Evaluation of Syllable Stress using Single Class Classifier

Abhinav Parate, Ashish Verma, Jayanta Basak

IBM India Research Laboratory, New Delhi, India
abparate@in.ibm.com, vashish@in.ibm.com, bjayanta@in.ibm.com

Abstract

Evaluation of syllable stress in speech utterances is an important and challenging task in the area of speaker evaluation. In this paper, we propose a method to classify correct utterances of English words based on the evaluation of the lexical syllable stress pattern. Here we use only correctly stressed utterances of the words as training samples since a statistically significant pool of incorrectly stressed utterances is difficult to obtain. The underlying assumption here is that the correct utterances of a word form a compact cluster or a collection of compact clusters (with speaker dependent variations) in a suitably chosen multi-dimensional attribute space. We experimentally demonstrate the effectiveness of the proposed method on several English words and also compare with the standard classifiers where samples from both correct and incorrect utterances were used.

1. Introduction

Analysis of stressed syllables in speech has been an area of research for a long time. In spoken English, the syllables can be stressed either due to prosodic stress or due to lexical stress upon the word being spoken. This paper focuses on the lexical stress which plays an important role for efficient communication. In particular, we address the task of classifying English words spoken by Indian speakers into correct and incorrect classes considering the stress on various constituent syllables. In English, every polysyllabic word has a primary stress syllable and sometimes one or more secondary stress syllables. The position of the primary and secondary stress syllables is fixed for a given word. As many of the Indian languages do not have intra-word stress, Indian speakers often incorrectly stress English words.

Stress is generally considered to be a manifestation of pitch (or fundamental frequency), duration and energy (or volume). Most of the studies in this area have used these three acoustic features and their various combinations. Stressed syllables often have higher values of these acoustic features as compared to the case when they are unstressed. However, the amount of variation depends upon the particular syllable being stressed, the vowel (also called nucleus of the syllable) contained in the syllable and the speaker.

A significant amount of work has been done for identification of stressed syllables. Silipo et al. have investigated the role of duration, amplitude and fundamental frequency in identifying stressed syllables in spontaneous speech. They have concluded that duration and energy are more important in identifying stressed syllables as compared to fundamental frequency. More importantly, they have shown the effectiveness of derived features, such as (energy × duration), as compared to using the individual features alone. They have also shown that consonants are least affected by stress while vowels are most affected. Jenkin et al. have used only vowel (or nucleus) based features and report the accuracy of neural networks, Markov chains and rule based classifiers in identifying stressed syllables [2]. Imoto et al. have used Mel Frequency Cepstral Coefficients (MFCC) in addition to fundamental frequency, energy and duration features for detection of sentence level syllable stress [3, 6]. They determine the respective weight of these features for classification of stressed syllables using Linear Discriminant Analysis (LDA).

In this paper, we evaluate primary stress pattern of isolated words where only one syllable is stressed in each word. Tepperman et al. have addressed this problem in the context of a language learning system [4]. They used Gaussian mixture models to model stressed and unstressed syllables and quadratic Bayes Discriminant function for classification. The experiments were performed on English words spoken by German and Italian speakers. They consider three basic acoustic features, viz., fundamental frequency, energy and duration [4], and other features derived from these basic features, such as, the ranges and slopes of fundamental frequency and energy within a syllable. They normalize the duration related parameters using many context-dependent rules which improved the performance by 5% to 10%. A word-independent model to evaluate words for lexical stress has been proposed in [7]. Authors use generalized probabilistic models for stressed and unstressed syllables and then combine the likelihood of being stressed for each of the constituent syllables in a word dependent naive-Bayesian framework. It is shown that the word-independent model performs slightly inferior as compared to the word-dependent models but with the advantage of scalability.

In this paper, we propose a novel classification method for evaluation of lexical stress in spoken English words based on a single class model. In this method, the classification is performed using a single class which is modeled by only the correctly stressed utterances of a word. Researchers have used single class model based techniques, such as, One class Support Vector Machines (OSVM) [8] and Positive Example Based Learning (PEBL) [9], for classification in other domains, such as web page classification. The similarities and differences of these techniques from the proposed method are highlighted in Section 2.

