CLASSIFICATION OF BISYLLABIC LEXICAL STRESS PATTERNS IN DISORDERED SPEECH USING DEEP LEARNING

Mostafa Shahin[1], Ricardo Gutierrez-Osuna[2], Beena Ahmed[1]

[1]Department of Electrical and Computer Engineering, Texas A&M University, Doha, Qatar
[2]Department of Computer Science and Engineering, Texas A&M University, College Station, Texas

ABSTRACT

Technology-based therapy tools can be of great benefit to children with developmental speech disabilities as they typically require sustained practice with a speech therapist for several years. Towards this aim, over the past 4 years we have developed speech processing tools to automatically detect common errors in disordered speech. This paper presents an automated technique to identify incorrect lexical stress. Specifically, we describe a deep neural network (DNN) that can be used to classify the four different bisyllabic stress patterns: strong-weak (SW), weak-strong (WS), strong-strong (SS) and weak-weak (WW). We derive input features for the DNN from the duration, pitch, intensity and spectral energy on each of the two consecutive syllables. Using these features, we achieve 93% correct classification between SW/WS stress patterns and 88% correct classification of the four bisyllabic patterns on speech from typically developing children, while we obtain 73.4% classification between SW/WS in disordered speech. These figures represent a two-fold reduction in error rates compared to our prior work, which used a DNN with differential features from consecutive syllables.

Index Terms— deep neural network, prosody, lexical stress, automated speech therapy.

1. INTRODUCTION

English is a stress-timed language in which lexical stress plays an important role in intelligibility. Children with a range of speech disorders, including childhood apraxia of speech (CAS), struggle to produce the correct lexical stress patterns [1]. Treatment for CAS involves extensive interactive sessions with a speech language pathologist (SLP) [2]. During treatment, the SLP guides the child on how to control stress levels in pairs of adjacent syllables [3]. Automatic speech therapy tools can facilitate this treatment process, making it more practical and cost-effective and allowing children to practice remotely in their own homes. In earlier work [4] we proposed a client-server architecture to facilitate remote speech therapy for children with CAS. The system consists of a tablet application where children can perform speech exercises, and a remote server running a speech processing engine to detect articulation errors [5], production delays and lexical stress (prosodic) errors.

In this paper, we focus on the detection of prosodic errors in children's disordered speech, specifically the detection of stress level variations between consecutive syllables. We present a deep neural network (DNN) classifier to distinguish between the four possible bisyllabic stress patterns in multi-syllabic English words, strong-weak (SW), weak-strong (WS), strong-strong (SS) and weak-weak (WW). Using features derived from the duration, pitch, intensity and spectral energy of two consecutive syllables, we achieve 88% correct classification of the four bisyllabic patterns on speech from typically developing children and 73.4% classification between SW/WS in disordered speech. This work is unlike the current literature on lexical stress, which has focused primarily on detecting the most stressed syllable in multi-syllabic words.

The rest of the paper is organized as follows. Section 2 briefly summarizes prior work on lexical stress. The methods used for feature extraction, the classifier architecture and the speech corpus used are provided in Section 3. The experiments conducted and results obtained are detailed in Section 4 and finally the conclusions presented in Section 5.

2. PRIOR WORK

The existing work on lexical stress mostly targets adult second language (L2) learning applications, where detecting the most stressed syllable in multi-syllabic words is an important problem. In [6], a system that used the RMS energy of specific frequency bands along with basic prosodic features to automatically detect the most stressed syllable within a word achieved an accuracy of 87%-89%. A separate binary SVM classifier was trained for each vowel to be classified as stressed or unstressed in [7]. In [8], different machine learning algorithms were used to classify different stress patterns in 3&4 syllabic words. When tested on one female user data, it achieved an accuracy of 83% - 88%. Given the success achieved with DNNs [9] in speech recognition, there is now interest in using them to detect lexical stress; a DNN network was developed in [10] to detect syllable stress level in L2 English speech.

In previous work, we computed a variability measure of different sets of acoustic features extracted from pairs of consecutive syllables and used them as inputs in three different shallow machine learning algorithms (SVM,

MaxEnt, MLP) [11] as well as a DNN classifier [12]. The DNN architecture outperformed the shallow classifiers by 5%, with an overall classification accuracy of 85%. In this paper, we use raw acoustic features rather than their differential version to train a DNN classifier to distinguish between the four possible bisyllabic stress patterns. Feeding the DNN with raw features results in a substantial improvement of 10% in classification accuracy compared to a DNN classifier using differential features. The performance of the DNN improves because the many hidden layers of the DNN networks operate as hierarchical feature detectors that capture higher-order correlations between the raw features as similarly observed in [13, 14].

