AUTOMATIC DETECTION OF SYLLABLE STRESS USING SONORITY BASED PROMINENCE FEATURES FOR PRONUNCIATION EVALUATION

Chiranjeevi Yarra[1] *Om D. Deshmukh*[2] *Prasanta Kumar Ghosh*[1]

[1]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
[2]Xerox Research Center India, Bangalore, India

ABSTRACT

Automatic syllable stress detection is useful in assessing and diagnosing the quality of the pronunciation of second language (L2) learners in an automated way. Typically, the syllable stress depends on three prominence measures – intensity level, duration, pitch – around the sound unit with the highest sonority in the respective syllable. Stress detection is often formulated as a binary classification task using cues from the feature contours representing the prominence measures. We observe that cues from a feature contour obtained by incorporating relative sonority levels in the prominence measures are more indicative of the syllable stress compared to those from the feature contours representing only the prominence measures. Based on this observation, we propose a new feature contour based on temporal correlation selected sub-band correlation with an optimal set of sub-bands, called sonorous sub-bands, to maximize the stress detection accuracy. Experiments on ISLE corpus show that, for German and Italian non-native English speakers, the syllable stress detection accuracies (87.53% and 86.26%) are higher when the proposed features are used compared to the baseline accuracies (85.81% and 83.17%) indicating the effectiveness of the sonority based prominence features.

Index Terms— Sonorous TCSSBC (S-TCSSBC), syllable stress detection, sonority based features, prominence measures, forward sub-band selection

1. INTRODUCTION

Automatic detection of syllable stress has been shown to be useful for evaluating pronunciation [1] [2] [3] in several applications including computer assisted language learning. It is also useful in providing feedback to the L2 learners by automatically identifying localized pronunciation errors [4] [5] through a language learning system. In general, stressed syllables appear to be perceptually more prominent than the unstressed ones. This is reflected through the changes in the prominence measures – intensity level, duration and pitch [6] [7]. Hence, typically, automatic syllable stress detection consists of two steps [8] [5] [3] [9]. In the first step, feature contours representing prominence measures are computed from speech signal for each syllable using short-time energy, syllable/syllable nuclei duration and fundamental frequency (f0). In the second step, features are derived by computing the statistics from the feature contours. These features are further used in a binary classifier that classifies a syllable as stressed or unstressed.

A number of features have been proposed in the literature to capture the variation in the prominence measures. For example, Tamburini introduced a set of features that accounts for the variations in f0, energy and syllable duration between the stressed and unstressed syllables [8]. In addition to these features Tepperman et al. have introduced a new set of features by incorporating contextual variations of the syllable nuclei under stressed and unstressed conditions [5]. However, these features are computed to capture each prominence measure separately. Verma et al. have used a set of features that captures the variations in the prominence measures in a combined manner [3]. Using these features, Deshmukh et al. have improved the syllable stress classification accuracy by using nucleus level clustering [9]. Li et al. have incorporated perceptual attributes in the prominence measures for the purpose of automatic stress detection [10]. Ferrer et al. have introduced a set of features from spectral tilt and log posteriors from Gaussian mixture models using Mel frequency cepstral coefficients (MFCCs) along with features that depend directly on the energy, f0 and duration [4]. Shahin et al. have used deep neural networks along with features computed based on Mel & Bark scale energies, f0 and duration [11]. Further, they have used a subset of features with DNN for classifying bisyllabic lexical stress in disordered speech [12].

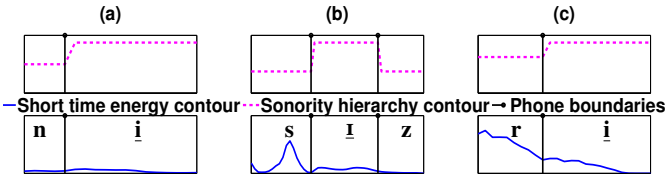

Fig. 1. Illustration of sonority and short-time energy contour for three unstressed syllables. The sonority is computed using sonorous hierarchy by dividing equal intervals between 1 and 0, considering sonority of the highest sonorous sounds (open vowels) as 1 and that of the least sonorous sounds (stops) as 0.

