SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education
20-21 September 2019, Graz, Austria

Comparison of automatic syllable stress detection quality with time-aligned boundaries and context dependencies

Chiranjeevi Yarra, Manoj Kumar Ramanathi, Prasanta Kumar Ghosh
Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India
{chiranjeeviy,manojkumar,prasantg}@iisc.ac.in

Abstract

Syllable stress is detected automatically using a classifier trained with stress labels and features computed based on acoustics within syllables. Typically, in real scenarios, syllable data is estimated considering an acoustic model (AM) and a lexicon. Thus, their quality affects the stress detection performance (accuracy). In this work, we analyse variations in the accuracies on ISLE corpus containing spoken English utterances from non-native speakers. In the analysis, we consider five AMs and five lexicons containing native English pronunciations augmented with different percentages of non-native pronunciations collected from the corpus. For each AM and lexicon combination, we estimate syllable data using two existing forced-alignment techniques and observe that the accuracies obtained with the features from both the data are comparable. Further, we propose a set of features based on context dependencies of the syllable nuclei. For all the combinations, the accuracies are higher when context based features are augmented with acoustic based features and the highest accuracy is obtained for the combination whose estimated syllable data has the least error. Among all five lexicons, the highest and the least accuracies for ITA & GER are obtained when the lexicons include all & none and none & all of the non-native pronunciations respectively.
Index Terms: Stress detection quality, context dependent features, estimated syllable data.

1. Introduction

Automatic detection of syllable stress has been shown to be useful for evaluating pronunciation [1–3] in several applications including computer assisted language learning (CALL). It is also useful in providing feedback to the second language (L2) learners by automatically identifying localized pronunciation errors [4, 5]. Typically, the stress detection task is performed as a classification problem in a supervised manner [3, 5–7] using a set of features representing a syllable and the respective stress labels (stressed and unstressed). In most of the existing works, the labels are obtained from a manual annotation process and the syllable data (both the syllable transcriptions and their time-aligned boundaries) is estimated using forced-alignment [4, 7–10]. Hence, the reliability of the stress detection task depends on the features and the quality of the syllable data.

Forced-alignment is performed with an automatic speech recognition (ASR) system considering an acoustic model (AM) and a pronunciation lexicon [5, 6, 8]. Thus, different combinations of AMs and lexicons could cause variations in the syllable data in terms of the forced-aligned syllable transcriptions as well as their boundaries. Usually, a pronunciation lexicon containing all pronunciation variants of L2 learners could result in a better quality syllable data from the forced-alignment. However, the availability of such lexicons is limited and identification of such pronunciations is also challenging.

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

Significance of the study: Since the quality of the syllable data depends on the AM and the lexicon used in the forced-alignment, it is required to analyse the effect of the syllable data estimated under different combinations of AMs and lexicons in the stress detection task. Further, in a typical forced-alignment process, a single AM is considered to represent a syllable nucleus in both the stressed and unstressed syllables. However, the speech acoustics belonging to a nucleus of the stressed syllable differ from those of an unstressed syllable. Thus, in order to account these, Ramanathi et al. [11] have proposed to perform forced-alignment on non-native English data considering syllable nucleus in stressed and unstressed syllables separately. For this, they have encoded respective stress labels on to a syllable nucleus and referred it as stress encoded syllable nucleus (SESN). Further, in the process, they have used a separate AM for each SESN to represent speech acoustics in it.

In addition, most of the existing works have proposed features to capture variations in the acoustics within syllables. Some works have considered the variations in f0, energy and syllable duration [5, 6, 8, 9]. Some other works have also used features based on spectral tilt and log posteriors from Gaussian mixture models using Mel frequency cepstral coefficients (MFCCs) [4] and perceptual attributes [10]. In addition to the acoustic features, some works have reported using contextual information in the syllables for the stress detection task. Tepperman et al. have proposed a set of acoustic features by incorporating nuclei and its context dependent variations [5] computed from heuristic observations [12, 13]. Similarly, Deshmukh et al. have performed the stress detection using acoustic features in multiple clusters depending on broad phoneme class of the nuclei [7]. However, the contextual information considered in these works are limited and heuristically applied. Thus, it is required to analyse the context dependencies further in a data-driven manner for the stress detection task.