3. METHODS

3.1. System description

Our proposed lexical stress classifier requires access to the speech signal along with the prompted word in the therapy exercise. As shown in Figure 1, the speech signal is first force-aligned with the predetermined phoneme sequence of the word; this is performed using a Hidden Markov Model (HMM) Viterbi decoder along with a set of HMM acoustic models trained from the same corpus. Once the time boundaries for each phoneme have been determined, the algorithm extracts a set of features from 10 msec non-overlapped frames of each syllable and then combines both raw and differential features of each pair of consecutive syllables. Lastly we train two DNN classifiers using the raw and differential feature sets and compare their accuracies.

3.2. Feature extraction

Lexical stress can be identified by variations in pitch, energy and duration between different syllables in a multi-syllabic word [15], with a stressed syllable having higher energy and pitch and longer duration than other syllables within the same word. Accordingly, we extract seven features ($f_1 - f_7$) related to these characteristics as listed in Table 1.

Feature	Description
f_1	Peak-to-peak TEO amplitude over syllable nucleus
f_2	Mean TEO energy over syllable nucleus
f_3	Maximum TEO energy over syllable nucleus
f_4	Nucleus duration
f_5	Syllable duration
f_6	Maximum pitch over syllable nucleus
f_7	Mean pitch over syllable nucleus
f_8	27 Mel-scale energy bands over syllable nucleus

Table 1: The extracted acoustic features

The energy based features (f_1, f_2, f_3) are extracted after applying the non-linear Teager energy operator (TEO), which provides a better estimate of the speech signal energy and also reduces noise [16]. The nucleus and syllable durations are determined from the force alignment process. The pitch values is estimated using the auto correlation method and the mean and maximum values computed over the duration of the nuclei [17]. These seven stress detection features [6-8, 10] are computed for each syllable, resulting in two values per bisyllabic pair. In addition, we also compute Mel scale energies for each frame of the nucleus.

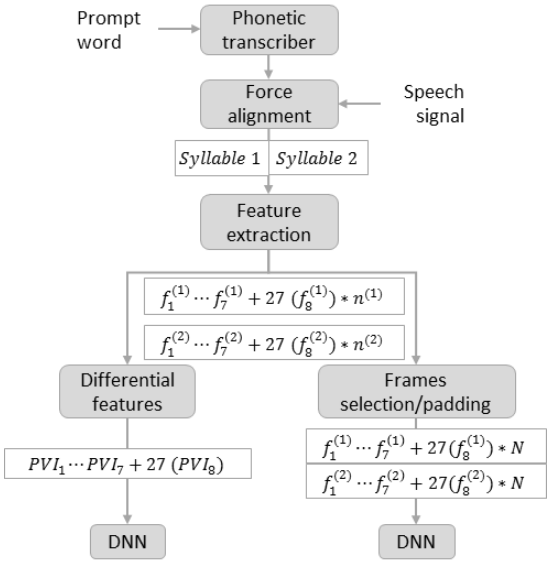

Figure 1: Block diagram of the classification process; $f_i^{(1)}$ and $f_i^{(2)}$ are the ith feature of the first and second syllable respectively, $n^{(1)}$ and $n^{(2)}$ are the number of frames of the first and second syllables' nuclei respectively, PVI_i is the pairwise variability index computed from the *i*th features of the two adjacent syllables and *N* is the number of input frames.

3.3. Raw features

As seen in Figure 1, to input the raw extracted features directly to the DNN, we concatenate the extracted features into one wide feature vector. Each syllable has 7 scalar values $f_1 - f_7$ and $27 * n$ Mel-coefficients where n is the number of frames in each syllable's vowel. To handle variable vowel lengths, we limit the number of input frames provided to the DNN to a maximum N frames for each syllable. This provides the DNN with a fixed length Mel-energy input vector and allows the DNN to use information about the distribution of the Mel-energy bands over the vowel. If the vowel length (n) is greater than N frames, only the middle N frames are used. If the length of the vowel (n) is smaller than N frames, inputs frames are padded to N frames. The final size of the input vector to the DNN is $2 * (7 + 27 * N)$ for a pair of consecutive syllables, with N tuned empirically.

