An Approach to Lexical Stress Detection from Transcribed Continuous Speech Using Acoustic Features

József Domokos, Adriana Stan, and Mircea Giurgiu

***Abstract* — This paper presents a first approach to the unsupervised learning and prediction of primary lexical stress starting from continuous speech data and its orthographic transcript. The approach is intended to be used in the development of text-to-speech synthesis systems for under-resourced languages. Our method is based on syllable nuclei approximation and stress detection using simple acoustic features. The evaluation is performed on 3.5 hours of speech uttered by a Romanian female speaker and results show an accuracy of 47.20% at word level and 58.61% at syllable level.**

***Keywords* — lexical stress, stress detection, stress prediction, text-to-speech synthesis.**

I. Introduction

Lexical stress represents one of the most important prosodic aspects of speech. Stress patterns in various languages determine the overall rhythm and melody of that language. While the prosodic accent is speaker or context dependent, lexical stress is generally set for each word, and in some cases it changes its meaning altogether: for example the heteronym: *address* - /ədrɛˈs/ - to speak to the crowd; /əˈdrɛs/ - a postal address.

Lexical stress is also an important component of text-to-speech synthesis systems--correct stress assignment over the utterance can significantly increase its naturalness and intelligibility. However, in most of the world's languages the resources required to train a lexical stress predictor are usually not available or insufficient. We therefore propose an approach to detect and learn the lexical stress position using only speech data and its orthographic transcript. This work is part of the SIMPLE^{4}ALL project [1], whose main objective is to build text-to-speech synthesis systems with little or no expert supervision and knowledge.

Across the wide variety of languages around the world, the lexical stress' placement varies largely. If for example in Czech, Finnish, Hungarian, and Slovakian the stress is always assigned to the first syllable of the word, for other languages, such as Romanian or English, there is no definite rule, and word etymology or morphology can alter it.

Previous studies related to lexical stress detection from acoustic features are mainly focus on isolated words and use supervised algorithms to predict. We index here some of the most relevant work, but emphasize the fact that we are using continuous speech data which was not purposely designed for these experiments.

One of the first studies regarding stress detection from spoken data is that of Aull and Zue [2]. Their study suggests that lexical stress can improve the accuracy of isolated word recognition. Their method determines the sonorant syllables from speech, and extracts simple acoustic features, such as duration, energy and fundamental frequency (F0). These features are then the basis for a reference feature vector constructed for each category: stressed and unstressed. Using the Euclidean distance, the algorithm then determines the classification of each syllable.

In [3] the authors train a Support Vector Machine (SVM) classifier with acoustic features such as duration, loudness, semitone and spectral emphasis extracted from a purposely designed speech database. Their results show an accuracy of 88.57% when an additional post-processing method is applied. The post-processing includes the assignment of a single stressed syllable per word.

SVM classifiers are also used in [4] and compared against decision tree classifiers as well. The study is performed for New Zealand English, and uses only the vowel segments from which they extract prosodic features, as well as vowel quality features. The vowel quality is measured in terms of articulatory features, and the study suggests that in the unstressed syllables, the vowels tend to have a reduced form. The reported accuracy for the stressed vowel detection is 84.72%.

Aside from the automatic detection of lexical stress in speech data as a complete method, we also index the acoustic features found to be relevant to stress detection in general. From the previous cited works, the essential attributes for stress detection were found to be: duration, intensity and vowel quality, in decreasing order of importance. However in [5], F0 level and variation, as well as vowel duration and amplitude are found to be correlated

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287678 (Simple4All).

Corresponding author József Domokos is with the Department of Communications, Technical University of Cluj-Napoca, Bariţiu street 26-28, 400027 Cluj-Napoca, Romania (phone: +40-264-401224; e-mail: domi@ms.sapientia.ro).

Adriana Stan, is with the Department of Communications, Technical University of Cluj-Napoca, Bariţiu street 26-28, 400027 Cluj-Napoca, Romania (phone: +40-264-401224; e-mail: Adriana.Stan@com.utcluj.ro).

