Circuits, Systems, and Signal Processing (2021) 40:6098–6120
https://doi.org/10.1007/s00034-021-01754-1

A Robust Speaking Rate Estimator Using a CNN-BLSTM Network

Aparna Srinivasan[1] · Diviya Singh[2] · Chiranjeevi Yarra[3] · Aravind Illa[4] · Prasanta Kumar Ghosh[4]

Received: 8 March 2020 / Revised: 13 May 2021 / Accepted: 18 May 2021 / Published online: 11 June 2021

Abstract

Direct acoustic feature-based speaking rate estimation is useful in applications including pronunciation assessment, dysarthria detection and automatic speech recognition. Most of the existing works on speaking rate estimation have steps which are heuristically designed. In contrast to the existing works, in this work a data-driven approach with convolutional neural network-bidirectional long short-term memory (CNN-BLSTM) network is proposed to jointly optimize all steps in speaking rate estimation through a single framework. Also, unlike existing deep learning-based methods for speaking rate estimation, the proposed approach estimates the speaking rate for an entire speech utterance in one go instead of considering segments of a fixed duration. We consider the traditional 19 sub-band energy (SBE) contours as the low-level features as the input of the proposed CNN-BLSTM network. The state-of-the-art direct acoustic feature-based speaking rate estimation techniques are developed based on

✉ Chiranjeevi Yarra
chiranjeevi.yarra@iiit.ac.in

Aparna Srinivasan
a2sriniv@ucsd.edu

Diviya Singh
diviya@ee.iitr.ac.in

Aravind Illa
aravindi@iisc.ac.in

Prasanta Kumar Ghosh
prasantg@iisc.ac.in

[1] Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093, USA

[2] Department of Electrical Engineering, Indian Institute of Technology Roorkee (IITR), Roorkee 247667, India

[3] Language Technologies Research Center, International Institute of Information Technology (IIIT), Hyderabad 500032, India

[4] Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India

19 SBEs as well. Experiments are performed separately using three native English speech corpora (Switchboard, TIMIT and CTIMIT) and a non-native English speech corpus (ISLE). Among these, TIMIT and Switchboard are used for training the network. However, testing is carried out on all the four corpora as well as TIMIT and Switchboard with additive noise, namely white, car, high-frequency-channel, cockpit, and babble at 20, 10 and 0 dB signal-to-noise ratios. The proposed CNN-BLSTM approach outperforms the best of the existing techniques in clean as well as noisy conditions for all four corpora.

Keywords Speaking rate estimation · Acoustic feature-based estimation · Convolutional neural network · Bidirectional long short-term memory networks · Robust estimation under noise

1 Introduction

In general, speaking rate is measured as the number of speech units per second and these speech units could be either phones or syllables. In most of the existing works, syllables have been considered as the speech unit. It has been shown that the number of syllables per second correlates well with oral fluency in the application of language assessment [14,16]. Speaking rate could be useful for several applications including automatic speech recognition (ASR) [49], detection of dysarthria [46], computer-assisted language learning (CALL) [13] and emotion recognition [62].

Speaking rate is typically estimated in two ways—(1) ASR-based approach and (2) direct acoustic feature-based approach. In the ASR-based approach, first, the syllable segments are predicted using phonetic forced-alignment [63] followed by syllabification processes [5]. Then, the predicted segments are used to estimate the syllable rate. However, this approach suffers from several limitations—(1) the speaking rate can be estimated only when speech with reference transcription is available [60], (2) in case speaking rate needs to be estimated from speech under different conditions (e.g., shouted, whispered, dysarthric speech), the ASR system, typically trained on normal speech, fails to provide correct syllable boundaries [23] and (3) ASR-based approach is computationally expensive [60]. On the other hand, in the direct acoustic feature-based approach, the speaking rate is estimated using the features derived based on the acoustic properties of the vowels, which, in general, correspond to the syllable nuclei [54,64]. This approach is computationally less expensive as compared to the ASR-based approach [50]. In this work, we consider the problem of acoustic feature-based automatic speaking rate estimation.

1.1 Significance of Speaking Rate

Speaking rate has been shown to be useful in several applications including pronunciation assessment [13], ASR [12,49] and dysarthria [10]. Apart from these, it has also played a role in problems including perception studies, age estimation etc.

In the assessment of pronunciation, for Dutch fluency, Cucchiarini et al. had observed that the speaking rate correlates well with the expert's rating [13]. Further, they concluded that the speaking rate was a good predictor of oral fluency. Similarly, Black et al. found that human evaluators considered speaking rate to be a critical cue in determining a child's overall reading ability [8]. In another work, they used speaking rate for automatic evaluation of non-native English pronunciation [7]. They observed that the evaluated pronunciation obtained using speaking rate was more accurate than those using pauses or goodness of pronunciation. Further, it was hypothesized that direct acoustic feature-based speaking rate estimation could be useful for the application of pronunciation assessment [60,61].

Speaking rate is also a critical element in ASR systems [12,49]. Richardson et al. had found that the sensitivity of ASR to speaking rate limited its potential to be effectively used for real-time applications [55]. They observed that the recognition accuracy of an ASR system is affected by the mismatch between the speaking rate in the training and testing conditions. In order to overcome this drawback and improve ASR accuracy, they proposed a normalization technique on the phoneme durations in an utterance for improving fast speech recognition [55]. Similarly, Bartels and Bilmes incorporated speaking rate in the recognition models [4]. They considered the speaking rate estimation proposed by Wang et al. [59,60] where direct acoustic feature-based approach was used. Zheng et al. used a rate specific acoustic model [67] to make their ASR systems robust to fast speech.

It has been observed that speaking rate is low for people with neurological disorders that affect speech motor control like Parkinson's induced hypokinetic dysarthria [10] or apraxia of speech like stuttering [36]. Jiao et al. proposed a direct acoustic feature-based speaking rate estimation technique for monitoring the progress of dysarthria [33]. They observed that the estimated speaking rate followed the deterioration in speaking ability caused by the disease. Further, it has been shown that speaking rate depends on various factors including individual character [3,48], age [2], demographic, cultural, psychological and physiological factors [2]. Thus, direct acoustic feature-based speaking rate estimation could be useful for automatic prediction of those factors.

1.2 Review of Existing Works

In general, acoustic feature-based speaking rate estimation has been addressed in two ways—(1) knowledge-driven methods and (2) data-driven methods.

In the knowledge-driven methods, first, a one-dimensional feature contour is derived from speech signal such that the peaks in the contour correspond to the syllable nuclei locations. Then, a peak detection algorithm is used to predict the number of syllables using which the speaking rate is estimated [61]. In these methods, both acoustic feature computation and peak detection are carried out in a heuristic manner. Pfau and Ruske used a smoothed modified loudness contour as an acoustic feature and estimated the peaks using a threshold and frame range-based peak detection strategy [54]. Zhang and Glass proposed a contour based on Hilbert envelope and used a rhythm guided peak counting algorithm to estimate the syllable nuclei locations [65]. They improved the peak counting by removing the peaks falling in the unvoiced regions using voiced and unvoiced (VuV) decisions from the estimated pitch values. Jong et al. estimated the speaking rate using intensity-based envelope with peak counting along with VuV decisions [35]. Similarly, Heinrich and Schiel used the short-time sound pressure energy as the acoustic feature and obtained the peaks by a simple thresholding mechanism [25].

In addition to these works, Kitazawa et al. proposed a wide-band spectrum of a speech signal as an acoustic feature. The speaking rate was estimated by measuring the dominant peaks in the spectrum [39]. However, Morgan and Fosler-Lussier had found that the wide-band spectrum contained a lot of noise and, therefore, it became difficult to estimate the speaking rate robustly [50]. Instead, they developed an approach based on the sub-band energies (SBEs) and performed peak counting on the normalized product of cross correlation between all pairs of the energies [50]. Further, Wang and Narayanan improved this work by proposing a feature contour called “Temporal Correlation Selected Sub-Band Correlation” (TCSSBC) [59,60]. Furthermore, they used the estimated pitch along with the TCSSBC contour and a peak detection strategy which involved a threshold for the distance between adjacent peaks.

