Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 78 (2016) 62–71

www.elsevier.com/locate/specom

A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection

Chiranjeevi Yarra[a], Om D. Deshmukh[b], Prasanta Kumar Ghosh[a,*]

[a] Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India
[b] Xerox Research Center India, Bangalore 560103, India

Received 30 April 2015; received in revised form 30 December 2015; accepted 23 January 2016
Available online 3 February 2016

Abstract

Acoustic feature based speech (syllable) rate estimation and syllable nuclei detection are important problems in automatic speech recognition (ASR), computer assisted language learning (CALL) and fluency analysis. A typical solution for both the problems consists of two stages. The first stage involves computing a short-time feature contour such that most of the peaks of the contour correspond to the syllabic nuclei. In the second stage, the peaks corresponding to the syllable nuclei are detected. In this work, instead of the peak detection, we perform a mode-shape classification, which is formulated as a supervised binary classification problem – mode-shapes representing the syllabic nuclei as one class and remaining as the other. We use the temporal correlation and selected sub-band correlation (TCSSBC) feature contour and the mode-shapes in the TCSSBC feature contour are converted into a set of feature vectors using an interpolation technique. A support vector machine classifier is used for the classification. Experiments are performed separately using Switchboard, TIMIT and CTIMIT corpora in a five-fold cross validation setup. The average correlation coefficients for the syllable rate estimation turn out to be 0.6761, 0.6928 and 0.3604 for three corpora respectively, which outperform those obtained by the best of the existing peak detection techniques. Similarly, the average *F*-scores (syllable level) for the syllable nuclei detection are 0.8917, 0.8200 and 0.7637 for three corpora respectively.

Keywords: Speech rate estimation; Syllable nuclei detection; TCSSBC; Mode-shape classification; Mode-shape feature vectors; Peak detection.

1. Introduction

Speech rate estimation and syllable nuclei detection are important problems in the areas of automatic speech recognition (ASR), computer assisted language learning (CALL) and fluency analysis. The ASR accuracy has been shown to improve by using the speech rate and syllable nuclei information in the recognition model (Bartels and Bilmes, 2007; Morgan et al., 1997) . In CALL, the features used for fluency analysis (Cucchiarini et al., 2000) or non-nativeness analysis (Hönig et al., 2012) are based on one or more combinations of speech rate and syllable nuclei locations. The problems of speech rate and syllable nuclei detection are closely related. The speech rate is typically estimated by counting the number of speech units per second. Most of the existing works in the literature use syllable as the speech unit (Heinrich and Schiel, 2011; Morgan et al., 1997; Wang and Narayanan, 2007). The speech rate estimation typically involves identification of the syllable nuclei locations followed by syllable rate computation (Reddy et al., 2013). Generally the approaches for the speech rate estimation and the syllable nuclei detection are based on either acoustic features (Heinrich and Schiel, 2011; Morgan et al., 1997; Reddy et al., 2013; Wang and Narayanan, 2007) or hidden Markov model (HMM) based recognition systems (Cincarek et al., 2009; Cucchiarini et al., 2000; Hönig et al., 2012; Yuan and Liberman, 2010).

The HMM based methods involve the identification of the phoneme/syllable boundaries using an ASR system. The estimated boundaries are then used to compute the syllable rate. HMM based approaches are used in the applications related to CALL where a good quality speech rate estimation is

[*] Corresponding author. Tel.: +91 80 2293 2694; fax: +91 80 2360 0444.
E-mail addresses: chiranjeevi.yarra@ee.iisc.ernet.in, chiranjeevi.yarra@gmail.com (C. Yarra), prasantg@ee.iisc.ernet.in (P.K. Ghosh).

http://dx.doi.org/10.1016/j.specom.2016.01.004
0167-6393/© 2016 Elsevier B.V. All rights reserved.

essential (Cucchiarini et al., 2000; Deshmukh et al., 2008; Hönig et al., 2012; Witt, 1999). However for accurate speech rate estimation, methods based on HMM are time consuming particularly when the reference transcription is not available and, hence, often not useful in real time applications (Wang and Narayanan, 2007). In contrast to the HMM based methods, the acoustic feature based methods are computationally less expensive (Morgan and Fosler-Lussier, 1998). The acoustic feature based methods are typically developed using acoustic properties of the vowels, which in general correspond to the syllable nuclei. Therefore, the vowel rate corresponds directly to the syllable rate (Pfau and Ruske, 1998; Yuan and Liberman, 2010).

A typical approach for estimating syllable nuclei locations, which is also useful for estimating syllable rate, involves two steps – (1) computing a short-time feature contour such that most of the peaks corresponding to the syllable nuclei locations, (2) detecting the peaks belonging to the syllable nuclei. Pfau and Ruske (1998) estimated the vowel locations based on prominent peaks in smoothed loudness contour. They proposed a peak identification strategy based on the steepness information around the local maxima. Zhang and Glass (2009) proposed a contour based on Hilbert envelope and used a rhythm guided peak counting to estimate the syllable nuclei. De Jong and Wempe (2009) used intensity based envelope with simple peak counting based on voicing decisions to estimate speaking rate. Landsiedel et al. (2011) proposed a contour based on long short term memory neural networks and identified peaks based on the region based selection above a threshold limit. Wang and Narayanan (2005, 2007) introduced a method by proposing a feature contour "temporal correlation and selected sub-band correlation (TCSSBC)", which involves computing a spectral and a temporal correlation; they also proposed a peak detection strategy which involves smoothing and a thresholding mechanism. A comprehensive comparative study of eight different methods for speech rate estimation has been summarized by Dekens et al. (2007), who found that the TCSSBC method performs the best for speaking rate estimation.

The methods addressed in the literature for both the problems focus on the feature computation as well as on the peak detection strategies. Wang et al improved the speech rate estimation accuracy by optimizing parameters in the TCSSBC feature contour computation and using a robust peak detection strategy (Wang and Narayanan, 2007). A modified version of the peak detection strategy is used by Reddy et al. (2013) along with perceptually motivated features. A neural network based syllabic peak detection was proposed by Howitt (2000). Most of the existing peak detection strategies are typically heuristic and rule based. A generic formulation for syllabic peak detection is necessary to overcome the limitations of the rule based approaches. We observe that the rule based peak detection strategies often fail to detect target peaks mainly because the target peaks do not always satisfy the heuristically designed rules. In that direction Jiao et al. (2015) proposed a convex optimization based speech rate estimation to avoid dependency on heuristic peak detection strategy. Faltlhauser et al. (2000) used the Gaussian mixture model (GMM) for classification of speaking rate into three categories – slow, medium and fast. Following this, they used the class probabilities to estimate speaking rate with the help of Neural Networks.