Mircea Giurgiu is with Department of Communications, Technical University of Cluj-Napoca, Bariţiu street 26-28, 400027 Cluj-Napoca, Romania (phone: +40-264-401224; e-mail: Mircea.Giurgiu@com.utcluj.ro).

with the word-level stress. In [6] the authors confirm these correlations for different languages: Brazilian Portuguese, English, Estonian, French, Italian and Swedish. The same authors also introduce in [7] the use of overall rise in intensity and spectral emphasis as correlates to focal accent in Swedish. A separate study for Brazilian Portuguese showed that syllable duration, total intensity, duration, F0 standard deviation and spectral emphasis are also a marker for stressed vowels with respect to the unstressed ones [8]. Spectral emphasis was also found relevant for Dutch [9].

To summarize the above studies, prosodic aspects of the central vowel of a syllable can be good indicators of the syllable's stress. And we adhere to these findings by employing similar features for our unsupervised method.

One important aspect to be noticed is that in all the related work, a training dataset was available, and features extracted from this dataset can be easily used in the testing or prediction scenario. However, in this paper our aim is to learn the features of stressed syllables in an unsupervised manner, and without any prior or expert knowledge. This means that given a speech corpus and its orthographic transcript, we define an algorithm which can automatically draw a separation line between stress and unstressed syllables, as well as making predictions for unseen words. The method comprises two main parts: syllable nuclei detection, and a feature extraction and classification.

Our reported results use Romanian as the main language, but our ongoing studies are expanded to other languages as well.

The paper is structured as follows: Section II introduces the method used to automatically determine the syllable nuclei from the speech data. Section III enumerates the features used for our unsupervised classification. The results and conclusions of the work are presented in Section IV and V, respectively.

II. SYLLABLE NUCLEI DETECTION

Lexical stress in most languages is considered to be directly related to the syllabic structure of a word. Within a syllable, the acoustic realization of the stress is in direct correlation with the central vowel, which is also called the syllable nucleus. Therefore, a first step in the unsupervised detection of stress from the acoustic data is to determine the syllable-level segmentation of it. But as syllable boundaries are in most cases, even in written language, vaguely defined, we turn our attention strictly to the approximation of syllable nuclei positions.

A very good method for the automatic detection of syllable nuclei is presented in [10]. The method determines the intensity peaks which are preceded and followed by intensity valleys or dips. The intensity is measured on a logarithmic scale. From the initially hypothesized nuclei, those who consist of unvoiced regions are then discarded. One of the flaws in this method is represented by the fact that unstressed syllable are sometimes omitted. However, although in our subsequent processing steps this might affect the accuracy of our method slightly, we do not consider it to be a major issue.

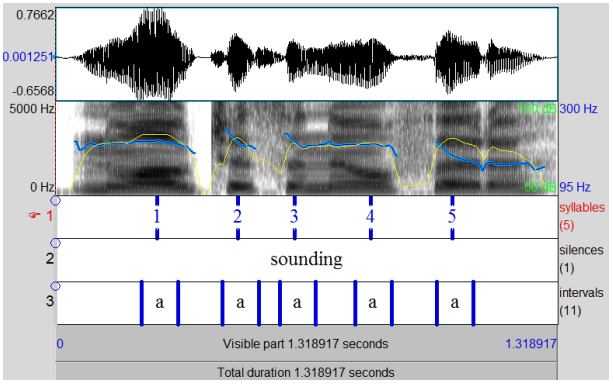

Fig. 1. Screenshot of the Praat syllable nuclei detection script.

In our initial evaluations of the script, the following parameter set gave the best results:

- silence threshold = -25[dB];
- minimum dip between peaks = 0.5[dB];
- Minimum pause duration = 0.3[sec].

As the script provides only the estimated position of the syllable nuclei as a time instance, we expand it to an interval of 80 [ms] centered on the estimated position, and extract the acoustic features from this time interval. Figure 1 shows a sample screenshot of the script's output. Tier *syllables* marks the detected syllable nuclei. Tier *silences* represents a voice activity detector, and determines if the current segment is speech or silence. Tier *intervals* shows the 80 [ms] window centered around the detected syllable nuclei.