On the other hand, only a few works exist that consider data-driven approaches, where statistical learning methods are used either for directly estimating the speaking rate from the acoustic feature, or to replace heuristic peak detection algorithms with data-driven approaches. For example, Yarra et al. proposed a mode-shape classification technique, wherein the smoothed TCSSBC contour was divided into single peak segments (modes) and a D-dimensional mode shape vector (MSV) was generated [61]. The MSVs were then classified as syllabic or non-syllabic using a support vector machine. Jiao et al. proposed a recurrent neural network (RNN) model for the estimation of speaking rate over a window of fixed duration of 1s. For this, they computed the features by using heuristic functions on the envelope modulation spectrum (EMS) and MFCC of the speech signal [33]. However, this approach has the limitation of using a fixed window duration of 1s. But, in general, the speaking rate often needs to be estimated for a given speech utterance of any length.

1.3 Motivation for the Proposed Approach

Generally, in comparison with the knowledge-driven methods, speaking rate estimation accuracy improves by using the data-driven methods [61]. However, existing works on data-driven approaches for speaking rate estimation are limited to using peak detection on features obtained heuristically from low-level representations. For example, the feature contour (TCSSBC) used in the data-driven peak detection proposed by Yarra et al. [61] was derived from 19 short-time SBE contours [59,60]. Similarly, in another work on data-driven peak detection proposed by Jiao et al. [33], the features were computed from EMS and MFCCs. We hypothesize that such heuristically derived features could degrade the performance of speaking rate estimation techniques. In order to illustrate this hypothesis, we consider an exemplary utterance “Don’t ask me to carry an oily rag like that.", taken from the TIMIT corpus in Fig. 1. For the illustrative example, we consider TCSSBC and the corresponding 19 short-time SBE contours. TCSSBC is considered to be one of the best features for speaking

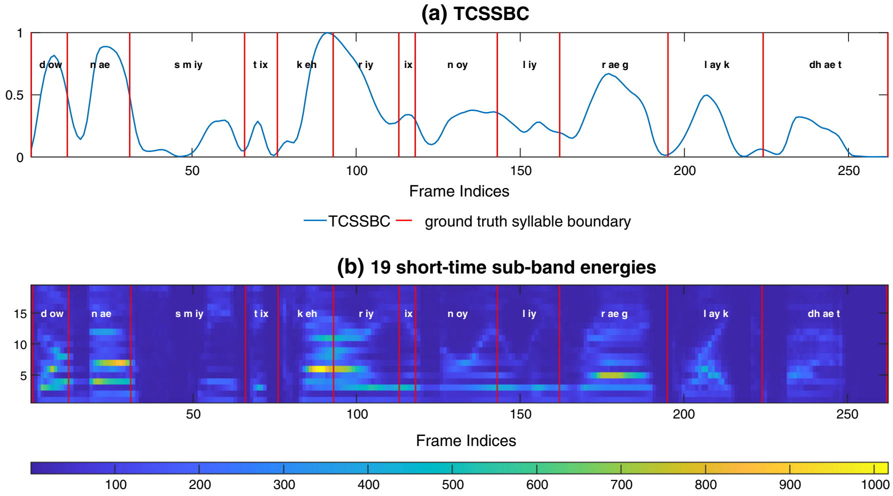

Fig. 1 TCSSBC and the 19 short-time SBE contours of the utterance “Don’t ask me to carry an oily rag like that." from the TIMIT corpus

rate estimation [15] and it has been widely used in many works [4,61]. TCSSBC uses a method called spectro-temporal correlation (STC) [51] to obtain a one-dimensional feature contour, in which correlations are first performed along the temporal axis followed by a correlation along the spectral axis. The STC computes both the spectral and temporal correlations independently.

In Fig. 1, the red vertical lines indicate the ground truth syllable boundaries. The utterance considered in Fig. 1 has 12 syllables within a duration of 2.62s. Thus, the ground truth speaking rate is 4.58 syllables/s. However, the estimated speaking rate is 4.20 syllables/s using both the heuristic approach proposed by Wang et al. [60] and the data-driven approach proposed by Yarra et al. [61]. This could be because of the absence of TCSSBC peak in the syllable segment “r iy". However, from Fig. 1b, it is observed that within the “r iy" syllable segment, there are sub-bands with high energies. Wang et al. had observed that consistent higher SBEs in a segment were indicative of a syllable nucleus [59], and in the computation of TCSSBC, this property was considered to obtain local peaks around the syllable nuclei [51]. However, the high energy bands in the syllable segment “r iy" are not successfully transformed into a strong peak in the TCSSBC contour. This could be because of the heuristic way of computing the TCSSBC with independent spectral and temporal correlations. Further, TCSSBC is computed in a manner agnostic to the peak detection strategy. Considering both of these, we believe that the performance of the speaking rate estimation could be improved by an approach that considers feature computation and peak detection jointly in a single computational framework. In order to achieve this, we propose an approach by exploring convolutional neural network-long short-term memory (CNN-LSTM) network-based models.

CNNs consist of a set of filters that are applied on low level features in order to learn the feature representations. They have been mostly used in the applications of image processing [42] and speech processing including speech recognition [11,53], and emotion recognition[29,45]. In speech recognition, CNNs are used for learning the feature representations from either the raw speech waveform or low level features such as MFCC, power spectrum from which the speech is decoded [11,53]. Similarly in speech emotion recognition, CNNs are used to learn the feature representations from the speech signal or low level features, which are then classified into different emotions [29,45]. However, training a model containing only CNNs requires fixed length sequences in the data set. But, in speaking rate estimation the length of the utterances is not uniform across the training data.

In order to handle variable length sequences, we propose to append LSTMs to CNNs. LSTMs are recurrent neural networks which have been used to learn both the long-term and short-term dependencies in the input sequences [26]. In this case, we propose to consider the input sequences as the local representations learnt by CNNs. There exist a few works which show the benefit of CNN-LSTM models for variable length feature sequences [52,68]. Further, bidirectional LSTMs (BLSTMs), which are a variant of LSTMs, have been used for learning the temporal dependencies in variable length sequences by exploring the similarities in forward and backward directions [22]. Since the task of speaking rate estimation could be similar in both directions, we propose to consider CNN-BLSTM models, for which there is no existing work to the best of our knowledge.

In this work, for speaking rate estimation, we build a CNN-BLSTM model (code is available online[1]) using 19 SBE contours. In addition to the 19 SBE contours, we use pitch values to suppress unwanted variations of SBEs in the unvoiced segments which might affect speaking rate estimation. Experiments are performed on four corpora namely, TIMIT [69], Switchboard [19], CTIMIT [9] and ISLE [47] considering the Pearson correlation coefficient between the estimated and the ground truth speaking rate as the performance measure. The performance of the proposed method is found to be better than the best of the two baseline schemes that do not consider representation learning, under all four corpora, when the models are trained with both TIMIT and Switchboard corpora. Also, the proposed method is found to be better than the baseline in noisy conditions under five additive noises at three SNRs of 20 dB, 10 dB and 0 dB, which shows the robustness of the proposed representations learning for the speaking rate estimation task.

The rest of the paper is organized as follows: the speech corpora details are described in Sect. 2, and the proposed approach is discussed in Sect. 3. The experimental results are analyzed and elaborated in Sect. 4. Finally, conclusions are discussed in Sect. 5.