Summary: For the study, we consider a supervised stress detection setup using an SVM classifier trained with a set of 19-dim binary features representing context dependencies added with existing acoustic based features [14]. We perform experiments on ISLE [15] corpus containing polysyllabic English words separately spoken by German and Italian speakers. We estimate the syllable data from a typical and the SESN based forced-alignment processes using five different AMs trained with the data from three corpora, namely, Librispeech (LS) [16], wall-street journal (WSJ) [17] and fisher-English (FE) [18] and six lexicons by augmenting randomly chosen word pronunciations from ISLE corpus at different percentages to a native English pronunciation lexicon. For all the AM and lexicon combinations, improvements in the accuracies are found when context based features are included compared to when acoustic features alone are used and the highest improvement is obtained for the lexicon that includes all word pronunciations from ISLE. For every combination of AM and lexicon, the accuracies obtained using the data from the two force-alignment processes are comparable.

10.21437/SLaTE.2019-15

2. Database

We use ISLE [15] 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. Phoneme transcriptions were available, which were obtained from forced-alignment followed by manual correction from a team of five linguists. They also labeled all the syllable nuclei with stress markings by assuring only one stressed syllable nucleus in each word. Using the phoneme transcriptions, we obtain time-aligned boundaries considering forced-alignment approach. Following this, we obtain the syllable data, referred to as ground-truth syllable data. In the experiments, we consider only polysyllabic words from the data. This result in a total of 7586 & 7791 and 8586 & 4648 words in the train and test data respectively for GER & ITA speakers.

3. Proposed setup for the study

The block diagram in Figure 1 shows the proposed stress detection setup for the study. It has four steps, out of which we analyse the effect of first and second steps in the stress detection task. In the first step, we apply the forced-alignment process on a speech signal using its sentence transcription to estimate phoneme transcriptions as well as aligned phoneme and word boundaries. Further, we syllabify the phoneme transcriptions and obtain syllable transcriptions and its time aligned boundaries. In the second step, we compute a binary feature vector for each syllable using its transcription. In the third step, we classify each syllable segment as stressed or unstressed with an SVM classifier using binary plus acoustic features. In the last step, we post-process the estimated stress markings to ensure that each polysyllabic word has only one stressed syllable.

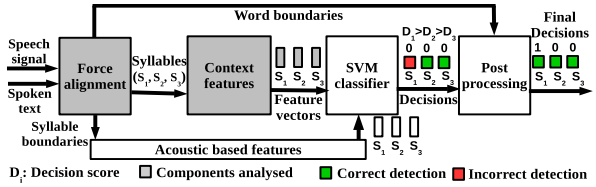

Figure 1: *Block diagram represents the steps involved in the proposed stress detection setup for the study. The gray color blocks indicate the components that are analysed in this study.*

3.1. Context dependent features

The acoustic based features that are derived from the syllable or the syllable nucleus are not only affected by stress prominence but also by context and phoneme class of the syllable nucleus. Umeda has observed that the duration of the vowel (typically, forms the syllable nuclei) changes depending on the phonemes on its either sides [13]. In order to account these variabilities, there exist works that use contextual information for the stress detection task [5, 7]. However, these are either heuristically applied or limited in using the context. In this study, we consider a 19-dim binary feature vector derived from the phoneme class of the nuclei and its context dependencies reported in Table 1 and perform the stress detection task in a data-driven manner. In the table, there are a total of 19 categories and for each category, a binary value (1 or 0) is encoded in the feature vector indicating the context category.

3.2. Syllable data estimation

We obtain two sets of syllable data using a typical and SESN based forced-alignment processes.