Unlike the existing works for syllable stress detection, we incorporate sonority motivated cues in the syllable prominence and propose a new sonority feature contour by means of spectro-temporal correlation (STC) [13] on short-time energy contours in selected sub-bands. Sonority is referred to as the carrying power of individual sounds either in a word or a longer utterance [6]. The carrying power is measured based on the sonorous hierarchy of various classes of sound. In the decreasing order of sonority they are: open vowels, close vowels, glides, liquids, nasals, fricatives, affricates and stops [6]. Among all sounds in a syllable, the highest sonorous sound represents syllable centers (typically vowels) by means of which syllable prominence is carried [6]. Hence, we hypothesize that the carrying power of these sounds should be included to reduce the variability in the cues for detecting the syllable prominence. We illustrated this variability with the help of Figure 1 for three exemplary unstressed syllables. From the figure, it is observed that the short-time energy contours have large variabilities across the unstressed syllables, but sonority contours have more consistent patterns in syllable nuclei (underlined phonemes in Figure 1) across illustrated syllables. Note that the sonorous hierarchy is independent of the stressed and unstressed syllables. Hence, direct use of sonorous hierarchy alone may not discriminate the stressed and unstressed syllables. We hypothesize that sonority cues could be combined with short-time energy reflecting the prominence measures in the feature contour, which could have lower variabilities and could discriminate the stressed and unstressed syllables better.

Sonority based features have been proposed in the literature for speech recognition and rhythm classification. For example, sonor-

ity features have been derived based on the formant structures [14] as well as regular patterns in magnitude spectrum along the time axis [15] [16]. Galves et al. proposed sonority features by computing the relative entropy between normalized spectra in consecutive frames for measuring speech rhythm variabilities [15]. However, due to normalization, it fails to include changes in the energy across utterances or words. Kocharov proposed sonority based measure for the automatic speech recognition by considering formant-like structure within each frame [14]; however, it does not consider the regular patterns in temporal domain.

We, in this work, assume that the sonority is related to the consistent temporal pattern in sub-band energies captured by STC [13] which was proposed by Wang et al. [17]. STC has been shown to be effective in exploiting the formant-like structures in the spectral domain [17] with the help of short-time energy contours of 19 sub-bands [18]. Nagesh et al. have also shown its effectiveness in capturing the regular patterns in the temporal domain with the help of the non-negative matrix factorization based activation profiles. In this work, we define sonority feature contour by STC of the sub-band energy profiles known as temporal correlation and selected sub-band correlation (TCSSBC) [17]. We assume that such a definition combines sonority cues with the short-time energy. However, TCSSBC is also known to introduce peaks at less sonorous regions, e.g., fricatives [19]. Hence, instead of using all sub-bands for TCSSBC, we select a few sub-bands to reduce its peaky nature in those regions. We call this modified TCSSBC as sonorous TCSSBC (S-TCSSBC).

In this work, we propose three sets of features for a syllable considering the statistics of the S-TCSSBC within that syllable. Further, we obtain an optimal set of features using a forward feature selection approach. We use the selected features in SVM classifier followed by a post-processing strategy for automatic syllable stress detection. Experiments are performed on ISLE [20] corpus containing polysyllabic words separately from German and Italian non-native speakers, for which the proposed approach outperforms the baseline scheme with an absolute improvement in the accuracy by 1.72% and 3.09% respectively. When the entire ISLE corpus is used for evaluation in a five-fold cross validation setup, the average accuracy turns out to be 92.87% with a standard deviation (SD) of 0.33%.

2. DATABASE

We use ISLE [20] corpus in all our experiments in this work. We use all 7834 utterances from 46 non-native speakers (23 German (GER) and 23 Italian (ITA)) learning English. Each speaker uttered approximately 160 sentences. Each utterance was phonetically aligned automatically with a forced alignment process and then those were corrected manually by a team of five linguists to reflect the speakers' pronunciation. They also labeled all the syllables with stress markings by assuring only one stressed syllable in each word. A total of 48868 syllables were marked as stressed and 16693 syllables as unstressed. We obtain the syllable transcriptions from the phone transcriptions using NIST syllabification software [21]. Based on the syllable transcriptions, we obtain the aligned syllable boundaries using aligned phone boundaries.