2 Database

We use ICSI Switchboard [19], TIMIT [69], CTIMIT [9] and ISLE [47] 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 speaking rate, ranging from 1.26 to 9.2 syllables per second. The audio in Switchboard corpus was col-

[1] https://github.com/diviya97/CNN-BLSTM-Speaking-Rate-Estimator.

lected through the telephone channel. A subset of 7300 speech segments (obtained in a manner similar to that of “spurt" as described in [50]), each of duration greater than 200ms, is used for our experiments. In Switchboard, syllable transcriptions as well as their time aligned boundaries are available. However, no phonetic transcription is available. TIMIT is a read speech database, which has phonetically balanced 6300 sentences spoken by 630 speakers with a speaking rate ranging from 1.44 to 8 syllables per second. All speech utterances of all sentences from TIMIT are used for our experiments. CTIMIT corpus is similar to TIMIT except that the audio was collected through the cell phone channel under various noisy conditions. All speech utterances of all 3370 sentences from CTIMIT corpus, spoken by 630 speakers, are used for our experiments. The speaking rate in CTIMIT sentences ranges from 1.87 to 8 syllables per second. ISLE corpus contains utterances from 46 non-native speakers (23 German (GER) and 23 Italian (ITA)) learning English. Each speaker uttered approximately 160 sentences. All speech utterances of all sentences from ISLE corpus are used for our experiments. The speaking rate in the ISLE sentences ranges from 0.64 to 7.81 syllables per second. In TIMIT, CTIMIT and ISLE, 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 [17]. Following the work by Wang et al., for the experimentation, silent segments in the initial and final parts of each sentence of all corpora are removed [60]. We use five noises, namely white, volvo, hfc, f16 and babble from NOISEX-92 database [58] in the experiments. Babble noise has the most non-stationary characteristics among all five noises considered in this work.

3 Proposed Approach

The proposed approach for speaking rate estimation consists of 2 components—computation of 19 SBE contours and the CNN-BLSTM model. Each of these components are explained in detail in the following subsections.

3.1 19 SBE Contours

3.1.1 Computation of SBE Contours

Figure 2 shows the block diagram that outlines the steps involved in the computation of 19 SBE contours following the work by Holmes et al. [27]. In the first step, the input signal is filtered by 19 second-order Butterworth bandpass filters separately. The center frequencies and the bandwidths of the bandpass filters are taken from the work by Holmes et al. [27] and are reported in Table 1. In the second step, the modulus function is applied on the filtered signals to ensure all the values in the filtered signals are positive. In the third step, the resultant signals are smoothened using a first-order Butterworth low-pass filter with cutoff frequency of 50 Hz. Finally, in the last step, energy values are computed in each window having duration of 20 ms with a window shift of 10 ms for all 19 smoothened signals separately.

Table 1 Center frequencies (f_c) and bandwidths (BWs) of the 19 sub-bands

Sub-band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
f_c (Hz)	240	360	480	600	720	840	1000	1150	1300	1450	1600	1800	2000	2200	2400	2700	3000	3300	3750
BW (Hz)	120	120	120	120	120	120	150	150	150	150	150	200	200	200	200	300	300	300	500

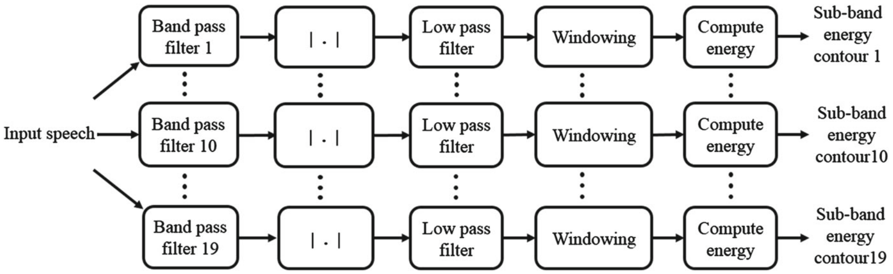

Fig. 2 Steps involved in the generation of 19 SBE contours

3.1.2 Benefit of SBE Contours

In the computation of SBE contours, Holmes et al. chose the bandpass filters such that the filter bands closely match the critical bands of human auditory system [27]. They also emphasized that the SBE contours could capture formant like structures around the syllable nuclei. Further, they found that the structures corresponding to fricatives and stops were different from those corresponding to the vowels. They also observed that when the speech was synthesized from the SBE contours, there was no loss of intelligibility in the synthesized speech. Considering these, we hypothesize that the spectral information in the 19 SBE contours could be useful to explore syllable nuclei locations for speaking rate estimation. We further analyze the variations in the spectral information using TIMIT corpus with the help of Fig. 3.

Figure 3 shows the average 19 SBE contours for all phonemes (excluding closures of six stop consonants “/b, d, g, p, t, k/") present in the TIMIT corpus. These are computed following the steps below:

1. Obtain the SBE contours specific to each phoneme segment using their respective time aligned boundaries available in the corpus.
2. Resample the SBE contours within every phoneme segment to 10 frames to obtain uniform length.
3. Average the energy values at every sub-band and frame locations, across all segments for each phoneme.

From the figure, it is observed that the sub-bands corresponding to higher energies vary across all phonemes. These sub-bands differ between vowels (which are typically considered as syllable nuclei [54]) and consonants. For example, in vowel “ae", the higher energies are observed around the 5th sub-band, whereas, in consonant “ch" they are around the 19th sub-band. These high energy sub-bands are also found to be different across different vowels. For example, considering vowels “ae" and “em", it is observed that, in “em", the higher energies are around the 1st sub-band in contrast to the higher sub-band energies in “ae". Further, it is interesting to observe that the temporal variations of the higher energies are not uniform across vowels. For example, in vowel “ah", higher energies can be seen around 5th sub-band at the frames in the beginning of the segment. However, at the end those energies are also found in 10th sub-bands. In contrast, in vowel “aw", at the end of segment high energies can be seen only around 5th sub-band, but, in the beginning, high energies are observed around 5th to 10th

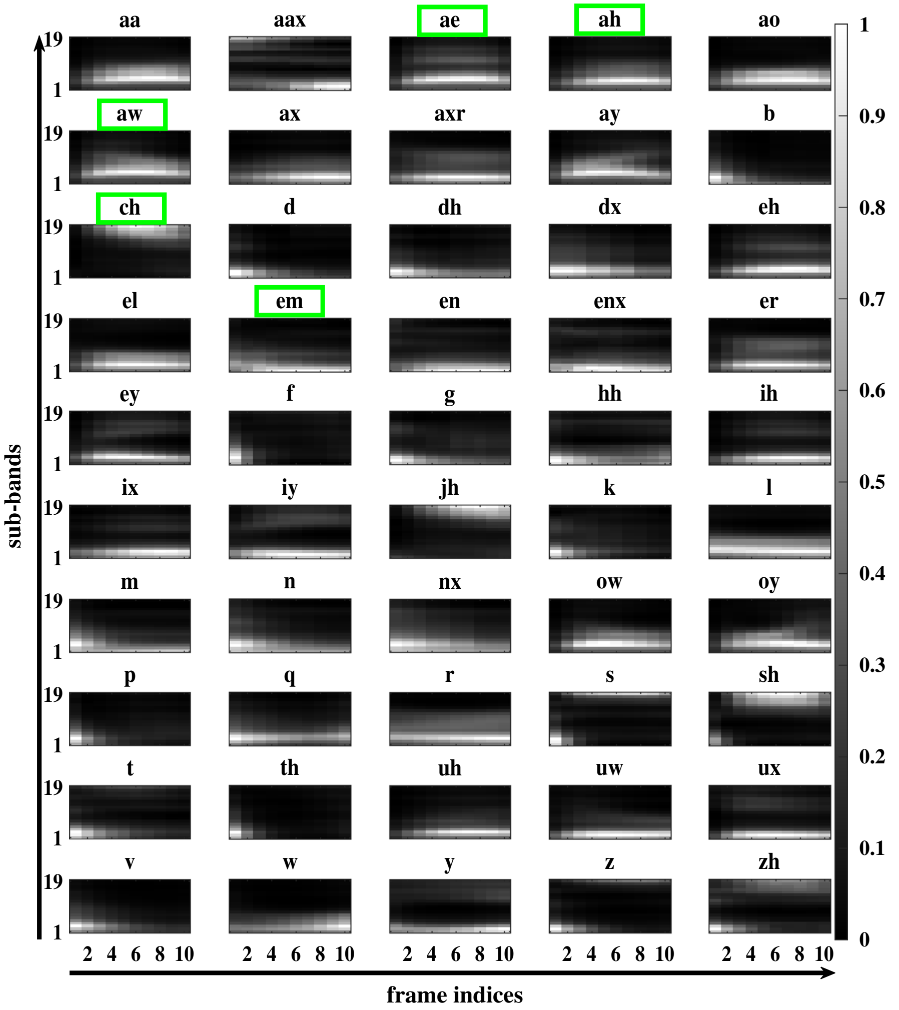

Fig. 3 Average 19 SBE contours for all phonemes present in TIMIT corpus resampled to 10 frames

