INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

ASR inspired syllable stress detection for pronunciation evaluation without using a supervised classifier and syllable level features

Manoj Kumar Ramanathi, Chiranjeevi Yarra, Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India

{manojkumar,chiranjeeviy,prasantg}@iisc.ac.in

Abstract

Automatic syllable stress detection is typically performed with a supervised classifier considering manually annotated stress markings and features computed within the syllable segments derived from phoneme transcriptions and their time-aligned boundaries. However, the manual annotation is tedious and the errors in estimating segmental information could degrade stress detection accuracy. In order to circumvent these, we propose to estimate stress markings in automatic speech recognition (ASR) framework involving finite-state-transducer (FST) without using annotated stress markings and segmental information. For this, we train the ASR system with native English data along with pronunciation lexicon containing canonical stress markings and decode non-native utterances as pronunciations embedded with stress markings. In the decoding, we use an FST encoded with the pronunciations derived using phoneme transcriptions and the instructions involved in a typical manual annotation. Experiments are conducted on polysyllabic words taken from ISLE corpus containing utterances spoken by Italian and German speakers and using the ASR models trained with three corpora. Among all the three models, the highest stress detection accuracies with the proposed approach respectively on Italian & German speakers are found to be 2.07% & 1.19% higher than and comparable to those with the two supervised classification approaches used as baselines.

Index Terms: Syllable stress detection, unsupervised approach, computer assisted language learning, ASR inspired modeling.

1. Introduction

Automatic syllable stress detection is an important component in the applications of computer assisted language learning (CALL) for learning second language (L2) [1, 2]. It is useful in evaluating L2 learners' pronunciation, identifying localized errors in their pronunciation and providing feedback to them during the learning [3, 4]. In most of the existing works, the automatic syllable stress detection is posed as a supervised classification problem and is performed in two steps [4, 5, 6, 7, 8, 9]. In the first step, features are computed heuristically for every syllable using phoneme transcriptions and its time-aligned boundaries from all utterances. In the second step, these features are used to classify a syllable as stressed or unstressed using a classifier trained using labeled data.

Tepperman et al. used prosodic features derived from fundamental frequency (f0), energy, duration to classify each syllable using Gaussian mixture models (GMMs) [4]. Deshmukh et al. used decision trees to classify similar prosodic features computed using nucleus level clustering [5]. Zhao et al. trained support vector machines (SVM) for classification using frame-averaged features and pitch-variation parameters computed using Rise/Fall/Connection model [6]. Li et al. trained multi-distribution deep belief networks (DBNs) with prosodic features for the classification [7]. Yarra et al. used SVM classifier considering the features based on the combination of sonority and energy [8]. However, most of these works train the classifier assuming availability of the stress labels on non-native English data that are obtained from manual annotation, which is often costly and cumbersome.

In order to avoid the manual annotation, Ferrer et al. [3] predicted the stress labels on non-native English data using a classifier trained with the features computed from native English data and respective canonical stress markings. However, these methods achieve lower performance than the approach proposed by Tepperman et al [4], which uses manual annotation. This could be because of the two step approach used in the stress detection task, in which, an error in estimating time-aligned boundaries propagates into the feature computation and affects the classifier and, hence, the stress detection performance. In order to circumvent this as well as the difficulties in the manual annotation, we propose to estimate the stress labels in an unsupervised manner without using the labels and the features.

In this work, unlike the previous approaches where each syllable was considered separately, we pose the stress detection task as sequence detection task in an automatic speech recognition (ASR) framework involving finite-state-transducer (FST). For this, we consider deep neural network-hidden Markov model (DNN-HMM) based ASR trained with the stressed and unstressed acoustic data separately using the native English data containing canonical stress markings. Further, for predicting stress label sequence in a non-native English word utterance, we perform decoding with the FST modified by embedding word specific multiple stress label sequences derived from the instructions used in the manual labelling of stress markings [10]. In order to know effectiveness of the proposed approach, we perform the experiments on polysyllabic words spoken by Italian (ITA) and German (GER) speakers taken from ISLE [10] corpus. We use three ASR models trained respectively with the entire training set of LibriSpeech corpus [11] (960 hours), a sub-set of training set of LibriSpeech (30 hours) and a subset of training set of Wall Street Journal corpus [12] (30 hours). We consider the supervised stress detection works proposed by Tepperman et al. [4] and Yarra et al. [8] as the first and the second baseline methods respectively. Among all the three ASR models, we achieve the highest stress detection accuracies of 85.24% and 87.00% on ITA and GER speakers, which are found to be 2.07% and 1.19% higher than and comparable to those from the first and the second baseline methods respectively.