3.4. Differential features

To produce one compact input feature vector for the DNN representing the variation between two consecutive syllables, we compute the pair-wise variability index (PVI) for each feature as in Figure 1 [18]. The PVI of two consecutive syllables for each feature f_i is given by

$$PVI_i = \frac{f_i^{(1)} - f_i^{(2)}}{(f_i^{(1)} + f_i^{(2)})/2} \quad (1)$$

where the subscript i indicates the feature index while the superscript (1-2) indicates the syllable index. Thus a single value for features $f_1 - f_7$ is computed from the two feature values obtained from the two consecutive syllables, e.g. the syllable's PVI vowel TEO maximum energy, PVI duration. As the spectral features, on the other hand, consisted of 27 Mel-coefficients for each frame of the syllable's vowel, we first average the energy in each frequency band over all frames to produce 27 values per syllable. We then compute the PVI of each of these 27 averaged energies. Thus in total, we obtain a total of 34 features to represent each pair of consecutive syllables.

3.5. Deep neural network (DNN)

The DNN in Figure 1 is trained using the mini-batch stochastic gradient decent method (MSGD) with adaptive learning rate. The learning rate starts with an initial value (typically 0.1) and after each epoch the loss in the error of the validation data set is computed. If the loss is greater than zero (i.e. the error increases) the training continues with the same learning rate. If the error continues increasing for 10 consecutive epochs, the learning rate is halved and the parameter of the classifier returned to the one that achieved minimum error. Training is terminated when the learning rate reaches its minimum value (typically 0.0001) or after 200 epochs, whichever is earlier. The performance of the DNN is then computed using a separate testing set. The input size of the DNN is dependent on the type of features (raw vs. differential) and when using raw features, the frame size. We also tune the number of hidden layers and number of units in each layer of the DNN empirically.

3.6. Speech corpus

Due to the limited disordered speech corpora available, we evaluated the algorithm on two speech corpora: the OGI kids' speech corpus [19] and disordered speech we collected from children with CAS. The OGI kids' speech corpus consists of recordings from 1,100 children ranging from grade 0 to 10, with each child pronouncing 200 single words and 100 full sentences. Only clear and correctly pronounced speech files are used to train, validate and test our classifier. We excluded full sentences in our study, focusing only on multi-syllabic single words for which lexical stress patterns could be classified.

Since the speech corpus was collected from native English speakers, we used the CMU English pronunciation dictionary to extract the phoneme sequence and assign stress levels to each syllable. Both primary and secondary stress syllables were marked strong, while the unstressed syllable was marked weak.

The system was tested as well against disordered speech we collected from 10 children with CAS aged 4 - 12 years, each pronouncing 15 isolated words: 10 with a SW pattern across the first two syllables (e.g., DInosaur) and 5 with a WS pattern (e.g., toMAto). These words were selected from the Nuffield Dyspraxia Programme-3 [20] therapy words list. Table 2 shows the data distribution over the training, validation and testing sets.

			Bisyllabic stress patterns			
		Children	SW	WS	SS	WW
Dataset	Training	370	6000	6000	4500	2000
	Validation	70	1000	1000	700	350
	Testing	70	1000	1000	700	350
	CAS	10	115	38	-	-

Table 2: Statistics of the different data sets

4. EXPERIMENTS AND RESULTS

4.1. Raw feature DNN

In a first experiment, we assessed the performance of the DNN using raw features from each syllable in a pair of consecutive syllables, as explained in Section 2.3. For this purpose, we trained two separate DNNs, one to classify between unequal bisyllabic stress patterns SW/WS, and a second one to classify between the four bisyllabic stress patterns SW/WS/SS/WW. Using a fixed vowel frame size of $N = 25$, we evaluated different DNN architectures with 1 to 6 hidden layers and 50 to 600 hidden units per layer. Results are summarized in Figure 2 for the 2- and 4-class DNN.

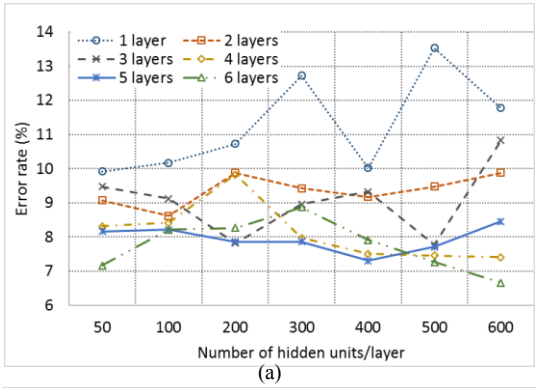

Figure 2: Classification rates as a function of the number of hidden layers and units/layer in the DNN for: (a) unequal bisyllabic lexical stress patterns SW/WS and (b) all bisyllabic lexical stress patterns SW/WS/SS/WW.