sub-bands. These observations suggest that sub-band energies within each phoneme have spectro-temporal variations. Further, these variations are specific to each vowel (syllable nuclei) as well as non-vowels (non-syllable nuclei). These variations, in general, significantly differ between syllable nuclei and non-syllable nuclei. Thus, capturing the spectro-temporal pattern specific to syllable nuclei could benefit the speaking rate estimation task. In the literature, the benefits of these spectro-temporal energy patterns have been explored for speaking rate estimation, for example, using TCSSBC [59–61].

As explained using Fig. 1, the independent nature of spectro-temporal correlation used in TCSSBC contour and the independent nature of contour formation and peak picking are potentially the causes for incorrect estimation of speaking rate values. In

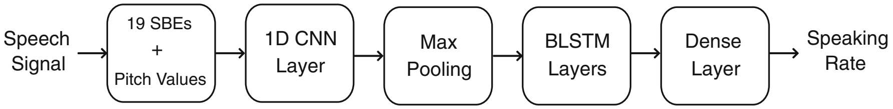

Fig. 4 Block diagram summarizing the steps of the proposed CNN-BLSTM network

order to jointly learn the peak picking and feature computation strategies, we propose to estimate the speaking rate directly using the CNN-BLSTM architecture considering a 20-dimensional feature vector obtained by concatenating the 19 SBEs and the pitch values. The inclusion of pitch is considered following most of the works on speech rate estimation [59,60,65], where VuV decisions estimated from pitch values have been used to suppress the unwanted variations of 19 SBEs in the unvoiced segments for better speech rate estimation. Considering this architecture, the speaking rate is estimated using a single network, which, we believe, learns both the feature contour and peak detection stages in a data-driven manner. Further, we also hypothesize that a single network learns representations constituting both spectral and temporal variations specific to syllable nuclei jointly in a data-driven manner.

3.2 CNN-BLSTM

This subsection presents a description of the proposed CNN-BLSTM model. In the proposed model, we use a combination of CNN and BLSTM as shown in Fig. 4. CNNs [1] are known for capturing local structures in a sequence while BLSTMs [44] deal with temporal dynamics by storing information from previous time steps in their internal state. We briefly present a review of CNN and BLSTM from literature, followed by the proposed approach for speaking rate estimation.

3.2.1 Convolutional Neural Networks (CNNs)

CNNs [1] belong to a class of deep neural networks that have been proven to be very effective in the field of image recognition and classification [34]. 1D CNNs have been applied to classify ECG signals [43] and recognize respiration patterns [37] and speech emotion [66].

1D CNNs [38] are capable of learning relevant features for the task in a data-driven manner from sequences of one-dimensional data. In this work, 1-d convolution filters are deployed to extract local temporal patterns similar to 2-d convolution filters in image processing tasks where filtering is performed in spatial domain. Unlike fully connected layers, CNNs perform operations by utilizing the temporal structures in the data to extract local features. This reduces the number of parameters that are required to be learned, thereby improving the efficiency of feature extraction. Consequently, it also turns out to be computationally efficient [18]. The feature maps x_j^l obtained through convolution operation by the l-th convolutional layer are elaborated below:

$$x_j^l = \sigma \left(\sum_{i \epsilon N_j} x_i^{l-1} \otimes w_{ij}^l + b_j^l \right) \tag{1}$$

where $\otimes$ denotes convolution operation and w_{ij}^l and b_j^l represent the weight and bias of the jth convolutional filter, respectively. N_j is the number of input feature maps and f is the activation function. The features extracted from CNN are further processed using max-pooling and batch normalization layers [41]. The feature maps from convolution layer are down-sampled using max-pooling. The features extracted by 1D CNN layer are normalized using a batch normalization layer before feeding it to ReLU [28] nonlinearity. Batch normalization makes the optimization landscape significantly smoother, thus helping in more stable gradient propagation, faster training [56], and also acts as a regularizer [31] to reduce over-fitting.

3.2.2 Bidirectional Long Short-Term Memory (BLSTM)

Recurrent neural networks (RNNs) [21] are a class of neural networks designed to recognize patterns in sequential data. It is a simple feed-forward neural network with a feedback. They use their internal states (memory) to capture information from previous inputs and use this information to predict the current output. RNNs have been known for better modeling the temporal structure of sequential data but are not effective when sequences are very long because the back-propagated error can decay or boost exponentially with increasing number of time steps resulting in vanishing or exploding gradient problem [6]. To deal with the vanishing gradient problem of RNNs, LSTM was proposed in [26].

Compared to the standard RNN, LSTM architecture has an additional state referred to as cell state (c_t), which is used to preserve long-term information. LSTM also has three multiplicative gates namely an input gate (i_t), an output gate (o_t) and a forget gate (f_t) to regulate the flow of information inside the LSTM unit. At time t, let x_t be an M-dimensional input and N be the number of memory cells in an LSTM layer which outputs hidden state $h_t \in \mathbb{R}^N$. Then, for each LSTM layer there are different types of weights, namely input weights $V_* \in \mathbb{R}^{NXM}$, recurrent weights $U_* \in \mathbb{R}^{NXN}$ and bias weights $b_* \in \mathbb{R}^N$, where '$_*$' corresponds to either cell state (c) or one of the multiplicative gates, i, o, f. The forward operations for an LSTM unit are governed by the following equations [24]:

$$\begin{aligned}
f_t &= \sigma(V_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma(V_i x_t + U_i h_{t-1} + b_i) \\
c_t &= c_{t-1} \odot f_t + i_t \odot \tanh(V_c x_t + U_c h_{t-1} + b_c) \\
o_t &= \sigma(V_o x_t + U_o h_{t-1} + b_o) \\
h_t &= \tanh(c_t) \odot o_t
\end{aligned} \tag{2}$$

where σ is a point-wise nonlinear activation function and $\odot$ denotes the element-wise multiplication of two vectors.

LSTMs process the input sequence in the forward direction and, thus, only make use of the past context to predict the current output. Bidirectional LSTMs (BLSTMs) [32] can access long-range context in both forward and reverse directions. It has been shown to outperform both unidirectional LSTMs and standard recurrent neural networks (RNNs) in phoneme classification task [20]. BLSTM uses two different LSTM layers, namely forward and backward. These two layers connect to the same output layer to generate the output information. Thus, this network takes into account both the past and future information of the sequential data to estimate the output variables.

We approach speaking rate estimation technique as a many-to-one regression problem, where the target output is a real number for the corresponding sequence of the frame level input features from a given speech utterance. The output from the BLSTM layer is fed to a dense layer to obtain the speaking rate as shown in Fig. 4. The target speaking rate for a speech utterance is computed by dividing the total number of syllable nuclei with the utterance duration. For a given speech utterance, we compute the frame level input features (19 SBEs plus pitch values) which are fed as inputs to the 1D CNN, in order to derive higher level representations. The output of the CNN layer is then down-sampled using max-pooling layer and fed to the BLSTM layers to model the temporal structure of the input sequences. We also use Dropout [40] between these layers to avoid over-fitting. We pass the output of the BLSTM layer through a dense layer having a single output neuron with linear activation. The output of the dense layer is then utilized to estimate the speaking rate for the given speech utterance.