2. Database

We use ISLE [10] corpus in all our experiments in this work. The corpus contains utterances from 46 non-native speakers (23 Italian (ITA) and 23 German (GER)) 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

http://dx.doi.org/10.21437/Interspeech.2019-2091

labeled all the syllable nuclei with stress markings by assuring only one stressed syllable, referred to as primary stress, in each word. In the experiments, we consider only polysyllabic words, of which, 4233, 1011 and 181 are bisyllabic, trisyllabic and quadrisyllabic words. These account to 5425 stressed (1) and 6798 unstressed (0) syllables respectively.

3. DNN-HMM based ASR system

A DNN-HMM based ASR system has three components – 1) acoustic model (AM), 2) pronunciation lexicon and 3) language model (LM) [13, 14]. The AM consists of an HMM and a DNN, where HMM and DNN represent the state transition probabilities for each context dependent phonemes and posterior probabilities of those states given speech acoustics respectively. The lexicon consists of multiple phoneme sequences representing pronunciations for each word. The LM consists of an n-gram model representing probability distribution of word sequences [15]. The parameters in the AM and LM are learnt independently, where for the former one, the lexicon is considered during the training. During decoding, the ASR uses an FST composed from the FSTs representing the three components. Typically, the FSTs for the AM, lexicon, LM and the composed FST are denoted as HC-FST, L-FST, G-FST and HCLG-FST [16]. It is to be noted that, when the decoding is performed for one word, the LM is not necessary, thus it is sufficient to consider HCL-FST.

In general, the syllable nuclei carry the prominence [17]. Hence, the AM could capture stress specific properties from the speech acoustics when it is trained using lexicon containing the phonemes encoded with stress prominence on the syllable nuclei. However, such training is not possible with a non-native English data since the stress markings are not readily available. On the other hand, it is possible with a native English data considering canonical stress labels on syllable nuclei. Hence, an HC-FST with native English data, referred to as native HC-FST, can be trained using such lexicon with stress encoded syllable nuclei (SESN).

4. Proposed approach

Figure 1 shows the four steps involved in the proposed approach and we illustrate these steps using an exemplary word " Tomorrow", which has the phonemes T, UW, M, AA, R, OW out of which the phonemes UW, AA & OW are the syllable nuclei of the syllables 'T UW', 'M AA R' & 'OW' [18]. For these syllable nuclei, the SESN with labels 1 and 0 are UW1, AA1 & OW1 and UW0, AA0 & OW0 respectively. The first step maps the phoneme sequence $\{p_1, p_2, \ldots, p_k\}$ in non-native word utterance containing N syllable nuclei into another phoneme sequence ensuring the mapped phonemes belong to the phoneme set used in constructing native AM. For the exemplary word, the k and N values are 6 and 3 respectively. The second step constructs an L-FST using N different phoneme sequences obtained from the modified phoneme sequence replacing the syllable nuclei in the modified sequence by either SESN with label 1 or 0. We refer these resultant phoneme sequences as SESN based phoneme sequences. For the exemplary word, there are 3 SESN based phoneme sequences obtained by replacing UW with either UW1 or UW0, AA with AA1 or AA0 & OW with OW1 or OW0. This is done by applying rules constructed from the typical instructions followed in manual labeling of stress markings. The third step performs composition of the constructed L-FST and HC-FST from the native AM to obtain HCL-FST. The fourth step decodes an SESN based phoneme sequence that has maximum likelihood among all the N sequences using DNN-HMM based ASR considering HCL-FST and the uttered speech signal of the respective word. At the end, the stress markings of the SESNs in the decoded SESN based phoneme sequence are declared as the estimated stress markings for the syllables containing those syllable nuclei. Let $\{T, UW1, M, AA0, R, OW0\}$ be the decoded SESN based phoneme sequence for the exemplary word, then the estimated stress markings are 1, 0, & 0 for the syllables 'T UW', 'M AA R' & 'OW' respectively.