The figures indicate that the DNN clearly outperforms the shallow (single hidden layer) NN in both classifiers. In the 4-class DNN, the error rate decreases from 20% for a single hidden layer with 50 units/layer down to 14% by adding one more hidden layer with 200 units. For the 4-class DNN, a minimum classification error rate of 12% is achieved using 4 hidden layers with 500 units/layer; for the 2-class DNN, the best error rate of 7% is obtained using 6 hidden layers with 600 units/layer.

Next, we examined the effect of vowel frame length (as used to extract the Mel energy input vector). Results are shown in Figure 3 for values of *N* from 10 to 30. For both DNNs (2-class and 4-class), error rates decrease initially with an increasing number of input frames till a minimum of 11% at 20 frames (200 msec) for the 4-class and 7% at 25 frames (250 msec) for the 2-class. After that, the performance degrades with an increasing number of frames.

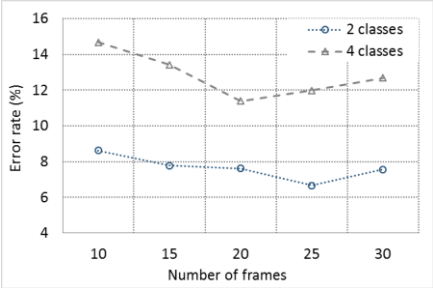

Figure 3: The classification error rate of 2- and 4-class DNN as a function of the number of input frames.

4.2. Comparison of raw and PVI feature DNN

In a second experiment, we examined the ability of the DNN to learn from raw features compared to a DNN with PVI features. Since PVI features only capture the variation in consecutive syllables, they cannot differentiate between SS and WW patterns. Hence, when using the PVI-based DNN, we combined both of these two classes into one class with syllable pairs of Equal stress. The minimum error rate for the PVI-based DNN was obtained using 6 hidden layers with 200 units/layer for the 2-class, and 5 hidden layers with 100 units/layer for the 4-class.

Figure 4 shows error rates for the DNN trained on raw features and the DNN trained on PVI features when classifying SW/WS syllables, whereas Figure 5 shows error rates for classification of SW/WS/SS/WW syllables. These results indicate that raw features increase the performance of both classifiers, with error rates dropping to 6.6% compared to 8.1% when using PVIs to classify SW/WS patterns. The benefits of raw features are more significant in the classification of the 4 classes, where they lead to a two-fold reduction in the overall error rate compared to PVI features.

Finally, we tested the developed system against the disordered speech corpus. As it contained only SW/WS patterns, we used the best performing, binary DNN classifier fed with raw features. The system correctly classified 73% and 75% from the SW and WS patterns with an overall accuracy of 73.4%. This performance degradation can be explained by the pronunciation errors in the disordered speech, resulting in inaccurate phone alignment. It is worth noting, earlier [21] we found that though the inter-rater reliability between two therapists marking lexical stress was 98% for typically developing children, it dropped down to 82% for children with CAS.

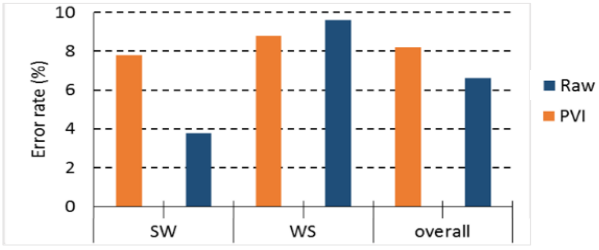

Figure 4: Comparison between the raw and PVI feature DNNs when classifying SW/WS bisyllabic stress patterns.

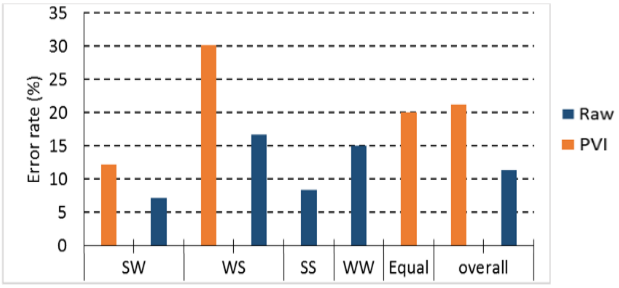

Figure 5: Comparison between the raw and PVI feature DNNs when classifying SW/WS/SS/WW bisyllabic stress patterns. (Equal = SS + WW)