3. PROPOSED APPROACH

Block diagram in Figure 2 shows the five steps involved in the proposed approach. In the first step, S-TCSSBC is computed using STC on a subset of sub-bands learnt during training phase. In the second step, sentence level S-TCSSBC is divided into N syllable segments, where N is total number of syllables in the utterance. In the third step, for each syllable segment, a set of sonority based features is computed. In the fourth step, each syllable is classified as stressed

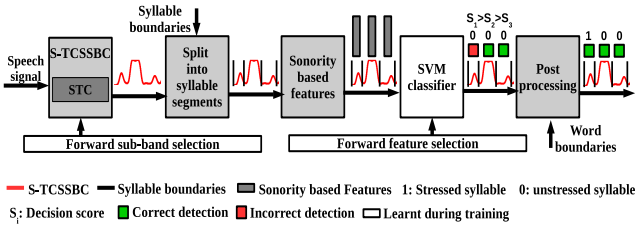

Fig. 2. Block diagram represents the steps involved in the proposed approach using a three syllabic word. In this example, all three syllables are classified as unstressed and hence the syllable with the highest score is declared as the stressed syllable following post-processing.

or unstressed with SVM classifier using a subset of features selected using forward feature selection approach. In the last step, the estimated stress markings are post-processed to ensure that each polysyllabic word has only one stressed syllable.

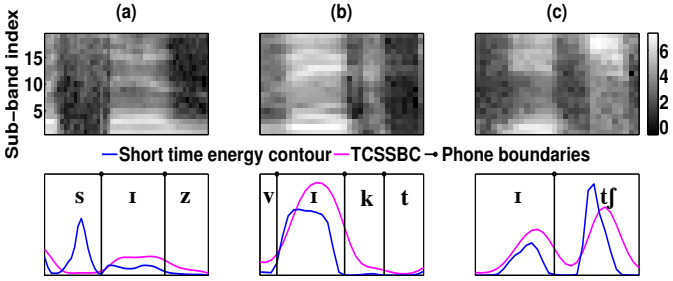

Fig. 3. Illustration of syllable segments explaining the advantage of STC in TCSSBC and the need for sub-band selection. Top row shows the energies of 19 sub-bands.

3.1. S-TCSSBC

TCSSBC is obtained by using STC on 19 sub-band energies, following the work by Wang et al. [17]. The TCSSBC of three exemplary syllable segments are shown in Figure 3 along with 19 sub-band energies used. From Figure 3a, it is observed that the TCSSBC has lower values than the short-time energy in the phoneme 's'. This is because sub-band energies are more irregular in phoneme 's' than in phoneme 'ɪ'. It is also observed that the peak strength of TCSSBC for phoneme 'ɪ' in Figure 3a is lower than that in Figure 3b. This could be because lower sub-band energy values in Figure 3a cause lower TCSSBC values than those in Figure 3b. Thus, TCSSBC appears to be more consistent in capturing syllable prominence unlike short-time energy. This is mainly due to combining the STC with short-time sub-band energies without considering short-time energy directly.

Although TCSSBC is more robust than short-time energy, we observe that it introduces peaks whenever there are consistent patterns in the sub-bands; which may not occur only in the high sonority regions such as syllable nuclei. This effect is illustrated with the help of Figure 3c using an exemplary syllable segment taken from ISLE corpus. From the figure, it is observed that the TCSSBC has the highest peak in phoneme 'tʃ'. This is due to the strong regular patterns in the 17-19th sub-band energies for 'tʃ'. This suggests that the peak strength only at high sonority regions (i.e., syllable nuclei) could be improved by removing such irrelevant sub-bands. We refer to these as non-sonorous sub-bands and eliminate them before computing the features. We identify such non-sonorous sub-bands using a forward sub-band selection approach to maximize the stress detection accuracy on the training set. We apply STC on the remaining sub-bands, called sonorous sub-bands, to compute S-TCSSBC

contour ($\mathscr{X}(m)$), where m denotes the frame index.