III. FEATURE EXTRACTION

Once the syllable nuclei have been approximated, and using the results of the surveyed studies from Section I, we extract the following acoustic features from the 80 [ms] window:

a) *fundamental frequency*, its maximum and minimum values, mean and standard deviation;
b) *intensity*, its maximum and minimum values, mean and standard deviation;
c) *fundamental frequency converted to semitones on the logarithmic scale*:

$$S = 69 + 12\,log_2(\frac{F0}{440}) \quad (1)$$

d) *duration of the syllable nucleus;*
e) *first three formant values;*
f) *harmonics to noise ratio (HNR)* - where PP% is the percentage of periodic signal's energy, and NP% is the noisy part of the signal. An HNR of 0[dB] represents a signal with equal amount of periodic and aperiodic energy. HNR is used as a measure of vowel quality, and it is computed as:

$$HNR = 10\,log_{10}(\frac{PP\%}{NP\%}) \quad (2)$$

g) *spectral tilt attributes*, computed as in [11]. We compute the following measures: **H1-H2** – difference of the amplitude of the first harmonic and the amplitude of the second harmonic, which is an indicator of the relative length of the opening phase of the glottal pulse; **H1-A1**, the

difference between the amplitude of the first harmonic and the strongest harmonic of the first formant, and represents the spectral tilt; **H1-A2**, the difference between the amplitude of the first harmonic and the strongest harmonic of the second formant, the spectral tilt at middle formant frequencies; **H1-A3**, the difference between the amplitude of the first harmonic and the strongest harmonic of the third formant represents the spectral tilt at higher formant frequencies.

These parameters comprise the feature vector used in the lexical stress detection algorithm described in the next section.

IV. Lexical Stress Detection

We again emphasize the fact that our method does not rely on an existing purposely designed training dataset, or any other form of expert knowledge. Therefore, the devised algorithm should be able to automatically determine the stressed and unstressed syllables from the speech data, and then create a text-based dataset for the lexical stress predictor.

Figure 2 shows the flow chart of the proposed method, and we present next its main steps. In order to build a text dataset for the stress predictor, the simple marking of stressed syllables in the audio data is not sufficient. We need to establish a direct correspondence between the speech-based stress marking and the text. For this we used self-trained acoustic models, and forced alignment. The forced alignment and acoustic model training use graphemes as basic units, and this again ensures that expert knowledge is not required. This step however, is most likely to introduce additional errors in our method, as the accuracy results depend highly on the amount of data available and the complexity of the letter-to-sound rules in a particular language.

Using the syllable nuclei detection method we then extracted the parameter sets from the 80 [ms] window, individually for each utterance.[1] Each utterance's acoustic features dataset underwent a Principal Component Analysis (PCA) feature reduction and transformation step. The results of the PCA analysis were then clustered using a simple k-means algorithm. A simple assignment of each cluster to the stressed and unstressed categories would be to assign the most populated cluster to the unstressed category. However, in utterances where there are more monosyllabic words, this hypothesis does not hold. Therefore, we test the stress detection algorithm using both clusters, and select the one which assigns the least number of stressed marks in polysyllabic words to be the one which represents the stressed syllables feature set. As a result, all the detected syllables from a certain utterance have an assigned stressed or unstressed marker.

Having the text alignment and the stress markings, for each word we determine the stressed syllable's central vowel, and build a text dataset for lexical stress prediction. The lexical stress prediction algorithm and results are not presented in this paper.