4 Experiments and Results

4.1 Experimental Setup

All experiments are carried out in a fivefold cross-validation setup, where 3 folds are used for training, 1 fold for validation, and remaining 1 fold for testing. Consequently, the CNN-BLSTM network is trained, validated and tested with 60%, 20% and 20% of the data, respectively, in a round-robin fashion. The results are reported in terms of average (standard deviation) Pearson correlation coefficient between the estimated and the ground truth speaking rate computed across all 5 testing folds from the test set considered.

4.1.1 Setup for the Proposed Approach

The 19 SBEs are computed as described in Section 3.1.1 using the speech filing system tool [30]. The pitch values are estimated using an algorithm based on normalized cross correlation function and dynamic programming [57] in a sub-routine provided by the speech filing system tool [30]. The obtained pitch values are zero in the regions which are estimated as unvoiced by the algorithm. In order to avoid discontinuities in the pitch trajectory between voiced and unvoiced regions, which might not be modeled well by the BLSTM, the pitch values in the unvoiced regions are obtained by linearly interpolating the pitch values in the voiced regions followed by applying a 5-point moving average filter. Each of the resultant 20 features (from 19 SBEs and pitch) corresponding to each speech utterance is normalized to have zero mean and unit standard deviation.

The proposed CNN-BLSTM model consists of a one-dimensional convolutional layer having 64 filters of length 5 with stride of 1. The representation that is learned by the convolutional layer is batch normalized and downsampled using a max-pooling layer with pooling size of 2. This is followed by 2 BLSTM layers, each consisting of 64 hidden units. A dropout of 0.2 is applied after each layer. Rectified linear units (ReLU) and hyperbolic tangent (tanh) are used as activation functions for the convolutional and BLSTM layers, respectively. The final layer consists of a single neuron with linear activation to predict the speaking rate. This model architecture is chosen from several combinations of number of layers, hidden units and filter lengths that resulted in the highest average correlation coefficient on the validation data. The model is trained using the Adam optimizer, and the mean squared error is chosen as the loss function. Further, the model is trained with early stopping criterion, wherein the training is stopped if the mean squared error on the validation data does not decrease for 3 consecutive epochs.

4.1.2 Setup for Baseline Approaches

The efficacy of the proposed approach is determined by comparing its performance with that of two baseline approaches. The baseline approaches considered are—(1) robust speaking rate estimation (RSRE) [60] and (2) online speaking rate estimation (OSRE) [33]. For the RSRE technique, following the work by Wang et. al. [60], the TCSSBC is computed using spectro-temporal correlation from 19 SBEs, from which the speaking rate is estimated. As outlined in the work by Jiao et. al. [33], the OSRE technique estimates the speaking rate for every 1s window of speech with 0.1s shift. Each window of speech is represented using a feature that comprises EMS, 13th order MFCC and their delta and delta-delta derivatives, and six statistical functions performed on each row of MFCC. The dimension of this feature was then reduced to 200 by principal component analysis (PCA), and we refer to the resultant feature as OSRE-ftr. A BLSTM model (referred to as OSRE-model) uses OSRE-ftr to predict the speaking rate. The BLSTM model is similar to the one used in the work by Jiao et. al. [33]. In general, to achieve a reliable speaking rate measure [50], the speaking rate is required to be computed for an entire speech utterance. However, the OSRE technique estimates the speaking rate for a fixed duration of 1s. Thus, in order to compute speaking rate for the entire utterance, we modify the OSRE technique as follows. The OSRE-ftr is computed following the work by Jiao et. al. [33] for an entire speech utterance. The OSRE-model is trained to predict the speaking rate for each speech utterance considering OSRE-ftr as the input.

4.1.3 Software Setup

For the experimentation, CNN-BLSTM models are trained and tested in Python using Keras 2.3.1 with Tensorflow 1.13.1 backend. For the OSRE technique, the Librosa

Table 2 Comparison of the proposed approach with RSRE and OSRE techniques

	TIMIT	Switchboard
Proposed	0.8330 (0.0089)	0.7635 (0.0123)
RSRE	0.6525 (0.0175)	0.6863 (0.0244)
OSRE	0.6677 (0.0217)	0.5094 (0.0126)

Table 3 Comparison of baseline OSRE-ftr and 19 SBEs considering the baseline OSRE-model on TIMIT and Switchboard corpus

	TIMIT		Switchboard	
	OSRE-ftr	19 SBEs	OSRE-ftr	19 SBEs
OSRE-model	0.6677 (0.0217)	0.7579 (0.0728)	0.5094 (0.0126)	0.7182 (0.0161)

library is used for computing the MFCC features and the Scikit-learn library is used for implementing PCA.

4.2 Results and Discussion

4.2.1 Comparison of Different Techniques for Speaking Rate Estimation

Table 2 shows the average (standard deviation) correlation coefficients for TIMIT and Switchboard corpora obtained using the proposed approach, RSRE and OSRE techniques. From the table, it is observed that, irrespective of the corpus, the proposed approach provides a significantly higher ($p < 0.01$, t-test) correlation coefficient than both the baseline techniques. A lower correlation coefficient in the case of RSRE could be because RSRE technique uses TCSSBC and a peak detection algorithm, both of which involve heuristics. Higher correlation coefficient with the proposed approach over OSRE indicates that the CNN-BLSTM in the proposed approach captures better spectro-temporal information from 19 SBEs for speaking rate estimation task than the spectro-temporal information learnt in OSRE using only BLSTM and heuristic OSRE-ftr. We further investigate the benefit of the CNN-BLSTM and 19 SBEs in the proposed approach, respectively, over only BLSTM (OSRE-model) and OSRE-ftr with the help of Table 3 along with results in Table 2.

Table 3 shows the average (standard deviation) correlation coefficient obtained on TIMIT and Switchboard corpus using 19 SBEs and OSRE-ftr considering OSRE-model. It is to be noted that the results using OSRE-model (with OSRE-ftr) in the table are identical to those in Table 2. From the table, it is observed that the correlation coefficients obtained with 19 SBEs are significantly higher ($p < 0.01$, t-test) than those with OSRE-ftr in both the corpora. This indicates that the 19 SBEs is better suited than OSRE-ftr for the task of speaking rate estimation. Further, higher correlation coefficients in Table 2 with the proposed approach than those in Table 3 under 19 SBEs indicate the benefit of the CNN layer. This suggests that the CNN considered in the proposed approach learns better spectro-temporal cues than just the BLSTM for the task of speaking rate estimation. Furthermore, from Tables 2 and 3, it is inter-

Table 4 Evaluation of the proposed approach trained on utterances from one corpus and tested with those from another

	Test corpus Switchboard
(a)	
TIMIT	0.5967 (0.0323)
TIMIT-switchboard	0.7765 (0.0180)
	Test corpus TIMIT
(b)	
Switchboard	0.7044 (0.0136)
TIMIT-switchboard	0.8291 (0.0081)

esting to observe that the correlation coefficients for TIMIT are higher than those for Switchboard when the proposed approach and OSRE as well as OSRE-model with OSRE-ftr and 19 SBEs are used. This could be because Switchboard has a wider range of speaking rates than TIMIT. Thus, large variability in speaking rate across the data might make it more difficult for the data-driven models to estimate the speaking rate accurately.