3.2. Sonority based feature computation

We propose a total of 20 features using S-TCSSBC in two sets (10 features in each set). For the first set, we use the syllable boundaries and refer to them as syllable level features. For the second set, we use syllable nuclei boundaries and refer to them as syllable nuclei level features. All the proposed features are divided into three categories – 1) 10-dim strength based features (SFs) , 2) 6-dim temporal variability based features (TFs), and 3) 4-dim area & duration based features (ADFs). In each category, half of the features are in syllable level and the remaining features are in syllable nuclei level.

3.2.1. Strength based features (SFs)

Typically, intensity at the stressed syllables is more than the unstressed syllables when the two other prominence measures (duration and pitch) for those syllables remain identical [6]. Since higher intensity leads to higher S-TCSSBC value, it could be used to discriminate these two classes. As strength based features, we compute *median, mean, geometric mean, range and SD* of S-TCSSBC values within each syllable segment. It is also expected that the S-TCSSBC values at the syllable nuclei are higher compared to the neighboring phonemes in the stressed syllables. Hence, we also compute the above five features within the syllable nuclei segments. Thus with these two 5-dim features, the dependence between the features in a syllable nuclei and the neighboring phonemes could be captured in a data driven manner.

3.2.2. Temporal variability based features (TFs)

In order to capture the shape of the S-TCSSBC peaks over time in a syllable, we compute three temporal domain features – *SD (σ), skewness (γ) and kurtosis (κ)* – using Equation 1. For this purpose, we consider normalized S-TCSSBC $\left(\mathscr{X}_1(m) = \mathscr{X}(m)/\sum_m \mathscr{X}(m)\right)$ as a probability mass function. However, Equation 1 depends on syllable segment lengths, which are not identical across syllables in general. In order to compensate the effect due to different durations, we resample the S-TCSSBC within each syllable to a fixed length of size 20 frames (found empirically) for computing these features. We also compute these three temporal features in the syllable nuclei level.

$$\begin{gathered}\mu = \sum_m m\mathscr{X}_1(m); \quad \sigma = \sqrt{\sum_m (m-\mu)^2 \mathscr{X}_1(m)}; \\ \gamma = \frac{1}{\sigma^3}\sum_m (m-\mu)^3 \mathscr{X}_1(m); \; \kappa = \frac{1}{\sigma^4}\sum_m (m-\mu)^4 \mathscr{X}_1(m)\end{gathered} \tag{1}$$

3.2.3. Area & duration based features (ADFs)

Similar to the intensity, the duration of a stressed syllable is typically higher than that for an unstressed syllable. This is also true for the respective syllable nuclei [6]. Hence, in general, area under S-TCSSBC could have higher values for stressed syllables than unstressed syllables. Considering duration and area under S-TCSSBC, we compute four features as follows: 1) ratio of the duration of a syllable and the duration of the word containing the syllable, 2) ratio of the duration of a syllable nuclei and the sum of the durations of all syllable nuclei in a word containing the syllable 3) ratio of the area under S-TCSSBC contour for a syllable and the area for the word containing the syllable, 4) ratio of the area under S-TCSSBC contour for a syllable nuclei and the area for the word containing the syllable. The area under a S-TCSSBC contour is computed by adding all the respective S-TCSSBC values.

3.3. Sonority based sub-band and feature selection

Sonorous sub-band selection is done to reduce peaky nature in the S-TCSSBC at the non-sonorous regions. On the other hand, feature selection is done to remove redundancy in the features. Instead of jointly doing sub-band and feature selection, we first select optimal sub-bands to obtain the proposed 20 features, which are further pruned using feature selection to maximize the stress detection accuracy. Optimal sub-bands for S-TCSSBC are obtained separately for syllable level 10-dim features and syllable nuclei level 10-dim features. The steps in the forward sub-band selection is outlined in Algorithm 1. Once the optimal sub-bands are obtained separately for syllable and syllable nuclei level features, forward feature selection is performed on all 20 proposed features following a feature selection algorithm used in the work by Prasad et al. [22].

Algorithm 1 Forward sub-band selection – inputs: $S = [s_1, s_2, ..., s_k]$ (all 19 sub-bands) and Ω (class labels).

