech rate estimation with the proposed strategies are more accurate than the best of the existing methods. Further investigations are required to study the use of the proposed method in current state-of-the-art ASR systems under different noise and SNR conditions. Future works also include the estimation of pruned sub-bands in unknown noise and SNR conditions considering a weighted combination of noise and SNR specific pruned sub-bands. Further, a direct estimation strategy could be developed for pruned sub-bands considering current state-of-the-art machine learning approaches, without any rule-based combination strategies on the noise and SNR specific pruned sub-bands.

[1]N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using online estimation of speaking rate," in *Proceedings of Eurospeech*, Rhodes, Greece (September 22–25, 1997), pp. 2079–2082.

[2]S. M. Chu and D. Povey, "Speaking rate adaptation using continuous frame rate normalization," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, Dallas, TX (March 14–19, 2010), pp. 4306–4309.

[3]S. Oviatt, "Predicting spoken disfluencies during human–computer interaction," Comput. Speech Lang. **9**(1), 19–35 (1995).

[4]C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," J. Acoust. Soc. Am. **107**(2), 989–999 (2000).

[5]M. Eskenazi, "An overview of spoken language technology for education," Speech Commun. **51**(10), 832–844 (2009).

[6]S. V. Pakhomov, L. S. Hemmy, and K. O. Lim, "Automated verbal fluency assessment," U.S. patent 9,576,593 (2017).

[7]M. Black, J. Tepperman, A. Kazemzadeh, S. Lee, and S. Narayanan, "Automatic pronunciation verification of English letter-names for early literacy assessment of preliterate children," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan (April 19–24, 2009), pp. 4861–4864.

[8]S. M. Ban and H. S. Kim, "Speaking rate dependent multiple acoustic models using continuous frame rate normalization," in *Proceedings of the Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Hollywood, CA (December 3–6, 2012), pp. 1–4.

[9]J. Epps and E. H. Choi, "An energy search approach to variable frame rate front-end processing for robust ASR," in *Proceedings of Interspeech*, Lisbon, Portugal (September 4–8, 2005), pp. 2613–2616.

[10]Z.-H. Tan and B. Lindberg, "A posteriori SNR weighted energy based variable frame rate analysis for speech recognition," in *Proceedings of Interspeech*, Brisbane, Australia (September 22–26, 2008), pp. 1024–1027.

[11]Z.-H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," IEEE J. Selected Topics Signal Process. **4**(5), 798–807 (2010).

[12]S. A. Borrie, M. Baese-Berk, K. Van Engen, and T. Bent, "A relationship between processing speech in noise and dysarthric speech," J. Acoust. Soc. Am. **141**(6), 4660–4667 (2017).

[13]Y. Jiao, V. Berisha, M. Tu, and J. Liss, "Convex weighting criteria for speaking rate estimation," IEEE/ACM Trans. Audio Speech Lang. Process. **23**(9), 1421–1430 (2015).

[14]H.-D. Huici, H. A. Kairuz, H. Martens, G. Van Nuffelen, and M. De Bodt, "Speech rate estimation in disordered speech based on spectral landmark detection," Biomed. Signal Process. Control **27**, 1–6 (2016).

[15]D. Wang and S. S. Narayanan, "Robust speech rate estimation for spontaneous speech," IEEE Trans. Audio Speech Lang. Process. **15**(8), 2190–2201 (2007).

[16]J. Yuan and M. Liberman, "Robust speaking rate estimation using broad phonetic class recognition," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, Dallas, TX (March 14–19, 2010), pp. 4222–4225.

[17]C. Heinrich and F. Schiel, "Estimating speaking rate by means of rhythmicity parameters," in *Proceedings of Interspeech*, Florence, Italy (August 27–31, 2011), pp. 1873–1876.

[18]T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA (May 12–15, 1998), pp. 945–948.

[19]T. Dekens, H. Martens, G. Van Nuffelen, M. De Bodt, and W. Verhelst, "Speech rate determination by vowel detection on the modulated energy envelope," in Proceedings of the 22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal (September 1–5, 2014), pp. 1252–1256.

