Automatic Detection of Abnormal Stress Patterns in Unit Selection Synthesis

Yeon-Jun Kim, Mark C. Beutnagel

AT&T Labs – Research, Florham Park, NJ, USA
{yjkim|mcb}@research.att.com

Abstract

This paper introduces a method to detect lexical stress errors in unit selection synthesis automatically using machine learning algorithms. If unintended stress patterns can be detected following unit selection, based on features available in the unit database, it may be possible to modify the units during waveform synthesis to correct errors and produce an acceptable stress pattern.

In this paper, three machine learning algorithms were trained with acoustic measurements from natural utterances and corresponding stress patterns: CART, SVM and MaxEnt. Our experimental results showed that MaxEnt performs the best (83.3% for 3-syllable words, 88.7% for 4-syllable words correctly classified) in the natural stress pattern classification.

Though classification rates are good, a large number of false alarms are produced. However, there is some indication that signal modifications based on false positives do little harm to the speech output.

Index Terms: speech synthesis, unit selection, lexical stress

1. Introduction

Anyone might have difficulty understanding a foreign speaker’s English. Even if a foreign speaker pronounces a word with the correct sequence of phones, it may still be difficult to recognize. One of the reasons is because a foreign speaker might not be aware of specific stress patterns in English words and put stress on the wrong syllables. In the same manner, a text-to-speech (TTS) synthesis system sometimes produces incorrect stress patterns, which makes a TTS system sound like a foreign speaker. An incorrect stress pattern is not only disruptive by itself, but also degrades intelligibility and naturalness of TTS synthesis.

English has strong-weak alternating rhythm and each word has its own specific stress pattern. While many languages have an entirely predictable stress pattern (e.g. either the first or the last syllable in a multi-syllable word), various stress patterns can be found in words from English and other Germanic languages[1]. Vowel identities can also be changed depending on the existence of stress, i.e. unstressed vowels in American English are often reduced to schwa, /ax/. Therefore, it is critical to predict and synthesize correct stress patterns in American English synthesis.

Previous work related to stress in speech synthesis has been concentrated in stress assignment to predict the correct stress patterns from given text [2] [3] [4]. Traditional parametric speech synthesis produces a stream of parameters from rules or from statistics based on a training corpus, and is guaranteed to produce the predicted stress patterns.

Unit selection synthesis, which can produce higher quality by concatenating natural speech segments with less signal processing, brings an unexpected complication. Acoustic units, chosen from various locations throughout the recorded corpus and concatenated in novel combinations, may convey the wrong lexical stress pattern even though the correct pattern was predicted by the TTS front-end in Fig. 1.

Ironically, as segmental problems due to allophones and unexpected segments in the unit database were reduced [5] [6], lexical stress gets more attention and listeners find unnatural stress more disruptive.

The speech produced by unit selection synthesis sometimes violates the listener’s expectations. Even if each unit’s stress, rhythm, etc. is appropriate for its local context, juxtaposing them with units from other contexts can interfere with the perceived stress. For example, a vowel with secondary stress from a louder word may overwhelm a primary stressed vowel from a softer word in different context.

The challenge is to mitigate such problems while still preserving the natural variations in recorded speech available to unit selection synthesis, i.e. not by strictly enforcing the predicted prosodic parameters (pitch, amplitude and duration) across all selected units.

10.21437/Interspeech.2010-77
Unit selection is typically implemented as a Viterbi search, and all decisions are strictly local. Cost functions evaluate the current candidate unit in relation to predicted features and adjacent units, which does not allow for any higher-level view.

In this paper, we introduce a post-processing module to detect abnormal stress patterns and remedy them in unit selection synthesis shown in Figure 1. First, we define possible stress patterns for English words and measure acoustic parameters of units in a recorded corpus. Human perception related to acoustic parameters is modeled using several machine learning algorithms.

2. Lexical Stress Patterns

A correctly produced sentence in English comes from the successful imposition of stresses at two levels: the correct syllable in a multi-syllabic word, *lexical stress*, and the correct placement within the sentence, *sentential stress* [7]. Determination of sentential stress is still an open problem because so many factors influence the placement of stress, including type of sentence, emotional status, context, and intentions, etc.

On the other hand, prediction of lexical stress is well-established and is the first step in prosody realization. However, mistakes in synthesizing the correct stress patterns for isolated words can still occur in unit selection synthesis. In this paper, we narrow our focus to the correlation between lexical stress patterns and acoustic realization in natural utterances.

Since the stress can be assigned to any syllable in a multisyllabic word in English, there are a number of stress patterns possible. In previous work by Clopper [8], she differentiated stress patterns solely by the position of the primary stress in a word. In addition to primary stress, our TTS front-end module also predicts secondary stress (from dictionary word lists and stress assignment rules [9] [10]). This allows a more natural stress pattern, but also allows for a wider range of errors.