Initialization: $S^s = \Phi$. $\mathscr{P}, \eta$ are as empty vectors. $\mathscr{I} = \{1, 2, ..., K\}$
for $l = 1 \text{to} K$ **do**
  for $i \in \mathscr{I}$ **do**
    $\mathscr{X} \leftarrow$ Compute STC using $[S^s \;\; s_i]$
    $\mathscr{F} \leftarrow$ Compute features using $\mathscr{X}$
    $\zeta_i \leftarrow$ Classification accuracy using $\mathscr{F}$ and Ω
  end for
  $\mathscr{P}_l \leftarrow \max_i \zeta_i \qquad \eta_l \leftarrow \arg\max_i \zeta_i \qquad S^s \leftarrow [S^s \;\; S_{\eta_l}]$
  $\mathscr{I} \leftarrow \mathscr{I} \setminus \eta_l$
end for
Return $\mathscr{P}$ and η

3.4. Post processing

The ISLE corpus was labeled assuming that each polysyllabic word has only one stressed syllable [5] [20]. To ensure this, we perform post-processing on all polysyllable words after every syllable is classified individually using SVM classifier. For the post-processing, we use the estimated labels and decision scores [23] from SVM classifier. In case the number of predicted stressed syllables in a word is different from one, we declare the syllable with the highest score as the stressed syllable.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

We consider unweighted accuracy (UA) and weighted accuracy (WA) as objective measures for evaluating the proposed approach. We consider the work by Tepperman et al. [5] as the baseline technique. We perform the experiments under two setups – 1) five fold cross validation setup 2) baseline setup [5]. In the five fold cross validation setup, we use three folds for training, one fold for forward feature selection and one fold for testing. We find the optimal sub-bands using one fold selected randomly from training set, in which half of data is selected (randomly) for SVM training and remaining for selecting the sub-bands. Once the optimal sub-bands are learnt,

all three folds in the training set are used to train the SVM classifier for feature selection. We use UA as the criteria for both sub-band and feature selection. In the baseline setup, we consider groups of the data from GER and ITA non-native speakers containing only polysyllabic words. Following the work of the baseline scheme [5], we use 1st-12th & 1st-13th speakers data for training and 13th-23rd & 14th-23rd speakers data for testing for GER & ITA respectively. We select the STC parameters identical to those in the work by wang et al. [24]. We use SVM classifier with RBF kernel for the classification task with the complexity parameter (C) equal to 1.0 and with kernel coefficient (γ) equal to 1/number of features. SVM classifier is implemented using Scikit-learn [25].

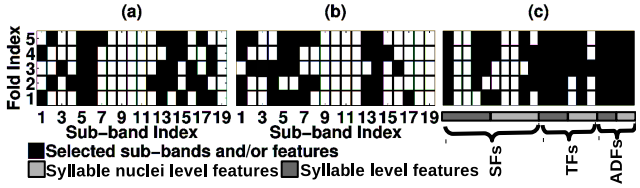

Fig. 4. Optimal sub-band and feature selection in five folds: a) optimal sub-bands for syllable level features, (b) optimal sub-bands for syllable nuclei level features, (c) optimal features.

4.2. Results and discussions

Figure 4 shows the optimal sub-bands and features selected using forward selection algorithm. From Figure 4a & b, it is observed that 9th-11th and 19th bands are consistently detected as non-sonorous sub-bands across all folds for both syllable and syllable nuclei level features. In addition, 8th and 12th & 18th sub-bands are detected as non-sonorous sub-bands for syllable and syllable nuclei level features respectively. On the other hand, 5th & 6th and 14th sub-bands are found to be sonorous sub-bands for syllable and syllable nuclei level features respectively. Otherwise the sonorous sub-bands vary across folds suggesting that the optimal sub-bands are data dependent. For comparison with baseline, we take union of all sonorous sub-bands in all folds – 1st-7th & 12th-18th sub-bands for syllable level features and 1st-8th & 13th-17th for syllable nuclei level features. We compute UA and WA using 10-dim syllable level features, 10-dim syllable nuclei level features and the combined 20-dim features. From Figure 5, it is observed that the UAs & WAs obtained using features from proposed S-TCSSBC are significantly ($p < 0.01$) higher than those from TCSSBC at all three levels – syllable, syllable nuclei and combined. These improvements with respect to TCSSBC suggests that the sonorous sub-bands are critical for accurate stressed syllable detection.