2. Proposed Approach

Motivation:

For English words, there exist only one correct primary stress pattern where only one syllable is usually stressed. Every other stress pattern is considered to be incorrect. Since incorrect utterances can happen due to various reasons such as incorrectly
stressed syllable and incorrect amount of stress, all possible samples of incorrect utterances potentially lend themselves to an infinite region in the multi-dimensional attribute space. It is therefore very difficult, in real life, to obtain a statistically significant pool of incorrect utterances to have a representative distribution. On the other hand, the correct utterances, due to their very nature, can be assumed to form a compact region or a set of compact regions in the same multi-dimensional attribute space subject to the speaker characteristics. Moreover, obtaining representative samples of incorrect utterances is difficult in real-life. It is not known a priori if a person would speak an incorrectly stressed word, and also the utterance needs to be evaluated by a domain expert; whereas the correct samples can be directly obtained from the linguists, model speakers or trainers. Also samples in this particular domain are not easily available, and are costly to obtain.

2.1. Single Class Model

A model of a correctly uttered word without using any incorrect utterances can be obtained by estimating the multi-dimensional shape of the correct class spanned only by the correct utterances of that word. Instead of directly computing the parametric shape (which is a computationally difficult problem), kernel-based non-parametric technique can be used for the density estimation of the single class. Various techniques exist in the literature which adopt this concept of estimating the density for classification such as one-class SVM (OSVM)[8] and PEBL [9]. In the case of OSVM, the attribute space is transformed into a higher dimensional space and the utterances in the higher dimensional space are enclosed in the smallest possible sphere subject to certain user-defined parameters. These spheres when mapped back to the original space, form the tight cluster boundaries. In PEBL, the single class classifiers are formed with the help of SVM where the margin between the correct class and the uniformly distributed incorrect samples is iteratively refined.

Here, we estimate the non-parametric density of the correct class for each word from the correctly stressed samples. For each test utterance, we obtain the density estimate of the correct class at the point representing the test utterance in the multi-dimensional space. We then make decision based on a bias (a local threshold) automatically decided by the LOO (leave-one-out) cross-validation on the training samples. We describe our method as follows.

Let \( x \) be a test utterance, and \( p(x|C_p) \) be the conditional density of \( x \) belonging to the correct class \( C_p \). In that case, \( x \) is assigned to the correct class if the likelihood ratio \( \frac{p(x|C_p)}{p(x|C_n)} \) is greater than a certain threshold, \( \theta \), where \( C_n \) is the class of incorrect utterances. In case the incorrect class is uniformly distributed, the label of a test utterance can be decided by comparing the local density with a given global threshold \( \theta \). However, with a constant threshold, the boundary of the correct class gets dilated around the dense region as shown in Figure 1, and it does not allow the correct class to be enclosed within a tight boundary. Note that, in the mapping convergence (MC) of PEBL algorithm, an assumption of uniform distribution about the incorrect class has been made and the margin is recursively refined to tightly bound the correct class. However, in PEBL it is assumed that unlabeled samples are easily available which is not true in our application domain. Therefore, recursive margin refinement is not possible in this case.

In order to obtain a tight boundary only with the correct utterances, we compare the conditional density of a test utterance with the minimum conditional density in the neighborhood of the test utterance. In other words, the threshold \( \theta \) is selected based on the nearest neighbors of the test utterance such that

\[
\theta = \min_{x_i \in C_p, x_i \in N_k(x)} \ p(x_i|C_p) + b
\]

(1)

where \( b \) is a bias and \( N_k(x) \) is the set of \( k \) nearest neighbors of the test utterance, \( x \), from the correct class. The test utterance is assigned to the correct class if the conditional density \( p(x|C_p) \) is greater than \( \theta \).

We estimate the conditional density based on the Parzen window estimate [10] with Gaussian kernels, such that at any point \( x \) in the multi-dimensional space,

\[
p(x|C_p) = \sum_{x_i \in C_p} K(x, x_i)
\]

(2)

where \( K(x, x_i) \) is a Gaussian kernel such that

\[
K(x, x_i) = \frac{1}{C} \exp\left(-\frac{D(x, x_i)}{c}\right)
\]

(3)

The parameter \( C \) is a normalizing constant, and \( \sqrt{D(.)} \) is a metric. In case of the Euclidian distance,

\[
D(x, x_i) = \|x - x_i\|^2
\]

(4)

However, we have considered the Mahalanobis distance [10], so \( D(.) \) is given as

\[
D(x, x_i) = (x - x_i)^T \Sigma^{-1} (x - x_i)
\]

(5)

where \( \Sigma^{-1} \) is the inverse covariance matrix of the correct class \( C_p \) estimated from the training utterances. The parameter \( c \) in Equation (3) defines the width of the kernel. The local threshold \( \theta \) can be expressed as

\[
\theta = \min_{x_i \in C_p, x_i \in N_k(x)} \sum_{x_j \in C_p} K(x_i, x_j) + b
\]

(6)

In finding out the nearest neighbors, we consider the metric \( \sqrt{D(.)} \) as in Equation (5) and take only the \( k \) nearest neighbors to define the set \( N_k(x) \). We obtain the bias \( b \) for a correct class by using the leave-one-out (LOO) cross-validation. Usually LOO based techniques are used to estimate parameters where data is available for both the classes. Here we use the samples only from one class. We therefore modify the LOO as follows.