Table 1: *List of the context and corresponding categories considered for obtaining a 19-dim binary feature vector*

Context	Category
Syllable nucleus type	Low, Mid, High, Lax and Tense
Phoneme preceding the syllable nucleus	Nasals, Voiced stops, Unvoiced stops, Voiced fricative, Unvoiced fricatives and remaining consonants
Phoneme following the syllable nucleus	Nasals, Voiced stops, Unvoiced stops, Voiced fricative, Unvoiced fricatives and remaining consonants
Word Position	Pre-pausal, Not pre-pausal

3.2.1. Typical forced-alignment

A state-of-the-art ASR system has three components – 1) AM, 2) pronunciation lexicon and 3) language model (LM) [19, 20]. The AM consists of a hidden Markov model (HMM) and a deep neural network (DNN), where each context dependent phoneme is modeled using states of an HMM and the DNN is used to compute the posterior probabilities of those states given speech acoustics. 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 [21]. The parameters in the AM and LM are learnt independently, where for the former, the lexicon is considered during the training. During decoding, all the three models are used. However, in the forced-alignment sentence transcriptions are known and, hence, the LM is not required. Thus, the forced-alignment is performed in a manner similar to decoding without LM.

3.2.2. SESN based forced-alignment

We perform SESN based forced-alignment by following the work proposed by Ramanathi et al. [11]. In this process, a syllable nucleus in the stressed and unstressed syllables is represented separately using SESN with label 1 and 0 respectively. As an example, for the syllable nucleus ɑ, the SESN with label 1 and 0 are $ɑ^1$ and $ɑ^0$ respectively. Further, the AMs are learnt separately for each SESN of every syllable nucleus and a lexicon is obtained by incorporating the pronunciation variants based on the SESN. We describe the process of obtaining updated lexicon with SESN using an exemplary word "Tomorrow". For this word, one of the pronunciation entries contains the phonemes t, u, m, ɑ, r, oʊ out of which the phonemes u, ɑ & oʊ are the syllable nuclei. For these syllable nuclei, the SESN with label 1 are u^1, $ɑ^1$ & $oʊ^1$ and with label 0 are u^0, $ɑ^0$ & $oʊ^0$. Considering these, a new pronunciation entry is obtained by replacing one syllable nucleus at a time in the existing entry with the respective SESN with label 1 and the remaining syllable nuclei with respective SESN with label 0. Thus, for each pronunciation entry a set of k entries is obtained, where k is the number of syllable nuclei in the pronunciation. For the exemplary pronunciation, three SESN based entries are obtained and those are {t, u^1, m, $ɑ^0$, r, $oʊ^0$}, {t, u^0, m, $ɑ^1$, r, $oʊ^0$} and {t, u^0, m, $ɑ^0$, r, $oʊ^1$}. This procedure is applied to all entries in the pronunciation lexicon.

3.3. Study summary

The stress detection accuracy is a function of acoustic and context based features, which are, in turn, a function of the syllable data. Further, the syllable data depends on forced-alignment process including the AM & lexicon used. Thus, it is non-trivial

to analyse each one individually. Besides that, in order to perform the stress detection task, stress labels are needed for estimated syllable data, which can be obtained from the manual annotation. In practice, it is costly and cumbersome to obtain labels for the estimated syllable data. Hence, it is required to analyse variations under mismatched train-test condition i.e., considering the test set from estimated syllable data and train set from the data assigned with ground-truth labels, for example, ground-truth syllable data. In order to address these, we propose to study their effects considering both acoustic and context based features with the following set-ups.

- To analyse the effect of AMs only, we consider the ground-truth syllable data obtained using different AMs.
- To analyse the mismatched condition, we consider estimated syllable data obtained using different AMs with one lexicon.
- To analyse the effect of the lexicon, we consider estimated syllable data obtained using the same AMs considered in set-up 2 and synthetically generated lexicons containing different percentages of non-native pronunciations.

4. Experiments and results

4.1. Experimental setup

We consider unweighted accuracy [5, 14] as an objective measure in the stress detection task. We use an 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 [22]. Following the work by Yarra et al. [14], we obtain the acoustic features for each syllable. We consider P2TK syllabifier [23] for the syllabification.