We, in this work, use TCSSBC as a short-time feature contour and perform mode-shape classification. In the vicinity of syllable nuclei locations TCSSBC contour typically has local maxima (Dekens et al., 2007; Howitt, 2000; Wang and Narayanan, 2007). Therefore, almost all syllables correspond to the peaks in the TCSSBC feature contour. However, some of the peaks corresponding to the syllable nuclei (referred to as syllabic peaks) are often less prominent compared to the peaks that do not correspond to any syllable (non-syllabic peaks). We hypothesize that the contour shape around each mode of the TCSSBC contour could be used for robust detection of target TCSSBC peaks.

We use a support vector machine (SVM) based binary classification method for distinguishing the syllabic mode-shapes from the non-syllabic ones. Note that although one mode-shape carries information about only one peak, we use the term 'mode-shape' instead of 'peak' because we exploit the shape of the TCSSBC feature contour around the peak for the binary classification. We propose different feature vectors spanning across multiple modes to represent each mode-shape of the TCSSBC feature contour. We also propose an automatic way of labeling each mode-shape – syllabic and non-syllabic – for training the SVM classifier. The effectiveness of the proposed mode-shape classification (MSC) approach is demonstrated using three large corpora, namely, Switchboard, TIMIT and CTIMIT. Experiments for both speech rate estimation and syllabic nuclei detection are performed on each corpus. The proposed MSC based syllabic peak detection approach achieves better performance in comparison to the best of the existing methods for both speech rate estimation and syllable nuclei detection.

The rest of the paper is organized as follows: Section 2 describes the corpora details, Section 3 discusses the details of the proposed MSC approach including TCSSBC feature contour computation, smoothing, mode-shape feature vector computation, labeling and classification procedures. Section 4 includes the experimental setup, results on various corpora and discussions. The conclusions are summarized in Section 5.

2. Database

We use ICSI Switchboard (Godfrey et al., 1992), TIMIT (Zue et al., 1990) and CTIMIT (Brown and George, 1995) 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 rate, 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. TIMIT is a read speech database, which has phonetically balanced 6300 sentences spoken by 630 speakers with a speech rate ranging

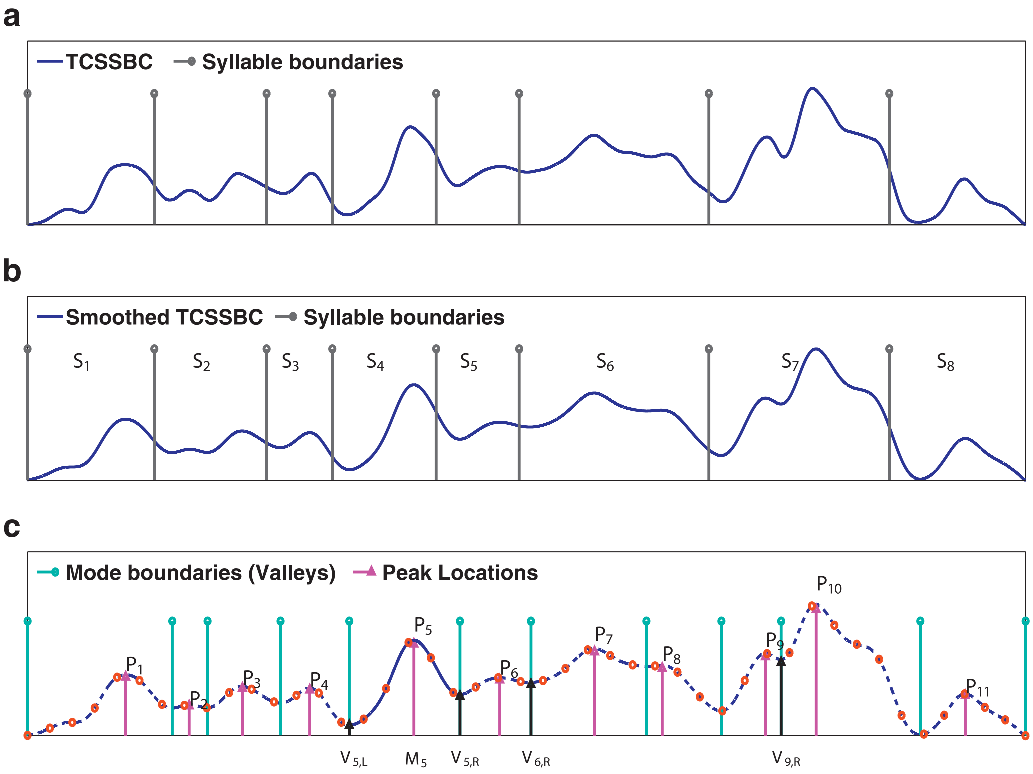

Fig. 1. An illustrative example describing location of the modes with respect to the syllable boundaries – (a) TCSSBC feature contour, (b) smoothed TCSSBC feature contour, (c) peaks and valleys in the smoothed TCSSBC feature contour. The syllables and peaks are indicated by S_i and P_i, respectively. The peak P_5 is located at M_5 and has nearest left neighboring minima is at $V_{5,\ L}$ and right neighboring minima is at $V_{5,\ R}$. The red dots indicate samples of the contour. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

from 1.44 to 8 syllables per second. All sentences from the 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 3370 sentences, 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. Using the available phonetic transcriptions, silences in the initial and final parts of each sentence of all corpora are removed.