Let the training set of only positive class samples be denoted by \( T \) and \( \epsilon \) be a small constant;
for every sample \( i \in T \),
choose \( b_i \) as the maximum value of bias \( b \) such that sample \( i \) is assigned to the positive class with a model constructed by the training set \( T - \{i\} \)
end

Sort the set \( \{b_i\} \) and let the sorted set be
\( B = \{b(1), b(2), \ldots, b(T)\} \);

Select the bias \( b_T \) for the training set \( T \) as \( b_T = b(k) \)
where \( k = |\epsilon[T]| \).

In other words, we select the bias for a training set consisting on only positive class samples in such a way that the model allows a fraction of the samples to be classified as negative. Experimentally, we observed that a choice of \( \epsilon \) in the range 0.05 – 0.1 provides the best balanced performance. Moreover, we observed that the performance of the classifier is not sensitive to the choice of the constant, \( c \), for a large range.

### 3. Feature Extraction

All the word utterances are time aligned with the corresponding phonetic spelling of the word using the acoustic models and the pronunciation dictionary of a speech recognition system. The acoustic models of this speech recognition system are trained on more than 600 Indian Speakers and the system provided 90% accuracy with speaker adaptation. Since for the purpose of stress analysis, syllable level acoustic features are required, a phone-to-syllable mapping for the word is used to get syllable level alignment of the utterance. Eight different acoustic features, shown in Table 1, have been used in this paper. As described in Section 1, average fundamental frequency, average energy and duration of the syllable are the three basic acoustic features. These three basic features for each of the constituent syllables of the word were normalized with the corresponding average values over the whole word utterance to remove any speaker dependent variations. Fundamental frequency and energy for a syllable were extracted every 10ms from a signal frame of 25ms multiplied with a Hamming window.

Sluijter et al. concluded that the effect of stress on the energy content of the speech signal is more prominent in the higher frequency band as compared to the lower frequency band [11]. Feature 4, i.e., filtered energy, is used to incorporate this phenomenon into the stress models. A high-pass butterworth filter, with a cut-off frequency of 4 kHz, was used to compute this feature for each of the syllables. Feature 5 and 6 are derived from the basic prosodic features, obtained by multiplying average energy and average fundamental frequency of the syllable with the duration of the syllable. Since lexical stress often associated with a rise and then fall in the prosodic features [12], features 7 and 8 capture the change in the average fundamental frequency and in the average energy respectively across two consecutive syllables.

### 4. Experimentation and Results

We performed the experiments on a speech database consisting of 13 different English words spoken by an average of 75 Indian speakers each. These utterances were recorded at a sampling rate of 22kHz in PCM wav format. Two human linguists labeled all these utterances as correct and incorrect considering the lexical stress pattern of the individual words.

We obtained Receiver Operating Characteristics (ROC) for 3-fold cross validation of all the words by varying the bias values from +1 to -1. In Fig. 2, the areas under two ROCs are 80.03% and 77.79% for values of constant, \( c \), 20 and 50 respectively. This shows that the performance of the proposed method is not much dependent upon the value of \( c \).

The performance of the proposed single-class method is shown in Table 2 for 3-fold cross validation. The last two columns of this table show the performances with two different distance metrics namely Mahalanobis Distance and Euclidean Distance, used in the Gaussian Kernel given in (3). The value of the constant \( c \) was chosen to be 20. The bias values were obtained using LOO cross-validation technique with \( \epsilon = 0.1 \). We also compare the proposed method with four different two-class classifiers, viz., Naive Bayes, C4.5 decision tree, k-nearest neighbor and SVM as in Table 2. In this comparison, we use a set of incorrectly stressed words in addition to the correctly stressed words in order to build these four standard classifiers. The performance of various classifiers was measured using WEKA tool [13].