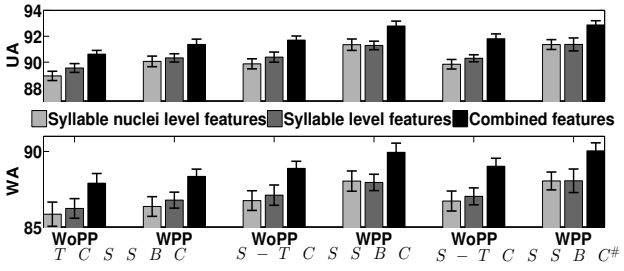

Fig. 5. Average accuracies (both UA and WA) across five folds without and with post processing (WoPP and WPP) (SD is indicated by error bars) using 20 features from TCSSBC, 20 features from S-TCSSBC and 16 selected features from S-TCSSBC (S-TCSSBC$^{\#}$).

From Figure 4c, it is observed that all ADFs & TFs except syllable nuclei level σ and κ are found to be optimal across all folds. This indicates that ADFs & TFs are data independent and have rich information for stress detection. It is also observed that syllable and syllable nuclei level mean and geometric-mean of SFs are not selected in any fold. This indicates that these features could be redundant in the presence of other features. Among three categories of features, it is observed that the ADFs are selected more consistently across folds followed by the TFs and the SFs. Similar to optimal sub-bands, instead of using fold specific features, we take union all optimal features across five folds for all experiments in this work. This, in turn, excludes four SFs resulting in a 16-dim feature vector. We compute UAs and WAs using this optimal set of features. From Figure 5, it is observed that both the UAs & WAs obtained using optimal features ((S-TCSSBC$^{\#}$)) are not significantly different from those using all features (S-TCSSBC) at all three levels – syllable, syllable nuclei and combined. This suggests that reduced set of features could be used for stress detection without any significant loss in the detection of syllable stress.

Table 1. UA obtained with baseline scheme and S-TCSSBC$^{\#}$ on GER and ITA test data. Bold entries indicate the highest UAs separately for GER and ITA.

	Baseline		S-TCSSBC$^{\#}$	
	WoPP	WPP	WoPP	WPP
GER	**85.57**	85.81	84.29	**87.53**
ITA	82.57	83.17	**83.73**	**86.26**

Table 1 shows the UA obtained on the baseline test setup for GER and ITA non-native speakers by the baseline technique [5] and the proposed sonority based features. From the table, it is observed that the UA obtained by the proposed S-TCSSBC$^{\#}$ is higher than that by the baseline technique for ITA under both WPP and WoPP and for GER under WPP. This indicates the effectiveness of the proposed sonority features for syllable stress detection. However, for GER under WoPP, the baseline technique has higher UA than the proposed method. This could be because baseline technique uses context rules in stress detection, which could be advantageous for stress detection in case of the German speakers [5]. However, these rules are applied using stress marking available in a dictionary, which, in general, may not be available for all non-native pronunciation variabilities. In this work, we assume availability of no such dictionary and avoid using any rule that may not be general for all non-nativity.

5. CONCLUSIONS

We propose a sonority based feature contour for automatic syllable stress detection task unlike a traditional short-time energy contour. The contour is computed by combining the sonority motivated cues with sub-band short-time energy contours reflecting prominence measures. We find an optimal set of sub-bands using a forward sub-band selection that maximizes the stress detection accuracy. We compute 20-dim feature vector from the sonority feature contour, out of which a sub-set of features are selected for stress detection using forward feature selection method. Experiments with ISLE corpus reveal that the proposed sonority based feature contour improves the syllable stress detection performance compared to the baseline technique. Further investigations are required to develop a better measure for sub-band selection that could result in an improved stress detection accuracy. Future works also include the use of the proposed features for the stress detection task in the native English speech as well as English from speakers with nativities other than German and Italian.

6. REFERENCES

[1] Abhishek Chandel, Abhinav Parate, Maymon Madathingal, Himanshu Pant, Nitendra Rajput, Shajith Ikbal, Om Deshmukh, and Ashish Verma, "Sensei: Spoken language assessment for call center agents," *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pp. 711–716, 2007.

