Acoustic Stress Detection in Isolated English Words for Computer-Assisted Pronunciation Training

Vera Bernhard1, Sandra Schwab2, Jean-Philippe Goldman3

1University of Zurich, Switzerland
2University of Bern, Switzerland
3University of Geneva, Switzerland
veralara.bernhard@uzh.ch, sandra.schwab@unibe.ch, jean-philippe.goldman@unige.ch

Abstract

We propose a system for automatic lexical stress detection in isolated English words. It is designed to be part of the computer-assisted pronunciation training application MIAPARLE (”https://miaparle.unige.ch”) that specifically focuses on stress contrasts acquisition. Training lexical stress cannot be disregarded in language education as the accuracy in production highly affects the intelligibility and perceived fluency of an L2 speaker. The pipeline automatically segments audio input into syllables over which duration, intensity, pitch, and spectral information is calculated. Since the stress of a syllable is defined relative to its neighboring syllables, the values obtained over the syllables are complemented with differential values to the preceding and following syllables. The resulting feature vectors, retrieved from 1011 recordings of single words spoken by English natives, are used to train a Voting Classifier composed of four supervised classifiers, namely a Support Vector Machine, a Neural Net, a K Nearest Neighbor, and a Random Forest classifier. The approach determines syllables of a single word as stressed or unstressed with an F1 score of 94% and an accuracy of 96%.

Index Terms: automatic prominence detection, lexical stress detection, computer-assisted language learning, computer-assisted pronunciation training, machine learning, supervised classification

1. Introduction

Although language learning websites and apps are very popular, they often lack prosodic training. The web application MIAPARLE, developed by [1], aims to address this gap and hence provides a range of tools to train the perception and production of lexical stress. Lexical stress describes the phenomenon that not all syllables within an uttered word are perceived as equally prominent [2]. In multi-syllabic English words, one syllable is always particularly salient and hence it carries so-called primary stress ([3], [4]). In English, stress can carry a grammatical function (e.g. difference in word class: ‘import’ - noun vs. ‘im’port - verb) and, according to [3] even more importantly, stress is also a crucial perceptual cue to detect word boundaries in speech.

Prosodically, or more specifically lexically stress, cannot be disregarded in language education as on the one hand, it is significantly affected by language transfer [5]. This means that the way prosody is produced in a learner’s mother tongue (L1) leads to inaccurate or erroneous production in the target language (L2). Lexical stress, poses difficulty for speakers whose L1 is a fixed-stress language like French when learning a free-stress language such as English [1]. In fixed-stress languages, stress placement is determined by a fixed rule for all words, whereas in free-stress language, stress patterns differ from word to word and might be assigned according to more complex rules. In English, for example, stress is often assigned according to the complexity of the syllables [4]. On the other hand, correct stress patterns also have a considerable impact on the intelligibility [6] and perceived fluency of a speaker [7].

On the CAPT (computer-assisted pronunciation training) website MIAPARLE, the correction of stress production is conducted by a lexical stress detector. It requests the L2 learner to pronounce a certain word and returns which syllable is stressed and whether this stress pattern is correct. Due to the importance of training and due to the current lack of English language support in the production exercise on MIAPARLE, this paper proposes a pipeline to detect stress patterns in isolated English words to provide immediate feedback to the learner within the web application.

2. Automatic Stress Detection

In the context of computer-assisted language learning, lexical stress detection is defined as identifying the stressed syllable within a spoken word [8]. Even though the entire word is taken into consideration, the classification is often executed on the level of a syllable, but the classifier assigns probabilities to each syllable and chooses the one with the highest score as the location of primary stress. This approach is adopted in many previous works (e.g. [8], [9], [10]).

Lexical stress detection is commonly addressed as a supervised machine learning problem. As a consequence, the truth labels, the correct stress patterns, as well as the data, usually syllables, are represented as feature vectors. Thus, the first step of preprocessing the training data includes the detection of syllables within the speech recordings. Syllable alignment is proven to be a difficult task which contributed significantly to the error of early models [11]. This was the reason that syllable boundaries were often annotated manually instead [12], [13] whereas more recent works rely on state-of-the-art syllable aligners (e.g. EduSpeak in [6]).

For the truth label, each syllable has to be annotated with its stress value. In case of the speech data being produced by natives, it is often assumed that they produce the correct stress pattern, and hence the truth labels can be retrieved from a dictionary ([8], [6]). Otherwise, the speech data has to be manually annotated, which is a time-consuming and therefore a costly task [6].

Each syllable is then represented by a feature vector consisting of diverse acoustic measures trying to capture
whether the syllable is carrying stress. However, how stress manifests acoustically is an intensively debated question. Research agrees that “[S]tress […] is never marked by a single acoustical property […]” [4]. Rather, stress is signaled by a complex interplay of relative changes in several acoustic dimensions [14].