[20]Y. Zhang and J. R. Glass, "Speech rhythm guided syllable nuclei detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan (April 19–24, 2009), pp. 3797–3800.

[21]N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," Behav. Res. Methods **41**(2), 385–390 (2009).

[22]D. Wang and S. Narayanan, "Speech rate estimation via temporal correlation and selected sub-band correlation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, PA (March 18–23, 2005), pp. 413–416.

[23]T. Dekens, M. Demol, W. Verhelst, and P. Verhoeve, "A comparative study of speech rate estimation techniques," in *Proceedings of Interspeech*, Antwerp, Belgium (August 27–31, 2007), pp. 510–513.

[24]C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection," Speech Commun. **78**, 62–71 (2016).

[25]S. Nagesh, C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "A robust speech rate estimation based on the activation profile from the selected acoustic unit dictionary," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Shanghai, China (March 20–25, 2016), pp. 5400–5404.
[26]A. A. Reddy, N. Chennupati, and B. Yegnanarayana, "Syllable nuclei detection using perceptually significant features," in *Proceedings of Interspeech*, Lyon, France (August 25–29, 2013), pp. 963–967.
[27]C. Landsiedel, J. Edlund, F. Eyben, D. Neiberg, and B. Schuller, "Syllabification of conversational speech using bidirectional long-short-term memory neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic (May 22–27, 2011), pp. 5256–5259.
[28]K. L. Brown and E. B. George, "CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Detroit, MI (May 9–12, 1995), pp. 105–108.
[29]V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," Speech Commun. **9**(4), 351–356 (1990).
[30]M. Huckvale, "Speech filing system: Tools for speech research," http://www.phon.ucl.ac.uk/resource/sfs (Last viewed 2/9/2019).
[31]J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, CA (March 23–26, 1992), pp. 517–520.
[32]K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 2158–2168 (2014).
[33]S. Gonzalez and M. Brookes, "PEFAC—A pitch estimation algorithm robust to high levels of noise," IEEE/ACM Trans. Audio Speech Lang. Process. **22**(2), 518–530 (2014).
[34]Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 7–19 (2014).
[35]P. Karjol, M. A. Kumar, and P. K. Ghosh, "Speech enhancement using multiple deep neural networks," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, Canada (April 15–20, 2018), pp. 5049–5052.
[36]A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun. **12**(3), 247–251 (1993).
[37]H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop*, Paris, France (September 18–20, 2000), pp. 181–188.
[38]B. Fisher, "tsylb2-1.1: syllabification software," https://www.nist.gov/itl/iad/mig/tools (Last viewed May 30, 2017).
[39]J. Holmes, "The JSRU channel vocoder," IEE Proc. F **127**(1), 53–60 (1980).
[40]T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," IEEE Trans. Audio Speech Lang. Process. **20**(4), 1383–1393 (2012).
[41]T. Gerkmann and R. C. Hendriks, "Noise power estimation based on the probability of speech presence," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY (October 16–19, 2011), pp. 145–148.
[42]J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise Reduction in Speech Processing* (Springer, New York, 2009), pp. 1–4.
[43]N. Krishnamurthy and J. H. Hansen, "Babble noise: Modeling, analysis, and applications," IEEE Trans. Audio Speech Lang. Process. **17**(7), 1394–1407 (2009).
[44]P. Howell and K. Kadi-Hanifi, "Comparison of prosodic properties between read and spontaneous speech material," Speech Commun. **10**(2), 163–169 (1991).
[45]C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech," J. Acoust. Soc. Am. **111**(6), 2862–2873 (2002).
[46]P. Lieberman and S. E. Blumstein, *Speech Physiology, Speech Perception, and Acoustic Phonetics* (Cambridge University Press, Cambridge, UK, 1988).
[47]H. J. Steeneken and F. W. Geurtsen, "Description of the RSG-10 noise database," Report IZF No. 3, TNO Institute for Perception, 1988.