3. Proposed mode-shape classification technique

The TCSSBC feature contour typically has peaky nature near the syllable nuclei. Fig. 1 shows a TCSSBC feature contour, its peaks and original syllable boundaries for an example speech segment taken from the Switchboard corpus having transcription *so she's a genuine chowperd*. There are a total of eight syllables (indicated by $S_i, i = 1, 2, \ldots 8$) whose boundaries are shown on the TCSSBC and smoothed TCSSBC feature contours in Fig. 1(a) and (b), respectively. The peaks of the smoothed TCSSBC contour are indicated by $P_i, i = 1, 2, \ldots, 11$ in Fig. 1(c). It is clear that there are one or multiple peaks within a syllable. The syllables with more than one peak are S_2, S_6 and S_7. In the syllable S_2, the value of the TCSSBC at P_2 (referred to as peak strengths) is lower than P_3; similarly P_8 and P_9 are lower than P_7 and P_{10} respectively in syllables S_6 and S_7. P_3, P_7 and P_{10} correspond to the syllable nuclei of S_2, S_6 and S_7 respectively and hence, they are referred as syllabic peaks while P_2, P_8 and P_9 are referred as non-syllabic peaks. Note that P_1, P_4, P_5, P_6, P_{11}, are syllabic peaks. While a syllabic peak is, in general, higher than the non-syllabic peaks within a syllable, it is not so across syllables. For example, the syllabic peak P_4 in S_3 is lower than a non-syllabic peak P_9 in S_7.

In the existing techniques for syllabic peak detection, the peaks of the TCSSBC feature contour corresponding to the syllable nuclei are typically identified by one or more combinations of the following rules – (1) using simple peak counting process (Dekens et al., 2007; Morgan and Fosler-Lussier, 1998), (2) thresholding the height of a peak relative to its larger neighboring minima (Wang and Narayanan, 2007), (3) thresholding the distance in time between two neighboring peaks (Wang and Narayanan, 2007). Thus, it is clear that most of these strategies are based on the heights and locations of the peaks of the TCSSBC feature contour. However, these strategies often fail to detect the target peaks. For example, rule 1 would fail to remove P_9 and keep P_3 as well as P_{10} with one threshold value. Similarly one threshold in rule 2 won't distinguish the peak P_9 from P_6, because the ratio of the heights of P_9 and its neighboring largest minimum ($V_{9,\ R}$) is more than the ratio of the heights of P_6 and its neighboring largest minimum ($V_{6,\ R}$). Considering (P_3, P_4) and (P_7, P_8) peak pairs, it is seen that rule 3 does not work since the gap between P_3 and P_4 is lower than that between P_7 and P_8

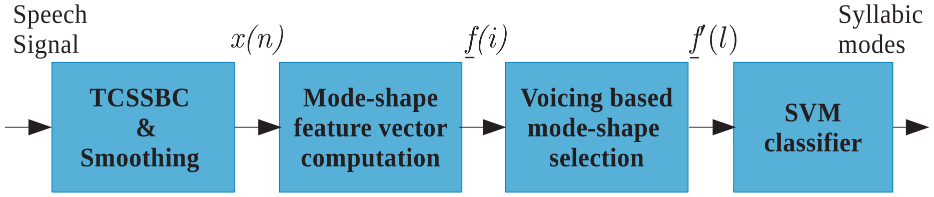

Fig. 2. Block diagram summarizing the steps of the proposed mode-shape classification (MSC) approach.

although P_3 and P_4 correspond to different syllables while P_7 and P_8 fall within one syllable. Thus rule based syllabic peak detection may not be always robust to the variations in the peak heights and locations. In contrast to employing several rules on the peak heights and locations, we hypothesize that the shape of the TCSSBC feature contour around each of its modes would be indicative of the syllabic peaks. We propose a multi-dimensional feature vector to represent each mode-shape, which are finally used for mode-shape classification.

We define a mode-shape as the segment of a TCSSBC feature contour between two consecutive valley locations of the contour. An exemplary mode-shape is indicated with a solid blue line in Fig. 1(c). The highlighted mode-shape is from location $V_{5,\,L}$ to $V_{5,\,R}$ containing the peak P_5. The multi-dimensional feature vector is proposed to capture the shape of the TCSSBC contour around each mode.

The proposed MSC method has four major stages as shown in Fig. 2. The first stage computes the TCSSBC feature contour from a speech signal followed by smoothing of the contour using a low-pass filter. The smoothed TCSSBC feature contour is denoted by $x(n)$, where n is the frame index. In the second stage, we represent the ith$(1 \leq i \leq Q)$ mode-shape of $(x(n))$ using a D-dimensional feature vector ($\underline{f}(i)$), where Q is the total number of modes in $x(n)$. In the third stage, a subset of mode-shapes are selected from the entire set ($\{\underline{f}(i);\ i \in \{1, 2, \ldots, Q\}\}$) based on the voicing decision. The selected subset of mode-shape feature vectors are denoted by $\underline{f}'(l), 1 \leq l \leq \bar{Q}$, where $\bar{Q}$ $(\leq Q)$ is the number of modes that fall within the voiced segments of the speech signal. In the last stage, SVM based binary classifier is used to categorize the selected mode-shapes into two groups – one representing the syllable nuclei and other comprising the remaining ones. The details of the all stages are given in the following subsections.

3.1. TCSSBC and smoothing

We have computed the TCSSBC feature contour following the steps outlined by Wang and Narayanan (2005) as follows:

1. 19 short-time sub-band energy contours ($y_i(n)$; $1 \leq i \leq 19$, where n is frame index) corresponding to non-uniform filter banks (Holmes, 1980) are computed using steps outlined by Huckvale (2000).
2. Using Eq. (1), temporal correlation is computed on each windowed sub-band energy contour ($y_i^w(n) = W(n)y_i(n)$) with a window shift of one frame. $W(n)$ is a Gaussian window of length K (Wang and Narayanan, 2005).
$$z_i(n) = \frac{1}{K(K-1)} \sum_{j=0}^{K-2} \sum_{p=j+1}^{K-1} y_i^w(n+j)y_i^w(n+p) \qquad (1)$$
3. At every frame, from all temporally correlated sub-band energies ($z_i(n)$), M highest energies are selected. Using these M components, sub-band correlation is computed with Eq. (2). By using sub-band correlation the syllable peak in the contour gets boosted (Morgan and Fosler-Lussier, 1998; Wang and Narayanan, 2005).
$$x(n) = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} z_i(n)z_j(n) \qquad (2)$$

The parameters (K, variance of Gaussian window and M) in Eqs. (1) and (2) are selected following work in Wang and Narayanan (2007).