In perceptual studies, it has been found that syllables with longer duration, higher intensity and higher fundamental frequency are perceived as stressed ([15], [4], [16]). Additionally, in English particularly, vowels are known to differ in vowel quality depending on stress ([17], [3]). Inspired by the perceptual studies, early attempts of automatic stress detection relied on simple correlates of duration, loudness and pitch ([12], [11], [13]).

The work of [11] and of [13] additionally explored some spectral features, namely spectral change, and spectral envelope and slope. Yet, both concluded that loudness and fundamental frequency were their most successful features and that they outperformed the spectral features. Including spectral features gained momentum again when [18] proposed a new spectral measure, namely spectral balance. They found that the increase in intensity in a stressed syllable is mostly located above 500 Hz. This finding was operationalized as the intensity difference in three frequency bands above 500 Hz and which was found to be close in reliability to duration measures. Spectral balance is adopted in several following works (e.g. [9], [10]) and has motivated similar spectral measures as in [6].

In the research of [14], they first introduced context-aware features by considering durational, loudness, and pitch features from the previous syllable. The idea of contextual features has become popular in the close past ([9], [6], [7]). Some studies (e.g.[19] or [17]) have attempted to implement a measure of vowel quality but it was classified as a weak correlate. In recent years, spectral information is commonly provided as mel-frequency cepstral coefficients (MFCC), which are reported to outperform the traditional features such as duration or intensity [6, 20].

Many different machine learning algorithms have been trained on the feature vectors. Common classifiers include Naive Bayes [9], Logistic Regression [9], Decision Trees ([12], [17], [9]), Gaussian Mixture Models ([8], [6]), Support Vector Machines ([17], [9], [10]) and different Neural Network architectures ([14], [6], [7], [20]). Defining the most successful approach is a difficult task, as the works differ widely in what training data they use, what and in which manner the features are calculated and how the models are evaluated. Only a few studies actually compare different algorithms: [17] concluded that Support Vector Machines outperform Decision Trees. Decision Trees are also challenged in the work of [6] where Decision Trees and also Neural Networks were exceeded by Gaussian Mixture Models. In contrast, Decision Trees perform well in the work of [9], where they achieved better results than Naive Bayes, Logistic Regression, and Support Vector Machines.

3. Material and Methods

3.1. Data

The training data, collected by Schwab and colleagues financed by the Lehkredit of the University of Zurich, comprises 1012 recordings of 92 different isolated American English nouns spoken by 6 English natives (three female and three male). Following the examples of [8] and [6], it was assumed that natives produced the words correctly, and the stress patterns were retrieved from a dictionary. In each audio file, the syllable boundaries were annotated manually. In total, the training data consists of 3054 syllables of which 67% are unstressed, and 33% are stressed syllables.

3.2. General Pipeline

In Figure 1, an overview of the pipeline, which is intended to be included within MIAPARLE, is depicted: The pipeline receives an audio recording of a single word spoken by a language learner together with the orthographic transcription. The transcription was performed by the ’G2P’ (Grapheme to Phoneme) service included in the BAS (Bavarian Archive of Speech Signals) Web Services [21]. Afterwards, with the help of the phonetic transcription, syllables are detected in the speech signal by applying two other BAS web services, ’MAUS’ [22] (phone segmentation) and ’PHO2SYL’ (phones to syllable) [21]. Each syllable is then represented by a feature vector consisting of measures calculated on the syllable nucleus. The linguistic reason for this is that “[…] large part of prosodic stress information is carried by the vocalic nucleus […]” [23]. In practical terms, detecting the syllable nucleus is easier than the entire syllable ([8], [7]) and it simplifies the normalization procedure [24]. To account for variances that are not due to stress placement, the feature vectors are then normalized and scaled before being sent to the classifier.

The classifier considers syllable by syllable and assigns each a probability score. Finally, a postprocessing step as in previous research, such as the works of [8], [9] and [10], was included to ensure that only one stressed syllable per word is detected. Both, when no syllable or several syllables are recognized as stressed, the syllable with the highest probability is classified as stressed.

3.3. Acoustic Feature Extraction

The first measure chosen to be calculated was duration, as it has been undisputedly considered a correlate of stress in previous literature. Additionally, loudness and pitch features were extracted from each syllable. Lastly, motivated by the statement of [18] “[…] spectral balance is a clear acoustic correlate of stress […]”, spectral information in the form of spectral balance was included. This work did not consider any features representing vowel quality as neither the research of [19] nor [17] showed convincing results.
Since a stressed syllable is defined by its differences in acoustic dimensions to the neighboring syllables, this work followed the approach of [14] and [10] and calculated for each feature a set of differential values to the normalized values of the previous and succeeding syllable. In case the syllable stands at the first or last position in the word and thus misses a left or right neighbor, the approach of [10] was copied: The mean value over all syllables of the feature replaced the missing neighboring feature.