### 5. Discussion

We can see from Table 2 that the standard two class classifiers generate very high false positive rates. This shows that the standard two class classifiers cannot correctly model the incorrectly stressed utterance class due to difficulty in obtaining statistically significant pool of negative samples. Using the same dataset, the proposed single class classifier generates much lower false positive rates at the expense of slight loss in true positive rates.
true positive rate comes with a false positive rate. However, the rate does not increase much with the decrease in the true positive rate of about 90% true positive rate comes with a 40% false positive rate. We are able to correctly classify words, where speaker-dependent variations are relatively lower. This is due to the fact that the shape of the correct class will be more compact for such words. For words, where speaker-dependent variations are higher, the performance of the proposed single class classifier performs comparably to the two-class classifiers in terms of the trade-off between true positive and false positive rates. As we observe from Fig. 2, the ROC curves get flattened at a true positive rate of about 90%. Table 2: Results for different classifiers. NB: Naive Bayes, DT: Decision Tree (C4.5), KNN: K-Nearest Neighbours (K=3), SVM: Support Vector Machine, MD: Proposed Classifier using Mahalanobis Distance, ED: using Euclidean Distance. TP: True Positive Rate, FP: False Positive Rate. W1: ANIMAL, W2: ATTORNEY, W3: AVAILABLE, W4: CONDITION, W5: DETERMINE, W6: EXPENSIVE, W7: INFORMATION, W8: OPPOSITE, W9: REMEMBER, W10: REPRESENTATIVE, W11: THERAPY, W12: TRADITIONAL, W13: TRANSACTION.

<table>
<thead>
<tr>
<th>Word</th>
<th>True Positive Rate</th>
<th>False Positive Rate</th>
<th>TP</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>96.24</td>
<td>96.71</td>
<td>97.18</td>
<td>97.65</td>
</tr>
<tr>
<td>W2</td>
<td>94.17</td>
<td>95.14</td>
<td>97.08</td>
<td>96.60</td>
</tr>
<tr>
<td>W3</td>
<td>95.20</td>
<td>92.75</td>
<td>91.30</td>
<td>95.65</td>
</tr>
<tr>
<td>W4</td>
<td>86.80</td>
<td>82.19</td>
<td>86.30</td>
<td>86.80</td>
</tr>
<tr>
<td>W5</td>
<td>87.93</td>
<td>87.93</td>
<td>91.37</td>
<td>93.10</td>
</tr>
<tr>
<td>W6</td>
<td>94.82</td>
<td>94.82</td>
<td>93.10</td>
<td>94.82</td>
</tr>
<tr>
<td>W7</td>
<td>74.64</td>
<td>80.28</td>
<td>84.50</td>
<td>80.28</td>
</tr>
<tr>
<td>W8</td>
<td>91.37</td>
<td>86.20</td>
<td>87.93</td>
<td>86.20</td>
</tr>
<tr>
<td>W9</td>
<td>96.82</td>
<td>96.82</td>
<td>96.82</td>
<td>96.82</td>
</tr>
<tr>
<td>W10</td>
<td>96.36</td>
<td>94.54</td>
<td>90.90</td>
<td>98.18</td>
</tr>
<tr>
<td>W11</td>
<td>87.62</td>
<td>94.32</td>
<td>95.87</td>
<td>95.87</td>
</tr>
<tr>
<td>W12</td>
<td>96.49</td>
<td>96.49</td>
<td>96.49</td>
<td>96.49</td>
</tr>
<tr>
<td>W13</td>
<td>99.01</td>
<td>99.01</td>
<td>99.01</td>
<td>99.01</td>
</tr>
</tbody>
</table>

Overall Results:

<table>
<thead>
<tr>
<th>Accuracy</th>
<th>TP</th>
<th>FP</th>
</tr>
</thead>
<tbody>
<tr>
<td>92.80</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td>93.66</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>94.60</td>
<td>0.99</td>
<td>0.86</td>
</tr>
<tr>
<td>94.97</td>
<td>0.93</td>
<td>0.33</td>
</tr>
<tr>
<td>90.37</td>
<td>0.80</td>
<td>0.89</td>
</tr>
<tr>
<td>78.62</td>
<td>0.80</td>
<td>0.89</td>
</tr>
</tbody>
</table>

We can see from Table 2 and Fig. 2 that on the average the proposed single class classifier performs comparably to the two-class classifiers. Moreover, due to the limited number of samples, the shape of the region changes whenever we consider a certain percentage as the training set.

The performance of the proposed technique will be even better on words for whom speaker-dependent variations are relatively lower. This is due to the fact that the shape of the correct class will be more compact for such words. For words, where speaker-dependent variations are higher, the performance of the proposed technique can degrade.

6. References


[9] Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang, “Pebi: Web page classification without negative samples,” IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 80–81, 2004.