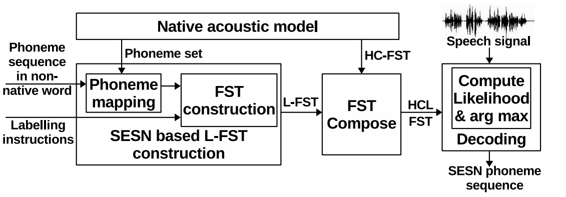


Figure 1: *Block diagram representing the steps involved in the proposed approach*

4.1. Background

In this work, we propose to select the best SESN based phoneme sequence based on the likelihood criteria. However, the likelihood of an acoustic observation sequence given AM would vary when there is a mismatch between the acoustic data considered in the observations and in the AM. Hence, it might affect the proposed criteria when acoustic observation sequence is from non-native speech but native speech is used to build AM. In order to investigate this, first, we describe the likelihood criteria and then analyze its effect on the stress detection task due to the mismatches between the non-native speech acoustics and the native AM.

Likelihood criteria: For a given phoneme sequence $\{p_1, p_2 ... p_k\}$, the likelihood corresponding to an acoustic observation sequence $\mathbf{O} = \{\mathbf{O}^{(i)}, 1 \leq i \leq k\}$, where $\mathbf{O}^{(i)}$ is an observation sequence of phoneme p_i, given native AM of those phonemes is defined as: $\mathcal{P}(\mathbf{O}) = \prod_{i=1}^{k} \mathcal{P}(\mathbf{O}^{(i)}|p_i)$. From the equation, it can be observed that the total likelihood of the observation sequence given the phoneme sequence is maximum when the likelihood of $\mathbf{O}^{(i)}$ given p_i is maximum for every phoneme. In order to compute the likelihoods for SESN based phoneme sequence, we propose to train the AMs for both the variants of SESNs i.e. SESNs with label 1 and 0. We indicate the AMs obtained for SESN with label 1 and 0 of the syllable nuclei phoneme p_i as (p_i, λ_i), where $\lambda_i \in \{1, 0\}$. It was studied that the likelihoods of acoustics within the segments of SESN with label 1 and 0 are higher with their respective AMs when both the acoustics and models belong to native speech compared to when they are not [19]. However, it is not clear how the likelihoods change when the acoustics are from non-native speech segments and AMs are from native speech. For this, we analyze the variations in the likelihoods of acoustic observation sequence belonging to syllable nuclei p_i given $(p_i, \lambda_i = 1)$ and $(p_i, \lambda_i = 0)$ separately.

Motivation for using native AM: Instead of the likelihood, we analyze variations using normalized likelihood, which is equal to the posterior probability $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i)$ shown in Equation 1 under equal likely prior on λ_i, with the help of Figure 2(a).

$$\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i) = \frac{\mathcal{P}(\mathbf{O}^{(i)}|p_i, \lambda_i)\ \mathcal{P}(\lambda_i)}{\sum\limits_{\lambda_i \in \{1,0\}} \mathcal{P}(\mathbf{O}^{(i)}|p_i, \lambda_i)\ \mathcal{P}(\lambda_i)} \quad (1)$$

Figure 2(a) shows the distribution of $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i)$ for all

p_i belonging to unstressed syllable nuclei computed using native AMs from LibriSpeech corpus [11] for $\lambda_i = 1$ and 0, where $\mathbf{O}^{(i)}$ is the acoustic observation sequence within syllable nuclei segments from ISLE corpus whose ground-truth stress label is 0. From the figure, it is observed that $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i)$ is higher when $\lambda_i = 0$. It is also observed that the percentage of number of syllable nuclei having $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i) < 0.5$ for $\lambda_i = 0$ is 17.20%. These together indicates that the $\mathbf{O}^{(i)}$ from non-native English data gives higher $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i)$ with native AMs when the stress labels on $O^{(i)}$ and p_i are matched than that when those are not. We also observe higher $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i)$ in the matched scenario when the acoustic observations are collected from the segments with ground-truth stress label as 1. These together suggests that the likelihood of the acoustic observation sequence $\mathbf{O}$ is maximum when the considered phoneme sequence contains syllable nuclei encoded with the ground-truth stress markings even for the non-native speech. However, these stress markings are unknown and those need to be estimated.