Setup for forced-alignment: We use DNN-HMM based AMs for the forced-alignment. The DNN-HMM models are learnt by following Daniel Povey's (Dan's [24]) implementation [25] available in the Kaldi speech recognition tool-kit [19] using five data sets – 1) 960 hrs of entire set from Libri-speech [16] (LS), 2) 30 hrs of sub-set randomly chosen from LS (LS-S), 3) 30 hrs of sub-set randomly chosen from Wall street journal [17] (WSJ), 4) 2000 hrs of entire set from Fisher English [18] (FE) and 5) 30 hrs of sub-set randomly chosen from FE (FE-S). The smaller sub-sets are considered to analyse the effect of data size used in learning AMs. Further, the lexicons are obtained by augmenting a native English lexicon with five different sub-sets randomly chosen from the entire set of word pronunciations by GER and ITA speakers collected from ISLE at the following percentages – 0, 25, 50, 75 and 100. The native English lexicon is obtained by combining the following four lexicons – CMU [26], TIMIT [27], Beep [28] and the lexicon used in preparing ISLE data. The phonemes in the combined lexicon are mapped to a set of 39 phonemes [26] following the phoneme mapping[1] available in the Kaldi tool-kit. It is to be noted that we have not learnt a DNN-HMM AM using ISLE corpus due to its limited amount of data.

Train and test sets: Similar to the work proposed by Yarra et al. [14] and Tepperman et al. [5], we use 1st-12th & 1st-13th speakers data for training and 13th-23rd & 14th-23rd speakers data for testing for GER & ITA respectively for the experiments based on the ground-truth syllable data. These sets are referred to as original train and test sets. However, under mismatched train-test condition, we use the original train set for the training but, for testing, we derive a sub-set from the above test set for each combination of AM and lexicon. This is because, in order to compute the accuracies, labels for the estimated syllable data are needed. Since those are not available, we propose to obtain those by assigning ground-truth labels to the estimated syllable data. This can be done only when the number of the syllables in a word in the ground-truth syllable data is identical to that in the estimated syllable data, although it is not guaranteed in the entire test set.

[1] https://github.com/kaldi-asr/kaldi/blob/master/egs/timit/s5/conf/phones.60-48-39.map

4.2. Results and discussion

4.2.1. Comparison based on AM

In order to know the effectiveness of AM, first, we analyse ASR performance using each AM and then study the stress detection accuracies. For the ASR, in order to avoid any variations based on lexicon entries, we consider a lexicon with only word pronunciations by GER and ITA speakers collected from the ISLE. The WERs obtained on the original test sets from GER & ITA speakers using LS, LS-S, WSJ, FE and FE-S based AMs are found to be 16.02 & 20.84, 27.41 & 36.02, 35.25 & 55.11, 24.75 & 31.73 and 40.18 & 50.73 respectively. From these values, it is observed that the WERs are significantly high in both GER and ITA speakers across all five AMs, although the lexicon contains all non-native pronunciations. This could be because of mismatches between the speech acoustics in the data used for AMs and those in non-native data. Further, between GER and ITA speakers, the higher WERs are observed in the case of ITA speakers. This indicates that the degree of such mismatches is more in the case of ITA speakers than those in the case of GER speakers. It is interesting to observe that, between the AMs from LS and LS-S, the higher WER is found when AM from LS-S is used. Similarly, a higher WER is found when the AM from FE-S is used compared to the AMs from FE. These together indicate that the AMs are less effective in compensating acoustic mismatches when the AMs are learnt from a smaller set of native English data compared to those from a larger set.

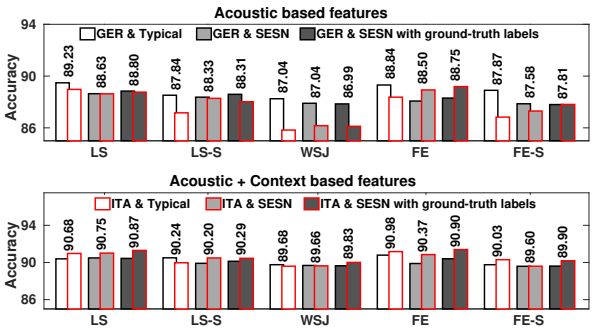

Figure 2: *Stress detection accuracies obtained with ground-truth syllable data using five types of AMs in three types of forced-alignment approaches – 1) typical 2) SESN and 3) SESN with ground-truth labels.*