Table 1: Lexical stress patterns in 3-/4-syllable words in the target speaker’s database. The primary stress is written in bold and upper cases, and the secondary stress in upper cases only in the examples.

<table>
<thead>
<tr>
<th>Stress pattern</th>
<th>No. of instances</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-syllable words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>010</td>
<td>3032</td>
<td>dePARTment</td>
</tr>
<tr>
<td>100</td>
<td>3489</td>
<td>Clitizen</td>
</tr>
<tr>
<td>102</td>
<td>2988</td>
<td>JACKsonVILLE</td>
</tr>
<tr>
<td>120</td>
<td>895</td>
<td>WESTMINster</td>
</tr>
<tr>
<td>201</td>
<td>515</td>
<td>ILLINOIS</td>
</tr>
<tr>
<td>210</td>
<td>1099</td>
<td>MONTana</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stress pattern</th>
<th>No. of instances</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-syllable words</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0100</td>
<td>1015</td>
<td>aMERICAN</td>
</tr>
<tr>
<td>0102</td>
<td>74</td>
<td>reLATIONSHIP</td>
</tr>
<tr>
<td>1000</td>
<td>71</td>
<td>TEMperature</td>
</tr>
<tr>
<td>1002</td>
<td>32</td>
<td>LibertyTOWN</td>
</tr>
<tr>
<td>1020</td>
<td>361</td>
<td>OperAtor</td>
</tr>
<tr>
<td>1200</td>
<td>29</td>
<td>PAINSTakingly</td>
</tr>
<tr>
<td>2010</td>
<td>1953</td>
<td>PENNsylvania</td>
</tr>
<tr>
<td>2100</td>
<td>283</td>
<td>MONGolia</td>
</tr>
</tbody>
</table>

We tagged our TTS voice database with the lexical stress patterns predicted by our TTS front-end. Table 1 shows the stress patterns of 3- /4-syllable words found in 20 hours of one female TTS voice database which includes many street and city names. Stress patterns consist of primary (‘1’), secondary (‘2’) stressed, or unstressed (‘0’) syllables. The following stress patterns are used as target classes for machine learning algorithms.

Even though any stress value can be assigned to any syllable in an English word, stress patterns in our recorded database are not evenly distributed, as shown in Table 1. Specially, we don’t have any 4-syllable word which has the primary stress in the final syllable. Another interesting result is that there are more 4-syllable words which have the primary stress in the second or the third syllable than ones which have the primary stress in the first syllable.

3. Acoustic Measures for Stress

It is widely-agreed that a stressed syllable is uttered with a greater amount of energy than an unstressed syllable [1]. The greater energy is realized in various acoustic forms in speech; increase in pitch (fundamental frequency), in amplitude or in length duration.

To learn how acoustic parameters are used to deliver lexical stress patterns by humans, pitch, amplitude and duration were measured quantitatively from a female TTS voice talent’s natural utterances. Prior to acoustic measurement, audio files in the unit database were energy-normalized by sentence in order to reduce unwanted variations from a series of recording sessions. Even though the TTS voice talent was asked to utter sentences in the consistent manner, some amount of variation cannot be avoided. Meanwhile, pitches and durations in speech were kept in the natural forms without modification.

Pitch and amplitude were both measured from speech files at 10 ms intervals and then averaged at the nucleus of the syllable. For amplitude measurement, log value were used rather than raw value. Durations of phone segments were computed from automatically segmented phone boundaries [11]. Another indication of stress is the rise in pitch that usually occurs caused by additional muscular activity. We modeled such phenomena with the slope of pitch (∆f/0), which was also computed in every half-phone.

In addition to features mentioned above, we included normalized values of the parameters which depend on phone identity: duration and amplitude. Some vowel sounds are known to have more acoustic energy than others due to the different degrees of mouth opening. Diphthongs tend to be longer than other vowels, for example, /ay/ in ‘time’ is typically longer than /aɪ/ in ‘Tom’ in comparable contexts. By introducing Z-score at the n-th syllable, \( Z(n) \), in Eq. (1), we can use stylized stress patterns independent of the phone’s intrinsic variations.

\[
Z(n) = \frac{(X_i(n) - \mu_i)}{\delta_i}
\]

where \( \mu_i \) and \( \delta_i \) are the mean and the standard deviation of one feature (e.g. duration) across all segments i of a given phone type in the target speaker’s database.

It is well known that the amplitude and the duration of a stressed syllable are increased compared to nearby unstressed vowels. However, as shown in Figure 2, it is difficult to draw a clear line between the stressed and the unstressed in actual data. Each plot shows the distributions of energy (a) or duration (b) at 10 ms intervals and then averaged at the nucleus of the syllable. It is not always the case that the stressed syllable have higher energy than the unstressed syllable.