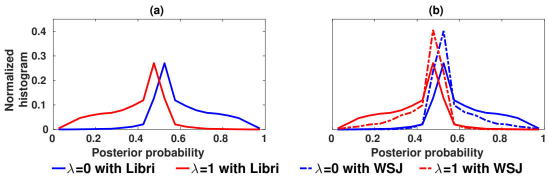
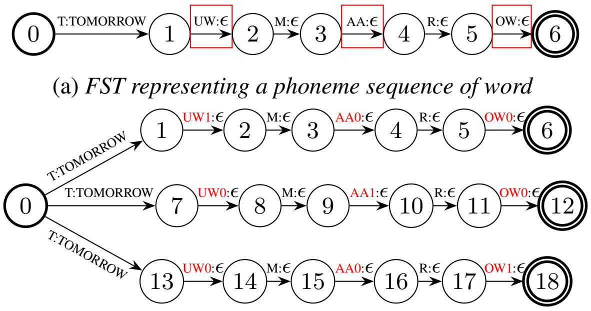


Figure 2: *Normalized histogram of the posterior probability for λ=0 and λ=1 when unstressed syllable nuclei acoustic segments are considered from non-native speech and AMs considered are trained on LibriSpeech (Libri) and Wall Street Journal (WSJ) corpora*



(a) *FST representing a phoneme sequence of word*

(b) *FST representing SESN based phoneme sequences of word*

Figure 3: *SESN based L-FST construction from a phoneme sequence. ϵ represents the null output label*

4.2. Rules for FST construction

In order to derive the SESN based phoneme sequence for the given phoneme sequence in a non-native utterance, we propose to modify syllable nuclei phonemes in the phoneme sequences by replacing with their respective SESN with label 1 or 0. Thus, after the replacement, the length of SESN based phoneme sequence is identical to that of the given phoneme sequence. In the manual labeling, typically, it was followed that there is only one stressed syllable nucleus in a word and remaining syllable nuclei are unstressed [10]. Thus, in deriving SESN based phoneme sequences we propose to replace one syllable nucleus at a time with the respective SESN with label 1 and the remaining with their respective SESN with label 0. Considering this constraint, it can be observed that a total of N SESN based phoneme sequences are obtained for an N syllable nuclei word from the given phoneme sequence. Also, it is easy to observe that the set of SESN based phoneme sequences also include the phoneme sequence encoded with ground truth stress labels, referred as ground truth phoneme sequence. Thus, we hypothesize that the SESN based phoneme sequence that matches with the ground truth phoneme sequence results in the maximum likelihood among all N SESN based phoneme sequences in the set. The maximum likelihood SESN based phoneme sequence can be decoded using DNN-HMM based ASR with HCL-FST composed from HC-FST and L-FST encoded with all of the N SESN based phoneme sequences.

4.3. SESN based L-FST construction

4.3.1. Phoneme mapping

In order to compute likelihood of an SESN based phoneme sequence, all the phonemes in the phoneme sequence should be present in the phoneme set considered for learning native AM. In order to ensure this, we replace phonemes that are present in the non-native English data but not present in the native AM by a phoneme in the phoneme set that is used in constructing the native AM [20] [21]

4.3.2. FST construction

Figure 3a shows the FST used to compute the likelihood in a typical DNN-HMM based system for the phoneme sequence of the word “Tomorrow”. It is to be noted that there is only one path in the FST with seven nodes and six transitions. In the FST, each transition has one input and one output symbol, where the input symbol on every transition is represented by each phoneme in the phoneme sequence according to their sequence of occurrence. Thus, the number of transitions in the FST are equal to the number of phonemes in the word. There are six phonemes in word “Tomorrow” and, hence, six transitions in FST in Figure 3a. Further, as this word has three syllables, there would be three SESN based phoneme sequences of length six.

In order to incorporate the SESN based phoneme sequences, we modify the FST by increasing the number of paths to N without altering the number of transitions as in the FST of Figure 3a by keeping a constraint of ‘no node should be common across these paths except the begin node’. In addition, in each path, we change the input symbol at the transitions corresponding to the syllable nuclei and consider the remaining input and output symbols identical to those in the FST in Figure 3a. Figure 3b shows the proposed SESN based L-FST, where the number of parallel paths is equal to three and the input and output symbols are identical to those in Figure 3a except the input symbols marked in red. The red marked input symbols at the marked locations are obtained considering the input symbols belonging to the respective transitions in the Figure 3a as follows: (1) Select one transition belonging to one of the syllable nuclei in each path ensuring that the transition is not selected in any other paths. Thus, all transitions corresponding to N syllable nuclei are selected with one transition selected in each of the N parallel paths. In Figure 3a, the red rectangular boxes indicate three such transitions. (2) Replace the syllable nuclei in the selected transitions with the respective SESN with label 1. In Figure 3b, the SESN with label 1 $\{UW1,\ AA1,\ OW1\}$ in the first, second and third paths are obtained by replacing the syllable nuclei $\{UW,\ AA,\ OW\}$ in the selected first, second and third transitions respectively from Figure 3a. (3) Replace the remaining syllable nuclei that are not considered in the step 2 in each path with the respective SESN with label 0. In Figure

3b, such SESN with label 0 in the first, second and third paths are $\{AA0,\ OW0\}$, $\{UW0,\ OW0\}$ and $\{UW0,\ AA0\}$ respectively.