In TCSSBC computation, like all short-time windowing methods, a larger window makes the contour to lose the fine temporal details. A smaller window provides more temporal detail but makes the contour noisy. Hence it results in noisy peaks, also called spurious peaks, and in turn renders spurious mode-shapes. This makes the peak detection problem challenging (Wang and Narayanan, 2005). So these spurious mode-shapes are removed by smoothing the contour. Typical range of articulation rate is 3–15 Hz which also determines the speaking rate (Crystal and House, 1990). Thus, TCSSBC feature contour is expected to have peaks (or mode-shapes) at a similar rate. However, the spurious mode-shapes in the TCSSBC feature contour introduces the noise, which could make the peak rate beyond the range of the articulation rate. These noisy peaks are removed by applying a low-pass filter with a cut-off frequency (F_c). The best choice of F_c is determined by varying it over the range of the articulation rate and maximizing the performance of the speech rate estimation over a development set. Note that even after smoothing, the TCSSBC feature contour contains mode-shapes that do not correspond to the syllable nuclei, e.g., P_2, P_8 and P_9 in Fig. 1(b). We design feature vector for each mode-shape such that using the mode-shape features these non-syllabic mode-shapes can be distinguished from the syllabic ones.

3.2. Mode-shape feature vector computation

All mode-shapes of the TCSSBC feature contour have neither identical shape nor identical number of samples of TCSSBC. So the cardinality varies across mode-shapes. These

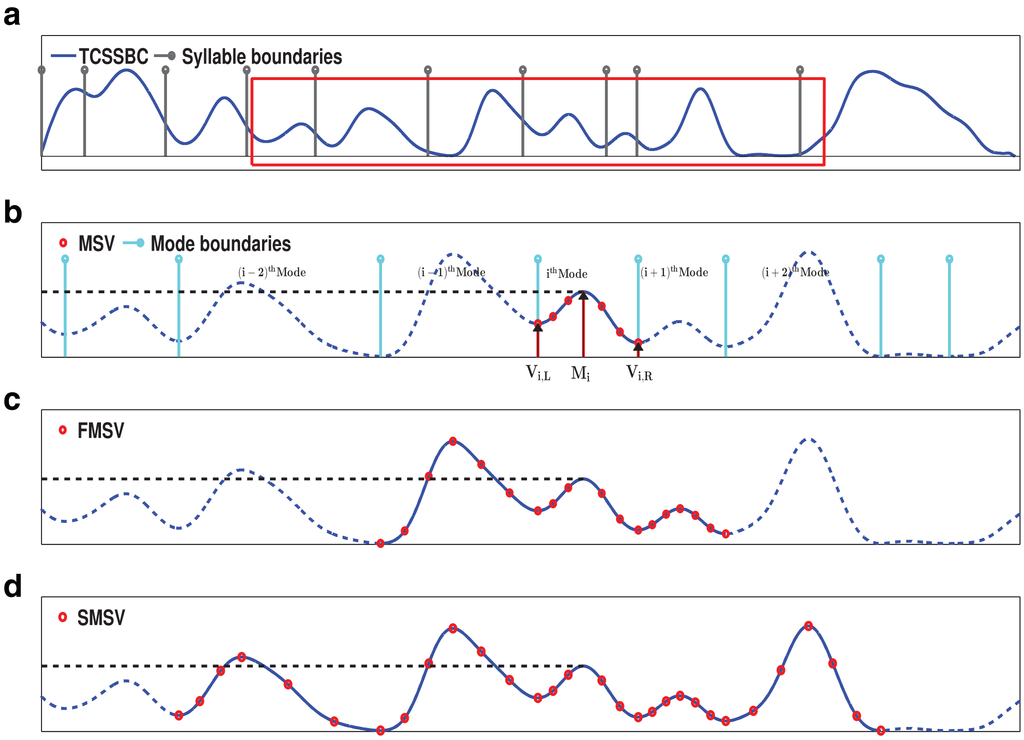

Fig. 3. Example describing the mode-shape feature vector computation. (a) TCSSBC feature contour with syllable boundaries, (b) Mode-shape vectors (MSV) with mode-shape boundaries (cyan lines), (c) first-neighborhood mode-shape feature vector (FMSV) and (d) second neighborhood mode-shape feature vector (SMSV). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

variations can be observed in Fig. 1(c); for example, mode-shapes corresponding to peaks P_1, P_5 and P_{11} have 7, 5 and 6 TCSSBC samples, respectively. The samples of a mode-shape describe the strength and shape of the respective mode. In order to compensate for the variability in the cardinality of the mode-shapes, we compute a fixed dimensional feature vector representing every mode-shape which is finally used for classification. The details of the mode-shape feature vector computation are described below.

We compute N(odd) values from every mode-shape which are used for computing a D-dimensional feature vector. The feature vector for the ith mode-shape is denoted by $\underline{f}_i$. The steps of the feature vector computation are explained in detail with the help of Fig. 3, which illustrates the TCSSBC feature contour of a sentence taken from the Switchboard corpus with transcription *everybody has their home phone number type job*. The syllable boundaries (black lines) of the sentence are indicated on the smoothed TCSSBC feature contour in Fig. 3(a). Part of the contour (marked by red rectangle) is shown in Fig. 3(b) to explain the steps of the mode-shape feature computation. The ith mode-shape has the mode locations at M_i and it spans from $V_{i,L}$ to $V_{i,R}$. The start location of the ith mode-shape is the same as the end location of the $(i-1)$th mode-shape; $V_{i,L} = V_{(i-1),R}$. It is also clear that $M_i - V_{i,L} \neq V_{i+1,L} - M_i$ due to the asymmetry in the mode-shape. The mode-shape feature vector is computed in two steps. In the first step, the left and right segments of the ith mode-shape around M_i are resampled to $(N-1)/2$ equally spaced points with spacing of $\delta_{i,L}$ and $\delta_{i,R}$ respectively. The resampling locations for the ith mode-shape are given as follows:

$$n_i(k) = \begin{cases} V_{i,L} + (k-1)\delta_{i,L}, & 1 \leq k \leq \frac{N-1}{2} \\ M_i + \left(k - \frac{N+1}{2}\right)\delta_{i,R}, & \frac{N+1}{2} \leq k \leq N \end{cases}$$
$$\delta_{i,L} = \frac{M_i - V_{i,L}}{(N-1)/2}, \; \delta_{i,R} = \frac{V_{i,R} - M_i}{(N-1)/2} \qquad (3)$$

In general, each resampling location need not match with a TCSSBC sample index. In such cases, we interpolate the TCSSBC samples to obtain the interpolated TCSSBC value $y_i(k)$ at location $n_i(k)$. This is done by using the closest TCSSBC values $(x(\lfloor n_i(k) \rfloor)$ and $x(\lceil n_i(k) \rceil))$[1] using linear interpolation technique.