4.2.2 Cross Corpus Performance

In order to test the capability of the proposed approach in terms of the range of estimated speaking rates, the CNN-BLSTM model is trained either on TIMIT or Switchboard separately or jointly and tested on Switchboard and TIMIT, respectively. The results obtained for this analysis are provided in Table 4. From the table, it is observed that when the proposed approach is trained on both TIMIT and Switchboard and tested on either one of them, the performance is significantly better ($p < 0.01$, t-test) than that from the proposed approached trained on either TIMIT or Switchboard and tested on vice versa. This could be because of more training data. Further, from Table 4a and b it is also observed that the performance on TIMIT is better than that on Switchboard when the proposed approach is trained on both corpora jointly. In addition to more training data, this could also be because, the speaking rate range in Switchboard encompasses the speaking rate range in TIMIT. This could also be the reason for the proposed approach trained on Switchboard and tested on TIMIT to perform better than the vice versa condition. In conclusion, it suggests that the range of speaking rates estimated by the proposed approach depends on the range of speaking rates seen in the training stage.

4.2.3 Performance on Noisy Corpora

Table 5 shows the average (standard deviation) correlation coefficient obtained using the proposed approach trained on TIMIT and Switchboard, respectively, and tested on noisy condition where noise is added to the TIMIT and Switchboard, respectively. From the table, it is observed that the correlation coefficients obtained with the

Table 5 Performance of the proposed approach under additive noise conditions

		TIMIT				Switchboard			
		20 dB	10 dB	0 dB	Avg	20 dB	10 dB	0 dB	Avg
Babble	Proposed	0.8233 (0.0091)	0.7336 (0.0301)	0.3273 (0.0484)	0.6281 (0.2255)	0.7530 (0.0249)	0.7334 (0.0218)	0.6305 (0.0224)	0.7056 (0.0596)
	RSRE	0.6391 (0.0104)	0.5040 (0.0192)	0.1666 (0.0359)	0.4366 (0.2069)	0.6598 (0.0304)	0.5722 (0.0273)	0.1159 (0.0232)	0.4493 (0.2481)
F16	Proposed	0.8272 (0.0102)	0.8003 (0.0098)	0.4944 (0.0425)	0.7073 (0.1580)	0.7506 (0.0240)	0.7402 (0.0194)	0.6283 (0.0142)	0.7064 (0.0601)
	RSRE	0.6366 (0.0132)	0.4935 (0.0260)	0.01386 (0.0250)	0.4229 (0.2177)	0.6618 (0.0282)	0.5722 (0.0329)	0.0546 (0.0061)	0.4295 (0.2780)
HFC	Proposed	0.8225 (0.0089)	0.7914 (0.0094)	0.5611 (0.0449)	0.7250 (0.1233)	0.7512 (0.0245)	0.7393 (0.0221)	0.6145 (0.0267)	0.7017 (0.0652)
	RSRE	0.6312 (0.0128)	0.4774 (0.0198)	0.1126 (0.0188)	0.4071 (0.2257)	0.6567 (0.0297	0.5728 (0.0073)	0.0157 (0.0222)	0.4151 (0.3013)
Volvo	Proposed	0.8321 (0.0095)	0.8304 (0.0087)	0.8062 (0.0154)	0.8229 (0.0163)	0.7554 (0.0239)	0.7533 (0.0256)	0.7512 (0.0245)	0.7533 (0.0229)
	RSRE	0.6523 (0.0126)	0.5751 (0.0139)	0.2797 (0.0271)	0.5024 (0.1671)	0.6581 (0.0325)	0.6359 (0.0314)	0.4271 (0.0311)	0.5737 (0.1116)
AWGN	Proposed	0.8269 (0.0081)	0.8071 (0.0080)	0.7016 (0.0307)	0.7785 (0.0595)	0.7479 (0.0251)	0.7151 (0.0227)	0.5987 (0.0294)	0.6873 (0.0705)
	RSRE	0.6253 (0.0173)	0.4728 (0.0284)	0.114 (0.0222)	0.4040 (0.2228)	0.6609 (0.0345)	0.6486 (0.0272)	0.5669 (0.0230)	0.6255 (0.0507)
Average	Proposed	0.8264 (0.0091)	0.7926 (0.0240)	0.5781 (0.1728)	0.7324 (0.1496)	0.7516 (0.0225)	0.7363 (0.0240)	0.6446 (0.0598)	0.7108 (0.0614)
	RSRE	0.6369 (0.0154)	0.5045 (0.0428)	0.1623 (0.0676)	0.4346 (0.2066)	0.6595 (0.0285)	0.6004 (0.0429)	0.2361 (0.2254)	0.4986 (0.2299)

Table 6 Performance of the proposed approach on unseen corpus

	Train corpus	Unseen test corpus	
		ISLE	CTIMIT
Proposed	TIMIT	0.6558 (0.0228)	0.4495 (0.0277)
	Switchboard	0.6445 (0.0152)	0.3430 (0.0261)
	TIMIT-switchboard	0.7201 (0.0169)	0.4174 (0.0308)
RSRE	–	0.6752	0.3466

proposed approach are significantly higher ($p < 0.01$, t-test) than those obtained with RSRE technique under all noise and SNR combinations. This indicates the robustness of the proposed approach over the baseline methods. It is also observed that as the SNR decreases, the average correlation coefficient also decreases. This is because a lower SNR causes more distortions in the SBE features. Comparing the average correlation coefficient across all the noises, it is observed that the best performance is obtained under AWGN and Volvo. This could be because both the noises are more stationary than the remaining three noises. Hence, these noises affect the 19 SBEs uniformly over time and those variations are effectively minimized. Similarly, the lowest average correlation coefficient is achieved when Babble noise is added. This could be because Babble noise is the most non-stationary noise among all the noises considered. It is also interesting to observe that for Switchboard, the least average correlation coefficient is observed for 0dB AWGN noise condition.

4.2.4 Performance on Unseen Corpus

Table 6 shows the average (standard deviation) correlation coefficient obtained for ISLE and CTIMIT corpus when the proposed approach has been trained on both TIMIT and Switchboard separately and jointly. For ISLE, when the proposed approach is trained on TIMIT and Switchboard separately, it is observed that the average correlation coefficient thus obtained is comparable to that obtained from RSRE. However, when the proposed approach is trained on both TIMIT and Switchboard corpora jointly, it is observed that the average correlation coefficient improves by 0.0449 and it is significantly higher ($p < 0.01$, t-test) than that obtained using RSRE. This might be due to the training and/or the fact that the range of speaking rate in ISLE corpus is better represented by both TIMIT and Switchboard data jointly.

Considering CTIMIT, it is observed that the average correlation coefficient obtained with the proposed approach trained on TIMIT is significantly higher ($p < 0.01$, t-test) than the correlation coefficient obtained through the RSRE technique. Furthermore, it is observed that when the proposed approach is trained on Switchboard corpus or TIMIT and Switchboard corpora, the average correlation coefficient is lesser than that obtained when the proposed approach is trained on TIMIT corpus alone. This is because CTIMIT is linguistically similar to the TIMIT corpus and is obtained by rerecording the utterances in TIMIT under noisy environments.