5. Experiments and results

Experimental setup: We consider unweighted accuracy [4, 8] as the objective measure for evaluating the proposed approach. We consider the works by Tepperman et al. [4] and Yarra et al. [8] as the baseline techniques and refer them as BL-1 and BL-2 respectively. We consider a test set of polysyllabic words identical to those used in the work by Yarra et al. [8] for both ITA and GER speakers. For these polysyllabic words, we obtain the phoneme transcriptions available in the ISLE corpus. We learn three native AMs separately using three data sets – 1) 960 hours of LibriSpeech (Libri) [11], 2) 30 hours of LibriSpeech (Libri-S) and 3) 30 hours of Wall Street Journal (WSJ) [12]. For training with these datasets, we use MFCC as features and we consider the phoneme set described in Section 4.3.1 by replacing syllable nuclei with SESN with label 1 and 0. The 30 hours of data in Libri-S is selected randomly from the Libri data to match with the data size of WSJ in order to study the effect of the data size on the stress detection performance. We use Kaldi speech recognition toolkit [13] to construct HC-FST from each native AM, to compose HCL-FST from HC-FST & L-FST and to decode the SESN based phoneme sequence.

Table 1: *Accuracies obtained for ITA and GER with the two baselines and the proposed approach using all three native AMs*

	BL-1	BL-2	Libri	Libri-S	WSJ
ITA	83.17	86.26	**85.24**	75.39	72.05
GER	85.81	87.53	**87.00**	79.32	75.53

Results and Discussion: Table 1 shows the stress detection accuracies obtained with the two baselines and the proposed approach using all three native AMs for ITA and GER speakers. In the Table, the bold entries indicate the highest accuracies achieved with the proposed approach among all the three native AMs for both sets of speakers. From the table, it is observed that the highest accuracy with the proposed approach is higher than that using the BL-1 separately for both the ITA and GER speakers. The absolute improvements over BL-1 are found to be 2.07% and 1.19% respectively on ITA and GER speakers. This indicates that the proposed approach achieves better performance without using the stress labels compared to the BL-1 that takes the advantage of the labels in a supervised manner. However, it is also observed that the highest accuracy obtained using the proposed approach are 1.02% and 0.53% (absolute) lower than the BL-2 for ITA and GER speakers respectively. This indicates that though the proposed approach does not use any labels, the accuracies are comparable with those of BL-2. These together suggest the effectiveness of the proposed approach for the stress detection task in an unsupervised manner.

From Table 1, it is also observed that the stress detection accuracy for both ITA and GER speakers using the proposed approach is higher with the native AM trained with Libri data set than that with Libri-S and WSJ data sets. It is interesting to observe that the amount of data considered in Libri-S and WSJ is lower than that considered in Libri. These together indicate that the stress detection accuracy with the proposed approach depends on the amount of data used in training the native AM. Further, we investigate this with the help of Figure 2(b) considering the native AMs trained with Libri and WSJ, similar to the illustration in Figure 2(a). Figure 2(b) shows the distribution of $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i)$ computed using native AM trained on Libri and WSJ data for $\lambda_i = 0$ & 1 for all the acoustic observations collected from syllable nuclei segments whose ground-truth stress labels are 0. From the figure, it is observed that the $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i)$ is higher for $\lambda_i = 0$ when native AM is from Libri data than that from WSJ data. Further, the percentage of number of syllable nuclei having $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i) < 0.5$ is found to be 26.59% with WSJ data, which is 9.39% higher than that with Libri data. A similar trend is observed with Libri-S as well, where we found $\mathcal{P}(\lambda_i|\mathbf{O}^{(i)}, p_i) < 0.5$ happens for 29.55% syllable nulcei. These together suggest that the drop in the accuracies is due to lesser amount of the training data for the native AM required for the stress detection task.