The average amplitude and duration in stressed syllables are slightly larger than those at unstressed syllables, but it is not a distinct bimodal distribution. We believe this is due to variation
4. Stress Pattern Classification using Machine Learning Algorithms

Our goal in this work is to model human perception concerning lexical stress patterns and make use of it to detect abnormal synthesized stress patterns. The perceptual-level data as heard by listeners is very expensive to collect. Instead of approaching human stress perception directly, we assume that how humans produce stress should be similar to how humans perceive stress and model the correlation between stress patterns and acoustic measurements.

To model human perception, we employed machine learning algorithms. All algorithms were trained with the given acoustic parameters from each syllable in a word and the corresponding stress pattern as a target class.

4.1. Machine Learning Algorithms

The machine learning algorithms used in this work came from WEKA which is a collection of machine learning algorithms for data mining tasks [12]. It also provides a graphical user interface so that it is convenient to develop and test learning algorithms.

CART Classification and regression tree, decides the target class with the given input variables. Quinlan’s C4.5 decision tree implementation was used.

AdaBoost+CART Adaptive Boosting, calls a weak classifier repeatedly and updates the importance of training examples to focus the misclassified instances. In this work, it is used in conjunction with CART algorithm.

SVM Support Vector Machine, maps the examples to the separate categories so that they are divided by a clear gap as wide as possible [13]. Implements John Platt’s sequential minimal optimization algorithm for training a support vector classifier.

MaxEnt Maximum Entropy, building and using a multinomial logistic regression model with a ridge estimator. Like many other regression models, it makes use of several predictor variables that may be either numerical or categorical.

4.2. Classification of Natural Utterance

All the machine learning algorithms were trained by supervised learning methods with acoustic measurements input parameters and stress patterns as the target class. Then, they were tested by 10-fold validation.

Table 2: Experimental results of natural stress patterns classification using machine learning algorithms

<table>
<thead>
<tr>
<th>Machine Learning Algorithm</th>
<th>Correctly Classified (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>74.8</td>
</tr>
<tr>
<td>AdaBoost+CART</td>
<td>81.3</td>
</tr>
<tr>
<td>SVM</td>
<td>81.6</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>83.3</td>
</tr>
</tbody>
</table>

In both 3- and 4-syllable word stress pattern classifications, MaxEnt outperforms the other algorithms, and correctly classified 83.3% stress patterns for 3-syllable words and 88.7% for 4-syllable words. All methods classified 4-syllable stress patterns correctly more often than 3-syllable patterns, but this may be due to the concentration of 4-syllable words in two categories (‘0100’ and ‘2010’). Distribution is more uniform in the 3-syllable words.

Table 3: Confusion matrix in stress pattern classification using MaxEnt for (a) 3-syllable and (b) 4-syllable words

(a)

(b)
Table 3 shows the confusion matrix when MaxEnt algorithm was used to classify stress patterns. From the experiment result, secondary stress brought more confusions, 9% of '100' patterns were misclassified into '102', and vice versa.

In stress pattern classification for 4-syllable words, the stress pattern '2010' far outnumbers other patterns. This resulted in the misclassification of a large fraction of '1020' stress patterns as '2010' shown at Table 3 (b).

4.3. Classification of Synthesized Utterance

When we apply models trained with natural utterances to classify the stress pattern of a synthesized word, models’ performance were degraded. They produced a huge number of false negatives which sounds reasonable to a native listener, but disagrees with the given lexical stress patterns.

In our experiment, we played misclassified synthesis words to a native listener and asked him to judge whether the misclassified pattern is truly off from the stress pattern that he expected, without knowing its confidence score. Figure 3 shows that more words truly violate human perception (true negative) when their confidence scores are higher. We claim that the confidence score from the classification algorithm is relevant to the listener’s perception. The confidence score will be more effective when the number of false alarms should be reduced.

![Figure 3: Number of detected instances](image)

5. Conclusions

Several machine learning techniques were used to model human perception of stress patterns, aiming to detect abnormal stress patterns in unit selection synthesis and remedy them using signal processing. Input data included raw and normalized feature values from a large database of high-quality recorded speech. The MaxEnt models produced the best results in classification of natural stress patterns.

Sample words were synthesized with AT&T Natural Voices™, a unit selection TTS system. Training data was derived from the same voice database used in synthesis. Synthesized words were classified by the models to detect synthetic words that violate their expected stress patterns.

One purpose of this work is to detect incorrect stress patterns after acoustic units are selected but before waveform synthesis. At that point, signal processing can be directed to modify the synthesis and produce an improved stress pattern (compared to the default speech output).

High numbers of false alarms were noted in classification of synthesized stress patterns. Future work will attempt to reduce confusion by taking phrase position and other contextual factors into account. Preliminary work indicates that unnecessary signal modifications caused by false alarms are not especially harmful to the speech output.

6. References