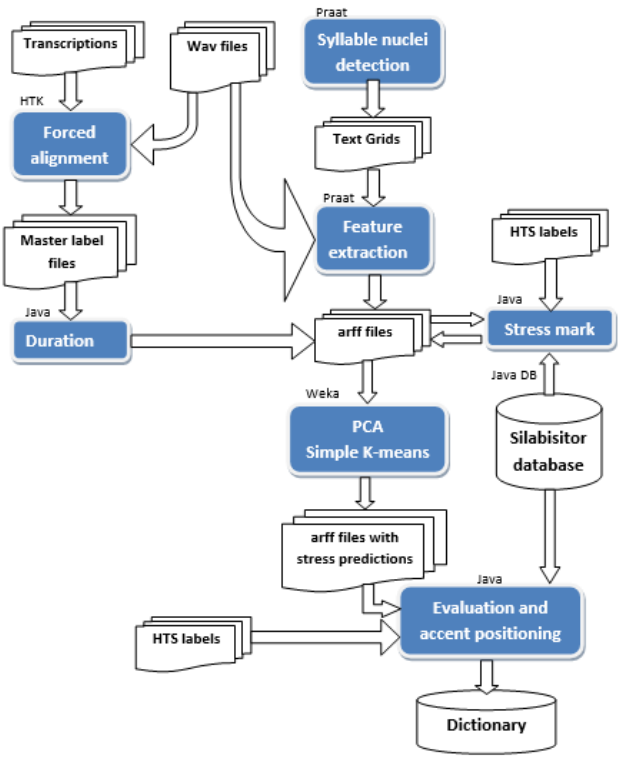

Fig. 2. Lexical stress detection system flow chart.

V. Experimental Results

In our experiments we used a 3.5 hour speech dataset uttered by a Romanian female speaker, a subset of the Romanian Speech Synthesis (RSS) corpus [12]. The results reported here use the following 5 subsets: *diph1*, *diph2* – which contain 983 utterances selected for diphone coverage; *rnd1, rnd2, rnd3* – contain 1493 randomly selected sentences from newspaper articles. The data is read with a flat intonation and was designed to be used in statistical parametric speech synthesis systems.

For the evaluation of the stress detection, we adopt a GOLD standard database for syllabification and stress assignment in Romanian, called "Silabisitor" [13]. The entire method is developed in Java and Praat, and uses the Weka machine learning toolkit [14]. The forced alignment at grapheme level is obtained using the HTK toolkit [15].

Tables 1 and 2 present the results for each speech subset in terms of word and syllable error rate, respectively, and using the grapheme-level alignment. The word error rate (WER) in Table 1 is computed as a function of the number of words which had a correct stress assignment (CSW), total number of words (TNrW), unverifiable words (UW)[2], and words which have multiple stresses detected (MAW):

$$WER = 1 - \frac{CSW}{TNrW - UW - MSW} \quad (3)$$

[1] We assume that the speech and text dataset is segmented into utterance-length chunks.

[2] Words which were not found in our GOLD standard stress assignment database.

TABLE 1: WORD ERROR RATES [WER] FOR THE RSS SUBSETS, USING GRAPHEME TRANSCRIPTIONS.

Speech subset	#words	CSW	UW	MSW	WER [%]
diph1	1 899	723	228	182	51.44
diph2	1 563	539	284	155	52.05
rnd1	2 950	1 119	340	255	52.48
rnd2	1 428	476	222	182	53.52
rnd3	1 583	546	254	172	52.81
Total	**9 423**	**3 403**	**1 328**	**946**	**52.40**

TABLE 2: SYLLABLE AND STRESS DETECTION ERROR RATES.

Speech subset	#syllables	#syllable nuclei	SyDER [%]	SyER [%]
diph1	3 078	2 698	12.34	41.33
diph2	2 680	2 125	20.70	42.73
rnd1	4 998	4 709	5.78	39.82
rnd2	2 660	2 007	24.54	42.35
rnd3	2 893	2 309	20.18	42.62
Total	**16 309**	**13 848**	**15.08**	**41.39**

TABLE 3: STRESS DETECTION ERROR RATES USING PHONE LEVEL ALIGNMENT.

Speech subset	SyER [%]
diph1	38.37
diph2	38.88
rnd1	40.92
rnd2	40.15
rnd3	40.06
Total	**39.82**

Table 2 also includes a comparison between the actual number of syllables in the speech data, and the detected syllable nuclei using the Praat script (SyDER). The stress assignment syllable error rate is computer against the detected number of syllable nuclei, as we cannot assume that the undetected syllable nuclei are classified either correct or incorrect. By doing this we also evaluate the efficiency of the syllable detection algorithm. It can be noticed that the error rates are high, both in terms of word (52.4%), as well as syllable (41.3%) level. However, this is to be expected, as most of the processing steps are unsupervised and a sum of assumptions are made about the acoustic behavior of syllable and stress realization.