The features were calculated using the Praat [25] (an open-source software for speech analysis) plugin Prosogram [26] or custom Praat scripts. Prosogram analyses pitch variations in speech and can measure prosodic features per syllable. For this purpose, it places great importance on applying stylization on the pitch to reflect human perception as accurately as possible [27].

3.3.1. Durational Features

As durational features, the nucleus and syllable duration were extracted. To reduce the effects of the individual speech rate and to counteract the fact that the final syllable of an utterance tends to be longer (a phenomenon called ‘preboundary lengthening’), the nucleus duration and the syllable duration were normalized by dividing it through the mean nucleus respectively syllable duration of the given word ([10], [18]). Vowels that constitute the syllable nucleus vary in length according to intrinsic differences, as pointed out by [24] and [17] (e.g. diphthongs are longer than ordinary vowels regardless of stress placement). Therefore, the average duration of each vowel phoneme was calculated on the entire dataset, and each nucleus duration was normalized further by dividing it by the corresponding mean vowel phoneme length. Both values, normalized for speech rate and normalized for vowel type, were added.

3.3.2. Loudness Features

To depict loudness, measures describing the energy in Decibel (dB) and amplitudes in Pascal have been used in previous literature. This work applied an amplitude feature, namely the root mean square amplitude motivated by [23] and [17], and peak intensity motivated by [4]. The mean values for both measures retrieved from all nuclei of a word were subtracted from the value per syllable to normalize for speaker differences and varying recording setup.

3.3.3. Pitch Features

Following several previous studies, peak and mean fundamental frequency were calculated ([11], [14], [23], [17], [8], [9], [10]). Since Prosogram also provided maximal F0 stylized and mean F0 in semitones (ST), these features were incorporated as well. The selection of a measure in semitones was also motivated by [10]. The maximum and mean F0 features were normalized by subtracting the corresponding average value over all nuclei of a word. Apart from average and peak F0, inspired by measures related to the movement of F0 in a syllable as in [13] or [7], Prosogram’s ‘trajectory’ features describing “the sum of absolute pitch interval of tonal segments [...] after stylization” [26] was added as well. In addition to the context-aware features, the difference in semitones between the end of the previous nucleus and the start of the current nucleus was calculated using Prosogram (called ‘intersyllab’).

Table 1: Baseline performance of all classifiers without and with postprocessing (pp).

<table>
<thead>
<tr>
<th>Classifier</th>
<th>F1</th>
<th>F1 PP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural Network</td>
<td>0.89</td>
<td>0.93</td>
</tr>
<tr>
<td>Nearest Neighbors</td>
<td>0.86</td>
<td>0.91</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.87</td>
<td>0.90</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.85</td>
<td>0.90</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.84</td>
<td>0.87</td>
</tr>
<tr>
<td>AdaBoost</td>
<td>0.81</td>
<td>0.85</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.74</td>
<td>0.79</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.76</td>
<td>0.76</td>
</tr>
</tbody>
</table>

3.3.4. Spectral Features

This work adopts spectral balance inspired by [18]. Therefore, the mean intensity in three frequency bands, namely between 500-1000 Hz, 1000-2000 Hz, and 2000-4000 Hz was determined.

3.4. Training a Classifier

Since a single best algorithm could not be identified in previous literature, eight different classifiers were implemented with the Python library Scikit-Learn [28] (the code and the final model can be found on GitHub: “https://github.com/vera-bernhard/stress-detector”): Decision Trees, Logistic Regression, Support Vector Machines, Naive Bayes and a simple Neural Network architecture, more precisely a Multilayer Perceptron, K Nearest Neighbor classifier and two Ensemble classifiers, namely Random Forest and AdaBoost. Since most machine learning algorithms are sensitive to different scales of the data, the training data was scaled with Z-score normalization [29].

To evaluate the classifiers, 10-fold cross-validation with a stratified 75:25 training-test split was performed. In this paper, the evaluation relied on F-Score (also called F1 score) as it is better suited to measure the model performance on an unbalanced data set. Stress annotated English data is always unbalanced because there can be one syllable carrying primary stress at most; with our data, 2/3 of the syllables are unstressed. Additionally, since the lexical stress detector devised in this work is part of a CALL application, it should avoid false corrections (i.e. minimize false negatives and false positives) [6].