In general, in all cases, the performance on ISLE is found to be significantly greater than that on CTIMIT. This could be because the utterances in CTIMIT corpus were

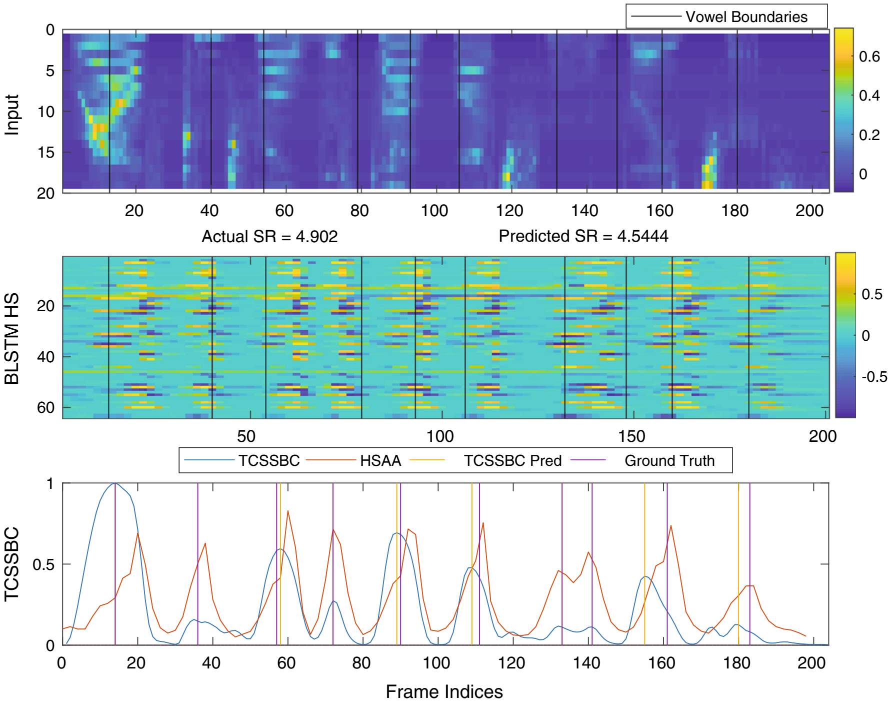

Fig. 5 Illustration of CNN-BLSTM hidden state representations

recorded in real noisy conditions which might disturb the spectro-temporal information in 19 SBEs. Overall, the correlation coefficients obtained on both ISLE and CTIMIT with the proposed approach are higher than those with RSRE technique when the proposed model is trained with both TIMIT and Switchboard corpora. This indicates the effectiveness of the proposed method under unseen corpus as well as real noisy conditions.

4.2.5 Illustration of Learned Representations by CNN-BLSTM

In order to understand the representations learned by the CNN-LSTM model, we analyze the hidden state outputs of the last BLSTM layer. Figure 5 illustrates the learned representation from the CNN-BLSTM model for an example utterance. The reference input SBEs to the CNN-BLSTM model are plotted in the top sub-figure along with the corresponding vowel boundaries. The hidden state (HS) outputs of the last BLSTM layer is plotted in the second sub-figure, where color intensity variations indicate the output values. Since the activation used is tanh, the output values are bound between +1 and −1. To verify whether the representations learned are similar to the TCSSBC, we also plotted TCSSBC contour in the last sub-figure along with hidden state outputs absolute average (HSAA). HSAA is computed by taking the absolute mean across all the hidden unit outputs. Interestingly, we observe that the peaks of TCSSBC match with the peaks of HSAA in a majority of the cases. Further, we observe that unlike TCSSBC, the heights of the HSAA peaks are more uniform and invariant to the values of the input SBEs. This could be due to the nonlinear and long-term temporal dependency modeling capability of the BLSTM, which, in turn, results in a better estimation of the speaking rate with the proposed CNN-BLSTM model.

5 Conclusions

The key difference between the proposed data-driven CNN-BLSTM-based speaking rate estimation technique and the existing knowledge-driven heuristic approaches is that the CNN-BLSTM network directly models the complex relation between the input 19 sub-band energy contours and the speaking rate, unlike feature engineering from 19 sub-band energies and peak picking in a knowledge-driven manner. The 19 sub-band energy contours are explored to obtain the spatiotemporal information that is indicative of syllable nuclei which is learnt by the CNN-BLSTM model in a data-driven manner. Experiments with four corpora, namely Switchboard, TIMIT, CTIMIT and ISLE under five additive noise conditions, reveal that speaking rate estimation with the proposed CNN-BLSTM model is more accurate than the best of the existing methods. Further investigations are required to study the use of the proposed method in the current state-of-the-art ASR systems under different noise and SNR conditions, and in the speaking rate estimation of emotional speech utterances. Future works also include developing models for speaking rate estimation directly from raw speech waveform. Additionally, transfer learning strategies need to be explored for accurate speaking rate estimation under mismatched train-test speech conditions, particularly that include native and non-native accent mismatches.

Declarations

Data availability Data sharing was not applicable to this article as no datasets were generated during the current study

References

1. S. Albawi, T.A. Mohammed, S. Al-Zawi, Understanding of a convolutional neural network, in *International Conference on Engineering and Technology (ICET)* (IEEE, 2017), pp. 1–6
2. J.D. Amerman, M.M. Parnell, Speech timing strategies in elderly adults. J. Phon. **20**(1), 65–76 (1992)
3. W. Apple, L.A. Streeter, R.M. Krauss, Effects of pitch and speech rate on personal attributions. J. Pers. Soc. Psychol. **37**(5), 715 (1979)
4. C.D. Bartels, J.A. Bilmes, Use of syllable nuclei locations to improve ASR, in *IEEE Workshop on Automatic Speech Recognition and Understanding* (2007), pp. 335–340
5. S. Bartlett, G. Kondrak, C. Cherry, On the syllabification of phonemes, in *Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2009), pp. 308–316
6. Y. Bengio, P. Simard, P. Frasconi et al., Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. **5**(2), 157–166 (1994)
7. M.P. Black, D. Bone, Z.I. Skordilis, R. Gupta, W. Xia, P. Papadopoulos, S.N. Chakravarthula, B. Xiao, M.V. Segbroeck, J. Kim, et al, Automated evaluation of non-native English pronunciation quality: Combining knowledge-and data-driven features at multiple time scales. in *Sixteenth Annual Conference of the International Speech Communication Association* (2015), pp. 493–497

8. M.P. Black, J. Tepperman, S.S. Narayanan, Automatic prediction of childrens reading ability for high-level literacy assessment. IEEE Trans. Audio Speech Lang. Process. **19**(4), 1015–1028 (2011)
9. K.L. Brown, E.B. George, CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1995), pp. 105–108
10. M.P. Caligiuri, The influence of speaking rate on articulatory hypokinesia in Parkinsonian dysarthria. Brain Lang. **36**(3), 493–502 (1989)
11. S.Y. Chang, N. Morgan, Robust CNN-based speech recognition with Gabor filter kernels, in *15th Annual Conference of the International Speech Communication Association* (2014), , pp. 905–909
12. S.M. Chu, D. Povey, Speaking rate adaptation using continuous frame rate normalization, in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (2010), pp. 4306–4309
13. C. Cucchiarini, H. Strik, 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)
14. N.H. De. Jong, R. Groenhout, R. Schoonen, J.H. Hulstijn, Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior. Appl. Psycholinguist. **36**(2), 223–243 (2015)
15. T. Dekens, M. Demol, W. Verhelst, P. Verhoeve, in *A comparative study of speech rate estimation techniques* I(nterspeech, 2007), pp. 510–513
16. T.M. Derwing, M.J. Munro, R.I. Thomson, M.J. Rossiter, The relationship between L1 fluency and L2 fluency development. Stud. Second. Lang. Acquis. **31**(4), 533–557 (2009)
17. B. Fisher, tsylb2-1.1: syllabification software. National Institute of Standards and Technology, https://www.nist.gov/itl/iad/mig/tools. Last accessed on 30–05–17 (1996)
18. K.J. Geras, A.R. Mohamed, R. Caruana, G. Urban, S. Wang, O. Aslan, M. Philipose, M. Richardson, C. Sutton, Blending LSTMs into CNNs, in *ICLR Workshop* (2016)
19. J.J. Godfrey, E.C. Holliman, J. McDaniel, SWITCHBOARD: telephone speech corpus for research and development, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1992), pp. 517–520
20. A. Graves, S. Fernández, J. Schmidhuber, Bidirectional LSTM networks for improved phoneme classification and recognition, in *International Conference on Artificial Neural Networks* (Springer, 2005), pp. 799–804
21. A. Graves, A.R. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013), pp. 6645–6649
22. A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. **18**(5–6), 602–610 (2005)
23. P. Green, J. Carmichael, A. Hatzis, P. Enderby, M. Hawley, M. Parker, Automatic speech recognition with sparse training data for dysarthric speakers, in *Eight European Conference on Speech Communication and Technology* (2003), pp. 3321–3324
24. K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, J. Schmidhuber, LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. **28**(10), 2222–2232 (2016)
25. C. Heinrich, F. Schiel, Estimating speaking rate by means of rhythmicity parameters. Interspeech (2011), pp. 1873–1876
26. S. Hochreiter, J. Schmidhuber, Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
27. J. Holmes, The JSRU channel vocoder, in *IEEE (Communications, Radar and Signal Processing)*, vol. 127 (IET, 1980), pp. 53–60
28. Z. Hu, Y. Li, Z. Yang, Improving convolutional neural network using pseudo derivative ReLU, in *Fifth International Conference on Systems and Informatics (ICSAI)*. (IEEE, 2018), pp. 283–287
29. Z. Huang, M. Dong, Q. Mao, Y. Zhan, Speech emotion recognition using CNN, in *22nd ACM International Conference on Multimedia* (ACM, 2014), pp. 801–804
30. M. Huckvale, Speech filing system: tools for speech research. http://www.phon.ucl.ac.uk/resource/sfs (2000)
31. S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in *International Conference on Machine Learning* (2015), pp. 448–456
32. S.K. Jemni, Y. Kessentini, S. Kanoun, J.M.Ogier, Offline Arabic handwriting recognition using BLSTMs combination, in *13th IAPR International Workshop on Document Analysis Systems (DAS)* (IEEE, 2018), pp. 31–36