5. CONCLUSION

In this paper we present a DNN classifier to detect bisyllabic lexical stress patterns in multi-syllabic English words. The DNN classifier is trained using pitch, energy and durational features extracted from pairs of consecutive syllables. The feature set of each pair of consecutive syllables is combined by 1) concatenating the raw features into one wide vector or 2) computing a variability index to produce one compact feature vector representing the variation in the features of the two syllables. Test results on children speech show that the DNN performs better when trained with raw features, as they provide more information than the abstract PVI values. In particular, using raw features reduced error rates by 50% on the 4-class problem (SW/WS/SS/WW) when compared to a DNN based on differential features. Our proposed 2-class DNN correctly classifies 93% and the 4-class DNN correctly classifies 89% of the bisyllabic lexical stress patterns in the test dataset of typically developing children and of 73.4% with disordered speech.

6. ACKNOWLEDGEMENT

This work was made possible by NPRP grant # [4-638-2-236] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

7. REFERENCES

[1] A. S.-L.-H. Association, "Childhood apraxia of speech," 2007.

[2] K. Forrest, "Diagnostic criteria of developmental apraxia of speech used by clinical speech-language pathologists," *American Journal of Speech-Language Pathology,* vol. 12, pp. 376-380, 2003.

[3] K. J. Ballard, D. A. Robin, P. McCabe, and J. McDonald, "A treatment for dysprosody in childhood apraxia of speech," *Journal of Speech, Language, and Hearing Research,* vol. 53, pp. 1227-1245, 2010.

[4] A. Parnandi, V. Karappa, Y. Son, M. Shahin, J. McKechnie, K. Ballard*, et al.*, "Architecture of an automated therapy tool for childhood apraxia of speech," in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, 2013, p. 5.

[5] M. Shahin, B. Ahmed, J. McKechnie, K. Ballard, and R. Gutierrez-Osuna, "A Comparison of GMM-HMM and DNN-HMM Based Pronunciation Verification Techniques for Use in the Assessment of Childhood Apraxia of Speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[6] J. Tepperman and S. Narayanan, "Automatic Syllable Stress Detection Using Prosodic Features for Pronunciation Evaluation of Language Learners," in *ICASSP*, 2005.

[7] J. Zhao, H. Yuan, J. Liu, and S. Xia, "Automatic Lexical Stress Detection Using Acoustic Features for Computer Assisted Language Learning," *Proc. APSIPA ASC,* pp. 247-251, 2011.

[8] Y.-J. Kim and M. C. Beutnagel, "Automatic assessment of american English lexical stress using machine learning algorithms," in *SLaTE*, 2011, pp. 93-96.

[9] L. Deng and D. Yu, "Deep Learning: Methods and Applications," *Foundations and Trends in Signal Processing,* vol. 7, pp. 197-387, 2014.

[10] K. Li, X. Qian, S. Kang, and H. Meng, "Lexical stress detection for L2 English speech using deep belief networks," in *INTERSPEECH*, 2013.

[11] M. Shahin, B. Ahmed, and K. J. Ballard, "Automatic classification of unequal lexical stress patterns using machine learning algorithms," in *Spoken Language Technology Workshop (SLT), IEEE*, 2012, pp. 388-391.

[12] M. Shahin, B. Ahmed, and K. J. Ballard, "Classification of lexical stress patterns using deep neural network architecture," in *Spoken Language Technology Workshop (SLT), IEEE*, 2014.

[13] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 21, pp. 1381-1390, 2013.

[14] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605,* 2013.

[15] R. Kager, "Feet and metrical stress," *The Cambridge handbook of phonology,* pp. 195-227, 2007.

[16] H. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 28, pp. 599-601, 1980.

[17] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.", Proceedings of the institute of phonetic sciences. Vol. 17. No. 1193. 1993.

[18] L. E. Ling, E. Grabe, and F. Nolan, "Q uantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English," *Language and speech,* vol. 43, pp. 377-401, 2000.

[19] K. Shobaki, J.-P. Hosom, and R. Cole, "The OGI kids' speech corpus and recognizers," in *Proc. of ICSLP*, 2000, pp. 564-567.

[20] P. Williams, H. Stephens, A. Williams, S. McLeod, and R. McCauley, "Nuffield Centre Dyspraxia Programme," *Windsor: Miracle Factory,* 2004.

[21] M. Shahin, B. Ahmed, A. Parnandi, V. Karappa, J. McKechnie, K. J. Ballard*, et al.*, "Tabby Talks: An automated tool for the assessment of childhood apraxia of speech," *Speech Communication,* vol. 70, pp. 49-64, 2015.