The mode-shapes representing the syllable nuclei may have different peak strengths. In order to capture only the mode shape information, we divide $y_i(k)$, $1 \leq k \leq N$ by $y_i((N+1)/2)$ (value at M_i) and discard the feature value corresponding to the peak[2] of the ith mode-shape to construct a $(N-1)$-dimensional feature vector $\underline{f}_i = [\frac{y_i(1)}{y_i\left(\frac{N+1}{2}\right)}, \ldots, \frac{y_i\left(\frac{N-1}{2}\right)}{y_i\left(\frac{N+1}{2}\right)}, \frac{y_i\left(\frac{N+3}{2}\right)}{y_i\left(\frac{N+1}{2}\right)}, \ldots, \frac{y_i(N)}{y_i\left(\frac{N+1}{2}\right)}]^{\mathrm{T}}$, where T denotes the

[1] $\lfloor x \rfloor$ and $\lceil x \rceil$ are the highest and lowest integers lower and higher than x, respectively.

[2] We observed that the speech rate estimation performance degrades when the peak strength is used as a feature in addition to the mode shape. Hence, we consider only the mode shape and discard the peak strength.

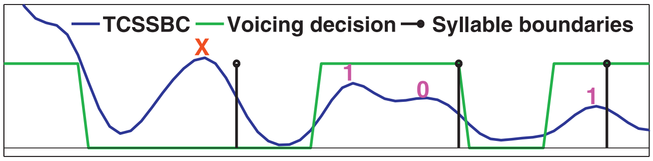

Fig. 4. An example TCSSBC feature contour segment describing the labeling approach. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

transpose operator. Fig. 3(b) shows the feature vector values (red dots) corresponding to the ith mode-shape, which capture the shape of the TCSSBC contour around the ith mode. We refer to this $D(= N - 1)$-dimensional vectors as the mode-shape feature vectors (MSV). In order to capture the shape of the TCSSBC feature contour over a longer segment around a mode, we compute a feature vector which represents the shape of a mode in relation to its neighboring mode-shapes. For this purpose, the feature values for the ith mode-shape are obtained by considering $(i - 1)$th and $(i + 1)$th mode-shapes. Each of these three mode-shapes are resampled to N values following Eq. (3). Note that $V_{i-1,R} = V_{i,L}$ and $V_{i,R} = V_{i+1,L}$. Thus, we obtain $3N - 2$ values after resampling the TCSSBC feature contour over $(i - 1)$th, ith and $(i + 1)$th mode-shapes. In order to capture only the shape information, we divide $3N - 2$ feature values by the TCSSBC value at M_i and discard the feature value corresponding the peak of the ith mode-shape (since it equals 1 after division) to construct a $3(N - 1)$-dimensional feature vector $\underline{f}_i^F = [\frac{y_{i-1}(1)}{y_i(\frac{N+1}{2})}, .., \frac{y_{i-1}(N-1)}{y_i(\frac{N+1}{2})}, \underline{f}_i^{\mathrm{T}}, \frac{y_{i+1}(2)}{y_i(\frac{N+1}{2})}, .., \frac{y_{i+1}(N)}{y_i(\frac{N+1}{2})}]^{\mathrm{T}}$. We refer to this $D(= 3N - 3)$-dimensional feature vector by the first neighborhood mode-shape feature vector (FMSV). Similarly, we compute a feature vector by resampling the TCSSBC feature contour over $(i - 2)$th, $(i - 1)$th, ith, $(i + 1)$th and $(i + 2)$th mode-shapes to a total of $5N - 4$ values and then dividing by the TCSSBC value at M_i resulting in a $D(= 5N - 5)$-dimensional feature vector, denoted by $\underline{f}_i^S = [\frac{y_{i-2}(1)}{y_i(\frac{N+1}{2})}, .., \frac{y_{i-2}(N-1)}{y_i(\frac{N+1}{2})}, \underline{f}_i^{F\,\mathrm{T}}, \frac{y_{i+2}(2)}{y_i(\frac{N+1}{2})}, .., \frac{y_{i+2}(N)}{y_i(\frac{N+1}{2})}]^{\mathrm{T}}$. We refer to this by the second neighborhood mode-shape feature vector (SMSV). The FMSV and the SMSV feature values are illustrated by red dots in Fig. 3(c) and (d), respectively.

3.3. Mode-shape labeling procedure

It is required to label each mode-shape as either a syllabic (1) or a non-syllabic (0) mode-shape to train the SVM based binary classifier. Note that while majority of the syllable segments contain one or multiple mode-shapes, the number of mode-shapes within a syllable could also be zero. The syllabic mode-shapes typically fall within a voiced segment which forms the syllable nuclei. Thus, we use voicing decisions to determine the label of each mode-shape. For illustration, we demonstrate the mode-shape labeling procedure using an exemplary TCSSBC feature contour segment in Fig. 4. The green plot indicates the voicing decisions. A segment with higher value of the plot is a voiced segment while that with lower value of the green plot is an unvoiced segment. All the mode-shapes, whose peaks belong to the unvoiced segments, are considered to be spurious and discarded (as shown by red cross in Fig. 4). If there are multiple mode-shapes within a voiced segment corresponding to a syllable nuclei, the mode-shape with the highest peak strength is labeled with 1 and the remaining (if any) mode-shapes are labeled with 0 (as shown in Fig. 4). The mode-shape feature vectors (MSV or FMSV or SMSV) thus selected using voicing decision and the corresponding labels (1 and 0) are used to train the SVM classifier which is further used to classify the mode-shapes of the TCSSBC feature contour of a test utterance.