Figure 2 shows the stress detection accuracies obtained using acoustic and acoustic plus context based features for GER and ITA using all five AMs. For this, in addition to the ground-truth syllable data from two types of forced-alignment approaches, we obtain another set of ground-truth syllable data with SESN based approach considering ground-truth phoneme transcriptions encoded with ground truth labels. From the figure, it is observed that the accuracies obtained using acoustic plus context based features are more than those using only

acoustic features for both GER and ITA across all three types of syllable data in every AM. This indicates the benefit of the proposed context dependent features. Further, the accuracies obtained using acoustic based features in the case of ITA speakers are lower than those in the case of GER for all three types of syllable data. Similarly, the accuracies are lower with AMs using LS-S, FE-S than those using LS and FE respectively and it is the least when WSJ based AM is used. These together indicate that the AM with better ASR performance on non-native data results in better stress detection accuracy. While observing the accuracies obtained using acoustic plus context based features, similar comparisons based on acoustic based features are observed only with respect to the AMs but not across GER and ITA. The accuracies obtained for ITA are higher than those for GER in most of the cases. This indicates that the reduction in the accuracies due to the lower ASR performance is compensated by using accurate context dependent features for ITA.

In addition, when the accuracies are averaged across the GER and ITA (shown above the each pair of bar), it is observed that the accuracies are similar across all three types of forced-alignment approaches for a given AM using acoustic and acoustic plus context based features separately. This indicates that the effect due to different types of forced-alignment approaches are negligible.

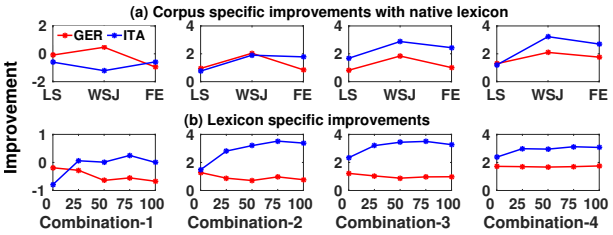

Figure 3: *Improvements in the accuracies under estimated syllable data conditions with respect to those under ground-truth syllable data.*

4.2.2. Comparison based on mismatched train-test condition

Figure 3a shows the improvements in the accuracies obtained with four sets of acoustic plus context based features using estimated syllable data from those obtained with acoustic based features using ground-truth syllable data. The four sets of features are obtained when the syllable boundaries & its transcriptions used in the respective acoustic & context based features are obtained from the following four combinations – 1) estimated data & none 2) estimated data for the both 3) estimated & ground-truth syllable data 4) ground-truth & estimated syllable data. For this, we consider the estimated syllable data obtained with the typical forced-alignment approach using the AMs from LS, WSJ and FE considering the native English lexicon. This because the accuracies are comparable across all three types of forced-alignment approaches, hence we use the typical force-alignment approach for the rest of the experiments. Further, AMs using LS-S and FE-S are not considered as they perform worse than LS and FE based AMs. From the figure, it is observed that the improvements are negative in most of the AMs only for the first combination. This could be due to the variations in the syllable boundaries in the estimated syllable data from the ground-truth syllable data. Further, the positive values of improvements in all the three AMs for the second combination indicate the benefit of the acoustic plus context based features even when the syllable data is estimated. In addition, the higher improvements with the fourth combination compared to those with the third combination suggest that the estimation of more accurate boundaries is important than the estimation of correct syllable transcriptions. It is also observed that the improvements in the case of ITA are lower than those in the case of GER. This could be because we observe that the variations in the boundaries and transcriptions are higher in the case of ITA speakers than in the case of GER speakers.

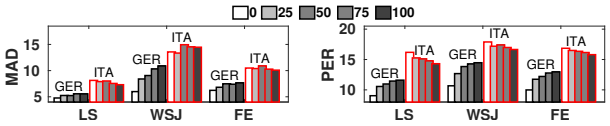

Figure 4: *MAD and PER obtained between estimated and ground-truth syllable data*