33. Y. Jiao, M. Tu, V. Berisha, J. Liss, Online speaking rate estimation using recurrent neural networks, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), pp. 5245–5249
34. N. Jmour, S. Zayen, A. Abdelkrim, Convolutional neural networks for image classification, in *International Conference on Advanced Systems and Electric Technologies (ICASET)* (IEEE, 2018), pp. 397–402
35. A. Jongman, R. Wayland, S. Wong, Acoustic characteristics of English fricatives. J. Acoust. Soc. Am. **108**(3), 1252–1263 (2000)
36. R.D. Kent, J.C. Rosenbek, Acoustic patterns of apraxia of speech. J. Speech Lang. Hear. Res. **26**(2), 231–249 (1983)
37. S.H. Kim, G.T. Han, 1D CNN based human respiration pattern recognition using ultra wideband radar, in *International Conference on Artificial Intelligence in Information and Communication (ICAIIC)* (IEEE, 2019), pp. 411–414
38. S. Kiranyaz, T. Ince, O. Abdeljaber, O. Avci, M. Gabbouj, 1D Convolutional neural networks for signal processing applications, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), pp. 8360–8364
39. S. Kitaazawa, H. Ichikawa, S. Kobayashi, Y. Nishinuma, Extraction and representation rhythmic components of spontaneous speech, in *Fifth European Conference on Speech Communication and Technology* (1997), pp. 641–644
40. B. Ko, H.G. Kim, H.J. Choi, Controlled dropout: a different dropout for improving training speed on deep neural network, in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2017), pp. 972–977
41. C. Laurent, G. Pereyra, P. Brakel, Y. Zhang, Y. Bengio, Batch normalized recurrent neural networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2657–2661 (2016)
42. S. Lawrence, C.L. Giles, A.C. Tsoi, A.D. Back, Face recognition: a convolutional neural–network approach. IEEE Trans. Neural Netw. **8**(1), 98–113 (1997)
43. D. Li, J. Zhang, Q. Zhang, X. Wei, Classification of ECG signals based on 1D convolution neural network, in *IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)* (2017), pp. 1–6
44. J. Li, Y. Shen, Image describing based on bidirectional LSTM and improved sequence sampling, in *IEEE 2nd International Conference on Big Data Analysis (ICBDA)* (2017), pp. 735–739
45. Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimed. **16**(8), 2203–2213 (2014)
46. H. Martens, G. Van Nuffelen, M. De Bodt, T. Dekens, L. Latacz, W. Verhelst, Automated assessment and treatment of speech rate and intonation in dysarthria, in *Seventh International Conference on Pervasive Computing Technologies for Healthcare* (ICST, Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2013), pp. 382–384
47. W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, C. Souter, The ISLE corpus of non-native spoken English, in *2000 Language Resources and Evaluation Conference* (European Language Resources Association, 2000), pp. 957–964
48. N. Miller, G. Maruyama, R.J. Beaber, K. Valone, Speed of speech and persuasion. J. Pers. Soc. Psychol. **34**(4), 615 (1976)
49. N. Morgan, E. Fosler, N. Mirghafori, Speech recognition using on-line estimation of speaking rate. Fifth Eur. Conf. Speech Commu. Technol. **4**, 2079–2082 (1997)
50. N. Morgan, E.Fosler-Lussier, Combining multiple estimators of speaking rate, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (1998), pp. 729–732
51. S. Nagesh, C. Yarra, O.D. Deshmukh, P.K. Ghosh, A robust speech rate estimation based on the activation profile from the selected acoustic unit dictionary, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), pp. 5400–5404
52. S.L. Oh, E.Y. Ng, R. San Tan, U.R. Acharya, Automated diagnosis of Arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Comput. Biol. Med. **102**, 278–287 (2018)
53. D. Palaz, M.M. Doss, R. Collobert, Convolutional neural networks-based continuous speech recognition using raw speech signal, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), pp. 4295–4299

54. T. Pfau, G. Ruske, Estimating the speaking rate by vowel detection, in *IEEE International Conference on Acoustics, Speech, and Signal Proessing (ICASSP)* (1998), pp. 945–948
55. M. Richardson, M. Hwang, A. Acero, X. Huang, Improvements on speech recognition for fast talkers, in *Sixth European Conference on Speech Communication and Technology* (1999), pp. 411–414
56. S. Santurkar, D. Tsipras, A. Ilyas, A. Madry, How does batch normalization help optimization? in *Advances in Neural Information Processing Systems* (2018), pp. 2483–2493
57. D. Talkin, A robust algorithm for pitch tracking (RAPT). Speech Coding Synth. **495**, 518 (1995)
58. A. Varga, 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)
59. D. Wang, S. Narayanan, Speech rate estimation via temporal correlation and selected sub-band correlation, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2005), pp. 413–416
60. D. Wang, S.S. Narayanan, Robust speech rate estimation for spontaneous speech. IEEE Trans. Audio Speech Lang. Process. **15**(8), 2190–2201 (2007)
61. C. Yarra, O.D. Deshmukh, P.K. Ghosh, A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection. Speech Commun. **78**, 62–71 (2016)
62. S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, Z. Deng, S. Lee, S. Narayanan, C. Busso, An acoustic study of emotions expressed in speech, in *Eighth International Conference on Spoken Language Processing* (2004), pp. 2193–2196
63. J. Yuan, W. Lai, C. Cieri, M. Liberman, *Using Forced Alignment for Phonetics Research* (Text, Speech and Language Technology. Springer, Chinese Language Resources and Processing, 2018)
64. J. Yuan, M. Liberman, Robust speaking rate estimation using broad phonetic class recognition, in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (2010), pp. 4222–4225
65. Y. Zhang, J.R. Glass, Speech rhythm guided syllable nuclei detection, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (2009), pp. 3797–3800
66. J. Zhao, X. Mao, L. Chen, Learning deep features to recognise speech emotion using merged deep CNN. IET Signal Proc. **12**(6), 713–721 (2018)
67. J. Zheng, H. Franco, A. Stolcke, Rate-of-speech modeling for large vocabulary conversational speech recognition, in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)* (2000), pp. 145–149
68. M. Zihlmann, D. Perekrestenko, M. Tschannen, Convolutional recurrent neural networks for electrocardiogram classification, in *2017 Computing in Cardiology (CinC)* (IEEE, 2017), pp. 1–4
69. V. Zue, S. Seneff, J. Glass, Speech database development at MIT: TIMIT and beyond. Speech Commun. **9**(4), 351–356 (1990)