4. Experiments and results

4.1. Experimental setup

We consider two separate objective measures for evaluating the speech rate estimation and the syllable nuclei detection. The objective measure for the speech rate estimation is the Pearson correlation coefficient (ρ) between the estimated syllable rate and the ground truth syllable rate across all test sentences. The measure for the syllable nuclei detection is the F-score which is computed following the work by Landsiedel et al. (2011). The F-score is computed at two different levels – phone level and syllable level. In the phone level computation, the vowel segment within a syllable is considered as the target segment representing the syllable nucleus. However, in the case of non-availability of phonetic boundaries (like Switchboard), syllable level F-score is computed in which syllable segments are treated as target segments. In the F-score computation, the peak locations of the estimated syllabic mode-shapes are considered as the estimated syllabic nuclei. It should be noted that the mode-shape labeling procedure used in this work does not affect the F-scores because the mode-shape labels are not used for computing F-scores. We compute the F-score by following the steps outlined by Landsiedel et al. (2011) – (a) one to one match between an estimated nucleus and a target segment is treated as correct (C_i); (b) target segments missed entirely is treated as deletions (D_i); (c) all the estimates in non-target segments are considered as insertions, but target segments containing more than one estimate are treated neither as correct nor as insertions; however, all but one estimates in these segments are evaluated as insertions (I_i). The insertions, deletions and corrects at the phone level are shown in Fig. 5 for an exemplary sentence. In the figure, target segments of each syllable are shown in the shaded regions and the estimated peaks are shown in red.

We consider the robust speech rate estimation (RSRE) technique (Wang and Narayanan, 2007) and the syllable nuclei detection using perceptually significant features (SNDPSF) technique (Reddy et al., 2013) as baselines for the speech rate estimation and the syllable nuclei detection respectively. Unlike Switchboard and TIMIT, we denoised the CTIMIT audio files using a spectral subtraction technique with smoothing constants α and β as 0.98 and 0.6,

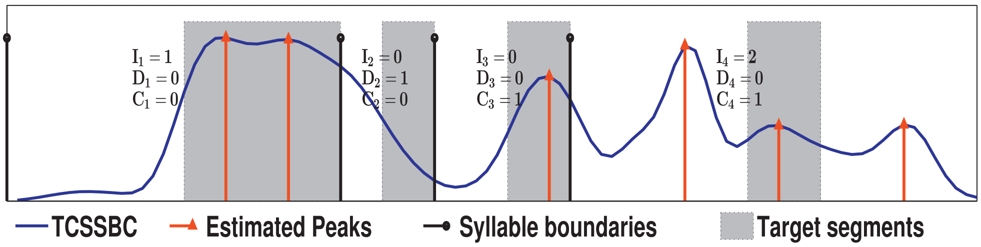

Fig. 5. An example illustrating insertions, deletions and corrects required for F-score computation. I_i, D_i and C_i denote the number of insertions, deletions and corrects for the ith syllable. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

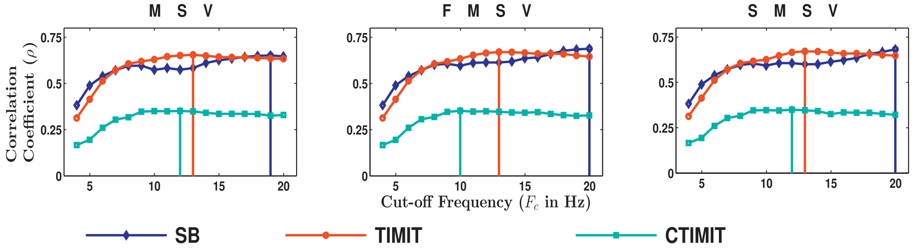

Fig. 6. Variation of the syllable rate estimation performance for different cut-off frequencies using different features in the MSC method. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

respectively (Lu and Loizou, 2008).[3] Data from each corpus is used for the experiments in a five-fold cross validation set up. We randomly divide the entire dataset into five groups out of which three are used for training, one for development and remaining one for testing. The TCSSBC feature contour is generated following the steps outlined by Wang and Narayanan (2005) for each sentence of three corpora. For smoothing TCSSBC feature contour, the cut-off frequency (F_c) of the low-pass filter is varied from 4 Hz to 20 Hz in steps of 1 Hz. We consider $N = 7$ for generating mode-shape vectors. For labeling the mode-shapes, the voicing decisions are obtained from the SFS software (Huckvale, 2000; Kleijn and Paliwal, 1995; Secrest and Doddington, 1983). The labeled set of mode-shapes are provided to the SVM classifier for training. An RBF kernel is used for SVM classifier with the complexity parameter (C) equal to 1.0. F_c for smoothing and the classifier parameters are optimized such that correlation coefficient (ρ) is maximum on the development set for each fold separately. With the optimal F_c the speech rate estimation and syllable nuclei detection are performed on all sentence of the test set. The syllable rate is computed by counting the number of syllabic mode-shapes obtained from the SVM classifier in each sentence. These mode-shapes are also used for representing the syllable nuclei.

4.2. Results and discussions

4.2.1. Optimal cut-off frequency

We compute ρ on the development set for each fold of three corpora. Fig. 6 shows the variations of ρ (averaged across all folds' development sets) with F_c for different types of feature vectors in the proposed MSC method separately. The optimal F_c values corresponding to the highest average correlation measures are indicated by vertical lines for all three corpora (blue for Switchboard, red for TIMIT and cyan for CTIMIT). For the MSV feature, these optimal F_c values vary across three corpora and this is also true for FMSV and SMSV. However, the F_c values corresponding to the highest ρ are similar across feature types for a corpus, which suggests that F_c in the proposed MSC method is more corpus dependent and less feature type dependent. The average and standard deviations of optimal F_c values across 5-folds are shown in Table 1. The standard deviations in Table 1 suggests that the optimal F_c is consistent across all folds. High standard deviation for the case of CIMIT suggests that the optimal F_c varies across folds.

Table 1
Average optimal F_c (standard deviations in brackets) in Hz across 5-folds for the speech rate estimation on the development set using the proposed MSC method for different features.

	Proposed MSC method		
	MSC (MSV)	MSC (FMSV)	MSC (SMSV)
Switch board	19 (0.7071)	20 (0.0000)	20 (0.0000)
TIMIT	13.2 (0.4472)	13.4 (0.8944)	13.4 (0.8944)
CTIMIT	11.6 (1.5166)	11.8 (2.4900)	12.2 (1.4832)

4.2.2. Speech rate estimation

We compute ρ on the test set for each fold of three corpora using RSRE as well as the proposed MSC approach with three types of feature vectors namely, MSV, FMSV and

[3] The performance on the CTIMIT is found to be worse by all algorithms compared to TIMIT and Switchboard. The denoising improves the performance.