Another source of errors for our method is the use of grapheme level alignments instead of phone level ones. Table 3 presents the results of our method when the alignment of the speech and text data is performed at phone level. A similar measures of the error rate as in Table 2 is used. The results are only slightly better than in the case of the grapheme level alignment. Though, this was to be expected as Romanian has very simple letter-to-sound rules. We hypothesize that for more complex languages, such as English for example, this difference should increase dramatically.

VI. CONCLUSIONS AND FUTURE WORK

This paper introduced a first attempt of unsupervised detection of lexical stress starting from continuous speech data and its orthographic transcript. The method comprises two main steps: a syllable and a stress detection algorithm, and uses simple acoustic features extracted from the speech data. It does not rely on any previous existing knowledge, and supports any language as long as the required data is available.

The results are modest at this point, with only around half of the syllables having assigned the correct stress category. However, by employing additional pre- and post-processing, as well as alternative clustering procedures, we expect the method to improve its performance. And, as stated in the scope of the paper, the stress detection text results to be used in the prediction of stress for text-to-speech synthesis systems purposes.

REFERENCES

[1] SIMPLE^{4}ALL project website: http://simple4all.org/

[2] A. Aull, V.W. Zue, "Lexical stress determination and its application to large vocabulary speech recognition", *in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '85*, pp. 1549 – 1552, Apr 1985.

[3] J. Zhao, H. Yuany, J. Liuy and S.H. Xia, "Automatic Lexical Stress Detection Using Acoustic Features for Computer-Assisted Language Learning", in *Proc. APSIPA ASC*, October 2011.

[4] H. Xie, P. Andreae, M. Zhang, P. Warren, "Detecting Stress in Spoken English using Decision Trees and Support Vector Machines", in *Proc. of the ACSW Frontiers 2004*, Volume 32, pp. 145-150

[5] J. Volín, L. Weingartová, "Acoustic correlates of word stress as a cue to accent strength", *Research in Language,* 2014, Vol. 12:2, pp. 175-183.

[6] A. Eriksson, P.A. Barbosa, and J. Åkesson, "Word stress in Swedish as a function of stress level, word accent and speaking style" in *Proc. Nordic Prosody 2012*, pp. 127–136.

[7] A, Eriksson, P.A. Barbosa and J. Åkesson, "The Acoustics of Word Stress in Swedish: A Function of Stress Level, Speaking Style and Word Accent" in *Proc. Interspeech 2013*, pp. 778–782.

[8] P.A. Barbosa, A. Eriksson, and J. Åkesson, "On the Robustness of some Acoustic Parameters for Signalling Word Stress across Styles in Brazilian Portuguese" in *Proc. Interspeech 2013*, pp. 282–286.

[9] R.C. van Dalen, P. Wiggers and L.J.M. Rothkrantz , "Lexical Stress in Continuous Speech Recognition", in *Proc. Interspeech 2006*, pp.

[10] N. H. de Jong and T. Wempe "Praat script to detect syllable nuclei and measure speech rate automatically", *Behavior Research Methods*, May 2009, Volume 41, Issue 2, pp 385-390.

[11] C. DiCanio. (2012). Spectral Tilt Script for Praat, available: http://www.haskins.yale.edu/staff/dicanio/scripts.html

[12] A. Stan, J. Yamagishi, S. King, M. Aylett, "The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate", in *Speech Communication*, vol. 53, no. 3, pp. 442-450, 2011.

[13] Software application for Romanian language word syllabification (Silabisitor): http://ilr.ro/silabisitor/

[14] WEKA 3: Data Mining software in Java: http://www.cs.waikato.ac.nz/ml/weka/

[15] The Hidden Markov Model Toolkit (HTK) for Speech Recognition webpage: http://htk.eng.cam.ac.uk/