Table 2: *Accuracies obtained for ITA and GER with the BL-2 and the proposed approach using all three native AMs for bisyllabic (B), trisyllalbic (T) and quadrisyllabic (Q) word*

	ITA			GER		
	B	T	Q	B	T	Q
BL-2	88.85	86.97	77.21	89.13	84.31	73.58
Libri	86.92	83.37	76.83	87.02	**89.27**	74.24
Libri-S	76.13	72.13	**79.88**	78.22	81.43	**83.33**
WSJ	71.70	72.90	72.56	75.04	76.93	75.25

Further, we analyze the stress detection performance using the proposed approach separately for the bisyllabic (B), trisyllabic (T) and quadrisyllabic (Q) words with respect to the best baseline (BL-2). Table 2 shows the stress detection accuracies for B, T and Q words separately for ITA and GER speakers with the proposed approach considering all the three native AMs and the BL-2. In the table, the bold entries indicate the accuracies where the proposed approach has a higher accuracy compared to that with BL-2. Although, the overall accuracy with BL-2 is higher than the proposed approach for both ITA and GER speakers, it is not so for all B,T and Q categories. For example, with Libri based native AM, the proposed approach yields higher accuracy in T and Q categories on GER speakers compared to BL-2. Similarly, with Libri-S based native AM, the proposed method does better than BL-2 in Q category in the case of ITA speakers. This suggests the benefit of the proposed approach compared to the BL-2.

6. Conclusion

Unlike using a supervised classifier and syllable level features in the stress detection task, we decode the stress markings on syllables in non-native word utterances using an ASR based framework involving FST. The FST is composed using an FST representing native AM and another FST constructed from a set of phoneme sequences containing SESN with labels 1 and 0. Three different native AMs are trained for the SESN with label 1 and 0 separately using native English data and a lexicon containing canonical stress markings. Experiments with ISLE corpus reveal that the proposed unsupervised approach performs comparably with the two supervised classification approaches. Further investigations are required to improve the stress detection task when there is less data to train the native AM. Future work also includes study of the stress detection performance when the phoneme sequence from the non-native utterance is not available. In the future, we also plan to investigate on first language specific tendencies for misplacing syllable stress.

7. Acknowledgement

We thank the Department of Science Technology, Government of India and the Pratiksha Trust for their support.

8. References

[1] A. Chandel, A. Parate, M. Madathingal, H. Pant, N. Rajput, S. Ikbal, O. Deshmukh, and A. Verma, "Sensei: Spoken language assessment for call center agents," *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pp. 711–716, 2007.

[2] A. Verma, K. Lal, Y. Y. Lo, and J. Basak, "Word independent model for syllable stress evaluation," *International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 1, pp. 1237–1240, 2006.

[3] L. Ferrer, H. Bratt, C. Richey, H. Franco, V. Abrash, and K. Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication*, vol. 69, pp. 31–45, 2015.

[4] J. Tepperman and S. 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.

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

[6] J. Zhao, H. Yuan, J. 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.

[7] K. Li, X. Qian, S. Kang, and H. Meng, "Lexical stress detection for L2 English speech using deep belief networks," *Proceedings of Interspeech*, pp. 1811–1815, 2013.

[8] C. Yarra, O. D. Deshmukh, and P. K. Ghosh, "Automatic detection of syllalbe stress using sonority based prominence features for pronunciation evaluation," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 5845–5849, 2017.

[9] J. Y. Chen and L. Wang, "Automatic lexical stress detection for Chinese learner's of English," *7th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 407–411, 2010.

[10] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," *Proceedings of Language Resources and Evaluation Conference (LREC)*, vol. 2, pp. 957–964, 2000.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.

[12] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," *Proceedings of the workshop on Speech and Natural Language*, pp. 357–362, 1992.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, and P. Schwarz, "The Kaldi speech recognition toolkit," *IEEE workshop on automatic speech recognition and understanding (ASRU)*, 2011.

[14] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, pp. 82–97, 2012.

[15] P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[16] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech Language*, vol. 16, pp. 69 – 88, 2002.

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

[18] B. 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.

[19] S. de Lemos, "What automatic speech recognition can tell us about stress and stress shift in continuous speech," *Proceedings on Speech Prosody*, pp. 984–988, 2018.

[20] R. Weide, *The CMU pronunciation dictionary*. Carnegie Mellon University, 1995.

[21] B. BabaAli, "Phones.60-48-39.map," *Available at https://github.com/kaldi-asr/kaldi/blob/master/egs/timit/s5/conf/phones.60-48-39.map, last accessed on 20-02-2019*, 2013.