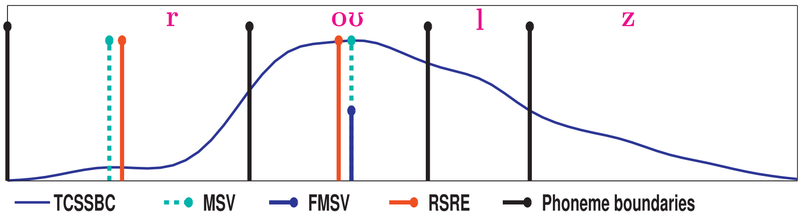

Fig. 7. Illustrative example for comparing the MSC and RSRE techniques for syllable nuclei detection. The word corresponding to the illustrated portion is '*roles*' (/r/oʊ/l/z/). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

Table 2
Average Pearson correlation coefficients (standard deviations in brackets) across 5-folds for the speech rate estimation using RSRE and the proposed MSC method using different features.

	RSRE	Proposed MSC method		
		MSC (MSV)	MSC (FMSV)	MSC (SMSV)
Switch board	0.6608 (0.0112)	0.6371 (0.0179)	**0.6761** (0.0205)	0.6701 (0.0171)
TIMIT	0.6909 (0.0191)	0.6768 (0.0157)	**0.6928** (0.0102)	0.6911 (0.0098)
CTIMIT	0.3546 (0.0409)	0.3588 (0.0347)	**0.3604** (0.0284)	0.3519 (0.0386)

Table 3
Average *F*-scores (standard deviations in brackets) across 5-folds for syllable nuclei detection using RSRE, SNDPSF and the proposed MSC using different features.

	Scheme	TIMIT	CTIMIT	Switchboard
Phone level	SNDPSF	0.6137 (0.0023)	0.5261 (0.0056)	
	RSRE	0.8321 (0.0052)	0.7113 (0.0047)	
	MSC (MSV)	0.8387 (0.0037)	**0.7238** (0.0059)	
	MSC (FMSV)	**0.8416** (0.0034)	0.7237 (0.0054)	
	MSC (SMSV)	0.8405 (0.0030)	0.7232 (0.0069)	
Syllable level	SNDPSF	0.6606 (0.0020)	0.5530 (0.0057)	0.0302 (0.0025)
	RSRE	**0.8246** (0.0018)	0.7394 (0.0024)	0.8907 (0.0024)
	MSC (MSV)	0.8170 (0.0031)	**0.7637** (0.0052)	0.8868 (0.0024)
	MSC (FMSV)	0.8200 (0.0040)	0.7608 (0.0086)	**0.8917** (0.0053)
	MSC (SMSV)	0.8189 (0.0045)	0.7620 (0.0079)	0.8885 (0.0060)

SMSV. The averages and standard deviations (shown in brackets) of ρ across all 5 folds are shown in Table 2. The highest averaged ρ value for each corpus is indicated in bold. The ρ values obtained with the proposed MSC method using the FMSV feature is significantly ($p < 0.01$, t-test) better than the baseline scheme for Switchboard corpus. ρ using the SMSV feature is also significantly ($p < 0.01$, t-test) better than the baseline scheme for Switchboard corpus. This implies that feature vector that represents the shape of a mode in relation to the neighboring mode-shapes improves the speech rate estimation compared to the shape of a mode in isolation.

4.2.3. Syllable nuclei detection

We compute the *F*-scores for the syllable nuclei detection at phone and syllable levels for each corpus. The *F*-scores averaged over all folds using the proposed MSC method along with RSRE and SNDPSF are shown in Table 3. As no phonetic transcription is available in the Switchboard corpus, we do not report *F*-score on the Switchboard in the phone level. We indicate the best *F*-score in bold for each level for all three corpora. The proposed MSC method significantly outperforms the SNDPSF ($p < 0.01$, t-test) at both the phone level as well as the syllable level for all three corpora. However it significantly ($p < 0.01$, t-test) outperforms RSRE only for CTIMIT at both phone and syllable level. The *F*-scores obtained by the proposed MSC method are not significantly different from those obtained by the RSRE method for the TIMIT and Switchboard corpora.

4.2.4. Discussions

Fig. 7 illustrates a portion of the TCSSBC contour for a sentence from the Switchboard corpus where the proposed MSC method performs better than the RSRE scheme. The illustrated portion contains only one word, namely 'roles', which has only one syllable. In the figure, we indicate the ground truth syllable nuclei at the phone level by marking the vowel boundary (black lines) and the syllable nuclei locations estimated by the RSRE (red line) as well as the MSV (cyan line) and the FMSV (blue line) of the proposed MSC method. The estimated nuclei locations using the FMSV and SMSV are identical for the illustrated segment; hence the nuclei location from the SMSV is not indicated in the figure. It is clear from Fig. 7 that the syllable nuclei location detected by FMSV correctly falls within the ground truth syllable nuclei boundaries. However, both MSV and RSRE detect an extra syllable nucleus which falls outside the ground truth syllable nuclei boundaries. This example indicates that a feature vector that captures the mode-shape along with the neighboring mode-shapes results in a better performance for the syllable nuclei detection task compared to a feature vector that only captures the shape of the target mode. This observation is consistent with that from the speech rate estimation performance (Table 2).

The performance of speech rate estimation and syllable nuclei detection are worse for CTIMIT compared to that for

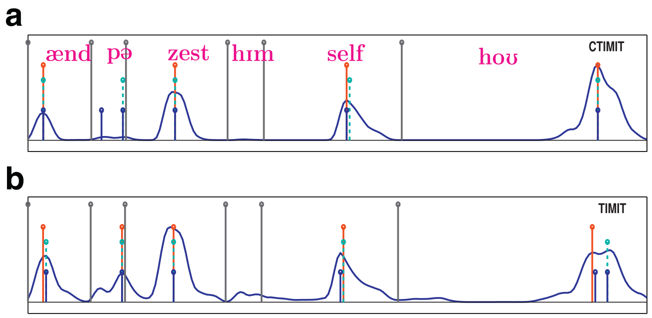

Fig. 8. Comparison of the syllable nuclei detection for an identical sentence from (a) CTIMIT and (b) TIMIT using different schemes. The sentence is "*and possessed himself how*".