From the eight classifiers, the four achieving the highest F1 score were chosen to be further investigated. This included, on the one hand, feature engineering - to discover the best performing feature combinations - and, on the other hand, optimizing the hyperparameters of the remaining four models.

Therefore, the performance of different feature group combinations was analyzed by training the four algorithms with only a certain feature group. Within a feature group, it was manually experimented whether removing potentially correlating features (e.g. maximum Pitch in ST and Hz) and different normalization strategies improved the model performance. After that, a more systematic grid search on the four models was performed to retrieve the best hyperparameters. Again, F1 was the evaluation measure for which the grid search was optimized.

As a last optimization experiment, the four improved classifiers were combined into a Voting Classifier. The idea of a Voting Classifier comprises that all four classifiers assign probabilities to an item, in our case a syllable, and the average overall probability defines which class is associated with the
item [29]. Voting Classifiers are known to outperform the individual classifiers they are composed of [29].

4. Results and Discussion

The Neural Network, Support Vector Machine, K Nearest Neighbor and Random Forest classifiers performed the best when implemented out of the box with all features (s. Table 1). The success of the postprocessing strategy can be seen in the same table, as each algorithm reached a higher F1 score when postprocessing was applied, which reflects the findings of [8].

From all features, the durational features proved to be the most successful. This ties in with the fact that durational measures were the uncontested feature in previous literature. The strength of pitch, loudness, and spectral features depends on which classifier is considered. For the Support Vector Machine, K Nearest Neighbor, and Neural Network, spectral features led to a better performance than pitch and loudness features. With the Random Forest, the pitch features surpassed loudness and spectral features. Nevertheless, the combination of pitch, loudness, spectral and, duration features yielded the best performance with all classifiers.

The juxtaposition of absolute, normalized and, context-aware features revealed that in agreement with the work of [10], the context-aware feature group outperformed the normalized and absolute features with each classifier if the groups are considered on their own. This finding is supported by the fact that stress is defined as a phenomenon on the level of the syllable in relation to its neighboring syllables. The best F1 score was reached with either the combination of absolute, normalized, and context-aware features or absolute and context-aware features. Correspondingly, both combinations were tested when further optimizing the classifier. Finally, removing some similar (e.g., vowel normalized vs. speech rate normalized duration features) or collinear features (e.g., F0 mean in Hz vs. F0 mean in ST) was explored. The removal of any feature led to a decrease in performance. As a consequence, all features were considered.

The F1 score of each of the four best performing classifiers could be slightly improved by 0.01-0.04 by applying grid search to optimize the hyperparameters. Then, they were combined into a Voting Classifier. The resulting Voting Classifiers, as predicted by [29], outperformed the single classifiers (s. Table 2). Unfortunately, no definite conclusion between the two feature sets could be drawn; both resulting classifiers reached an F1 score of 0.94. Since a model with less features is usually preferred because less time for training is required, the Voting Classifier trained on absolute and context-aware features was chosen to be implemented in the lexical stress detection pipeline. Therefore, the final model achieves an F1 score 94% and an accuracy of 96%. The classifier performs slightly better at detecting unstressed than stressed syllables (97% of unstressed and 94% of stressed syllables are detected correctly (s. Figure 2). This can probably be accredited to the fact that unstressed syllables are overrepresented in the training data as described when introducing the data.

5. Conclusion

This paper proposes a pipeline that successfully detects syllables in recordings of isolated English words and identifies the syllables carrying primary stress. For syllable alignment, three different tools from BAS Web Services were exploited. The lexical stress detector was trained with a supervised machine learning algorithm: A Voting Classifier composed of four different classifiers, including a Support Vector Machine, a Neural Net, a Random Forest and a K Nearest Neighbor classifier, achieved a 94% F1 score and 96% accuracy on the provided training data. The final classifier employed absolute and context-aware duration, pitch, loudness, and spectral features, which in combination led to the best performance. The durational features proved to be the most reliable feature group to detect lexical stress in this task setup. Apart from feature selection, a postprocessing step that ensured that only one stressed syllable per word is found contributes substantially to the quality of the classification.

Future work should investigate how the proposed pipeline performs on non-native and noisy data since this was not evaluated in the presented work but is the use case scenario on MIAPARLE. Moreover, the proposed pipeline only detects primary stress. Detecting different degrees of stress could be a topic of future research. Furthermore, the pipeline could also be improved by taking mispronunciations of a language learner on MIAPARLE into consideration. At the current state, if a vowel is mispronounced, it is still force aligned with the correct pronunciation in the preprocessing steps. Lastly, exploring Deep Neural Network architectures, which are able to represent time series (for example, recurrent neural networks (RNN) as in [20]) and employing spectral information from MFCCs may constitute the object of future studies.
6. References