4.2.3. Comparison based on lexicon

Figure 3b shows the improvements in the accuracies as computed in Figure 3a averaging across those from estimated data using all three AMs. In this figure, the averaged improvements are computed for the estimated data from five augmented lexicons. From the figure, it is observed that the averaged improvements increase (decrease) with the percentages of augmentation for ITA (GER) respectively. This indicates that adding non-native pronunciation helps in the case of ITA but not the case for GER. In order to investigate this, we compute the sum of mean absolute difference (MAD) between the boundaries and phoneme error rate (PER) between the transcriptions in the estimated and ground-truth syllable data. Figure 4 shows MAD and PER for GER and ITA using estimated syllable data from the three AMs and all five augmented lexicons. From the figure, it is observed that the MAD and PER increase with the percentage of augmented non-native pronunciations for GER speakers. This could be because the pronunciations of GER speakers are similar to the native English pronunciations. Hence, a bigger lexicon with more non-native pronunciations could result in lower performance due to the confusions caused by more number of pronunciations per word. On the other hand, the MAD and PER decrease with the percentages for ITA speakers. In this case, a bigger lexicon supports in estimating non-native pronunciations which are not available in the native English lexicon. These together suggest that the variations in the averaged improvements as observed in Figure 3b are inversely proportional to the variations in the MAD and PER with the augmentation percentage. Thus increasing the size of lexicon could not improve the accuracies for all types of non-native speakers.

5. Conclusions

We analyze how the accuracy of a supervised technique for stress detection task is affected by the quality of the estimated syllable data, which is used for computing the acoustic features along with the proposed context features for stress detection. We conduct experiments on ISLE corpus considering estimated data from two types of forced-alignment approaches using five AMs and five lexicons, where the lexicons are generated by adding five different percentages of non-native pronunciations of GER and ITA. We observe that the accuracies from both the forced-alignment approaches are comparable and those are higher when the acoustic features are augmented with the proposed context features. Among all the combinations, the highest and the least stress detection accuracies are obtained, when the lexicons contain 100% and 0% of non-native pronunciations respectively for ITA and vice-versa for GER. Further investigations are required to develop better strategies to obtain accurately aligned boundaries.

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

[7] 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.

[8] 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.

[9] 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.

[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] Manoj Kumar Ramanathi, Chiranjeevi Yarra, and Prasanta Kumar Ghosh, "ASR inspired syllable stress detection for pronunciation evaluation without using a supervised classifier and syllable level features," *Submitted to Interspeech, Available at: https://spire.ee.iisc.ac.in/samples_stress/stress_detection_IS_19.pdf*, 2019.

[12] Rodolfo Delmonte, Mirela Petrea, and Ciprian Bacalu, "Slim prosodic module for learning activities in a foreign language," *Fifth European Conference on Speech Communication and Technology*, 1997.

[13] Noriko Umeda, "Vowel duration in american english," *The Journal of the Acoustical Society of America*, vol. 58, no. 2, pp. 434–445, 1975.

[14] Chiranjeevi Yarra, Om D Deshmukh, and Prasanta Kumar Ghosh, "Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5845–5849, 2017.

[15] 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.

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.

[17] Douglas B Paul and Janet 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.

[18] Christopher Cieri, David Miller, and Kevin Walker, "The Fisher Corpus: a resource for the next generations of speech-to-text.," *4th international conference on Language Resources Evaluation*, vol. 4, pp. 69–71, 2004.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, and Petr Schwarz, "The Kaldi speech recognition toolkit," *IEEE workshop on automatic speech recognition and understanding (ASRU)*, 2011.

[20] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al., "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal processing magazine*, vol. 29, pp. 82–97, 2012.

[21] Peter F Brown, Peter V Desouza, Robert L Mercer, Vincent J Della Pietra, and Jenifer C Lai, "Class-based n-gram models of natural language," *Computational linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[22] 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.

[23] Joshua Tauberer, "P2TK automated syllabifier," *Available at https://sourceforge.net/p/p2tk/code/HEAD/tree/python/syllabify/, last accessed on 14-03-2018.*

[24] Piero Cosi, "A KALDI-DNN-based ASR system for Italian," *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–5, 2015.

[25] Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur, "Parallel training of DNNs with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.

[26] Robert Weide, "The CMU pronunciation dictionary, release 0.6," *Carnegie Mellon University*, 1998.

[27] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[28] A Robinson, "BEEP pronunciation dictionary," *Retrieved from World Wide Web: ftp://svr-ftp. eng. cam. ac. uk/pub/comp. speech/dictionaries/beep. tar. gz*, 1996.