TIMIT using all three feature vectors considered in this work. Although CTIMIT contains a subset of the TIMIT audio samples, they are played at the transmitter side under noisy conditions and rerecorded the same at the receiver end of a cellular channel. In order to investigate the influence of this acoustic difference on the speech rate and syllable nuclei detection in detail, we consider an identical sentence (as shown in Fig. 8) from both TIMIT and CTIMIT and compare the detected syllabic nuclei for both the cases using the RSRE and the proposed MSC method using MSV and FMSV features. From Fig. 8, it is clear that the peak strengths in the TCSSBC feature contour of the CTIMIT sentence are equal or lower compared to the corresponding peak strengths in the TCSSBC feature contour of the TIMIT sentence. In the cases of "/p/ə/" and "/h/ɪ/m/" syllables the peak strengths of the TCSSBC feature contour are not significant for the CTIMIT sentence. This could be a potential reason for the drop in speech rate estimation and syllable nuclei detection performance. It is clear that neither RSRE nor the proposed MSC method detect the fourth syllable (/h/ɪ/m/) indicating that the detection of the respective syllable is equally challenging in both CTIMIT and TIMIT. In the FMSV case the missing syllable is compensated by the extra syllables estimated at other locations (/p/ə/ in CTIMIT and /hoʊ/ in TIMIT case) in the speech rate estimation, which indicates that the false positives sometimes would improve ρ but cause *F*-score to degrade. However in the CTIMIT case, the proposed MSC method detects all the syllable peaks which are detected by all the methods in TIMIT case. On the other hand, RSRE misses one syllabic peak (/p/ə/) in CIMIT compared to that in TIMIT resulting in worse speech rate as well as syllable nuclei detection in CTIMIT compared to TIMIT. This could be because of heuristic approach of the peak detection in the RSRE method.

5. Conclusions

We propose a mode-shape classification technique for both speech rate estimation and syllable nuclei detection problems. Each mode-shape of the TCSSBC feature contour is represented by a feature vector capturing the shape of the mode in relation to the neighboring mode-shapes. Using this proposed feature vectors, each mode-shape is classified as either a syllabic or a non-syllabic mode-shape. Experiments with three large corpora namely, Switchboard, TIMIT and CTIMIT reveal that the proposed MSC technique performs better than the best of the existing methods for both speech rate estimation and syllable nuclei detection task. The hyper-parameters of the classifier are found to vary depending on the chosen corpus. Further investigation is required to develop a mode-shape feature vector that could result in a high accuracy for both the tasks in a corpus-independent manner.

References

Bartels, C.D., Bilmes, J.A., 2007. Use of syllable nuclei locations to improve ASR. In: Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding, pp. 335–340.

Brown, K.L., George, E.B., 1995. CTIMIT: a speech corpus for the cellular environment with applications to automatic speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 105–108.

Cincarek, T., Gruhn, R., Hacker, C., Nöth, E., Nakamura, S., 2009. Automatic pronunciation scoring of words and sentences independent from the non-natives first language. Comput. Speech Lang. 23 (1), 65–88.

Crystal, T.H., House, A.S., 1990. Articulation rate and the duration of syllables and stress groups in connected speech. J. Acoust. Soc. Am. 88 (1), 101–112.

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

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

Dekens, T., Demol, M., Verhelst, W., Verhoeve, P., 2007. A comparative study of speech rate estimation techniques. In: Proceedings of INTERSPEECH, pp. 510–513.

Deshmukh, O.D., Joshi, S., Verma, A., 2008. Automatic pronunciation evaluation and classification. In: Proceedings of INTERSPEECH, pp. 1721–1724.

Faltlhauser, R., Pfau, T., Ruske, G., 2000. On-line speaking rate estimation using Gaussian mixture models. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp. 1355–1358.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 517–520.

Heinrich, C., Schiel, F., 2011. Estimating speaking rate by means of rhythmicity parameters. In: Proceedings of INTERSPEECH, pp. 1873–1876.

Holmes, J., 1980. The JSRU channel vocoder. IEE Proc. F: Commun. Radar Signal Process. 127 (1), 53–60.

Hönig, F., Batliner, A., Nöth, E., 2012. Automatic assessment of non-native prosody – annotation, modelling and evaluation. In: Proceedings of International Symposium on Automatic Detection of Errors in Pronunciation Training, pp. 21–30.

Howitt, A.W., 2000. Automatic Syllable Detection for Vowel Landmarks (Ph.D. thesis). MIT, Cambridge, MA.

Huckvale, M., 2000. Speech filing system: Tools for speech research URL http://www.phon.ucl.ac.uk/resource/sfs. (accessed 11.01.16).

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

Kleijn, W.B., Paliwal, K.K., 1995. Speech Coding and Synthesis. Elsevier Science, Inc.

Landsiedel, C., Edlund, J., Eyben, F., Neiberg, D., Schuller, B., 2011. Syllabification of conversational speech using bidirectional long-short-term memory neural networks. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 5256–5259.

Lu, Y., Loizou, P.C., 2008. A geometric approach to spectral subtraction. Speech Commun. 50 (6), 453–466.

Morgan, N., Fosler, E., Mirghafori, N., 1997. Speech recognition using on-line estimation of speaking rate. In: Proceedings of European Conference on Speech Communication and Technology, Eurospeech, vol. 4, pp. 2079–2082.

Morgan, N., Fosler-Lussier, E., 1998. Combining multiple estimators of speaking rate. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 729–732.

Pfau, T., Ruske, G., 1998. Estimating the speaking rate by vowel detection. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2, pp. 945–948.

Reddy, A.A., Chennupati, N., Yegnanarayana, B., 2013. Syllable nuclei detection using perceptually significant features. In: Proceedings of INTERSPEECH, pp. 963–967.

Secrest, B.G., Doddington, G.R., 1983. An integrated pitch tracking algorithm for speech systems. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 8, pp. 1352–1355.

Wang, D., Narayanan, S.S., 2005. Speech rate estimation via temporal correlation and selected sub-band correlation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 413–416.

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

Witt, S.M., 1999. Use of Speech Recognition in Computer-assisted Language Learning (Ph.D. thesis). University of Cambridge.

Yuan, J., Liberman, M., 2010. Robust speaking rate estimation using broad phonetic class recognition. In: Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing, pp. 4222–4225.

Zhang, Y., Glass, J.R., 2009. Speech rhythm guided syllable nuclei detection. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 3797–3800.

Zue, V., Seneff, S., Glass, J., 1990. Speech database development at MIT: TIMIT and beyond. Speech Commun. 9 (4), 351–356.