[2] Junhong Zhao, Hua Yuan, Jia Liu, and S Xia, "Automatic lexical stress detection using acoustic features for computer assisted language learning," *Proceedings of Asia Pacific Signal and Information Processing Association (APSIPA) Annual Summit and Conference (ASC)*, pp. 247–251, 2011.

[3] Ashish Verma, Kunal Lal, Yuen Yee Lo, and Jayanta Basak, "Word independent model for syllable stress evaluation," *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 1, pp. 1237–1240, 2006.

[4] Luciana Ferrer, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication*, vol. 69, pp. 31–45, 2015.

[5] Joseph Tepperman and Shrikanth Narayanan, "Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners.," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 937–940, 2005.

[6] Alan Cruttenden, *Gimson's pronunciation of English*, Routledge, 2014.

[7] Andrew Rosenberg, Erica Cooper, Rivka Levitan, and Julia Hirschberg, "Cross-language prominence detection," in *Proceedings of Speech Prosody*. ISCA, 2012.

[8] Fabio Tamburini, "Prosodic prominence detection in speech," *Seventh International Symposium on Signal Processing and Its Applications*, vol. 1, pp. 385–388, 2003.

[9] Om D Deshmukh and Ashish Verma, "Nucleus-level clustering for word-independent syllable stress classification," *Speech Communication*, vol. 51, no. 12, pp. 1224–1233, 2009.

[10] Kun Li, Shuang Zhang, Mingxing Li, Wai Kit Lo, and Helen M Meng, "Prominence model for prosodic features in automatic lexical stress and pitch accent detection.," *Proceedings of Interspeech*, pp. 2009–2012, 2011.

[11] Mostafa Ali Shahin, Beena Ahmed, and Kirrie J Ballard, "Classification of lexical stress patterns using deep neural network architecture," *Spoken Language Technology Workshop (SLT), 2014*, pp. 478–482, 2014.

[12] Mostafa Shahin, Ricardo Gutierrez-Osuna, and Beena Ahmed, "Classification of bisyllabic lexical stress patterns in disordered speech using deep learning," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6480–6484, 2016.

[13] Supriya Nagesh, Chiranjeevi Yarra, Om D Deshmukh, and Prasanta Kumar Ghosh, "A robust speech rate estimation based on the activation profile from the selected acoustic unit dictionary," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5400–5404, 2016.

[14] Daniil A Kocharov, "Sonority measure for automatic speech recognition," *International Conference on Speech and Computer (SPECOM)*, pp. 359–362, 2006.

[15] Antonio Galves, Jesus Garcia, Denise Duarte, and Charlotte Galves, "Sonority as a basis for rhythmic class discrimination," *International Conference on Speech Prosody*, pp. 323–326, 2002.

[16] Robert Fuchs, *Speech Rhythm in Varieties of English*, Springer, 2016.

[17] Dagen Wang and Shrikanth Narayanan, "Speech rate estimation via temporal correlation and selected sub-band correlation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 413–416, 2005.

[18] JN Holmes, "The JSRU channel vocoder," *IEE Proceedings F (Communications, Radar and Signal Processing)*, vol. 127, no. 1, pp. 53–60, 1980.

[19] Chiranjeevi Yarra, Om D Deshmukh, and Prasanta Kumar Ghosh, "A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection," *Speech Communication*, vol. 78, pp. 62–71, 2016.

[20] Wolfgang Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, and Clive Souter, "The ISLE corpus of non-native spoken english," *Proceedings of Language Resources and Evaluation Conference (LREC)*, vol. 2, pp. 957–964, 2000.

[21] Bill Fisher, "tsylb2-1.1: syllabification software," *National Institute of Standards and Technology, Available online: https://www.nist.gov/itl/iad/mig/tools, last accessed on 07–09–16*, 1996.

[22] Abhay Prasad and Prasanta Kumar Ghosh, "Automatic classification of eating conditions from speech using acoustic feature selection and a set of hierarchical support vector machine classifiers," *Proceedings of Interspeech*, pp. 884–888, 2015.

[23] Alexander Gelbukh, *Computational Linguistics and Intelligent Text Processing*, Springer, 2011.

[24] Dagen Wang and Shrikanth S Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.