ISCA Archive
http://www.isca-speech.org/archive

Speech and Language Technology in Education (SLaTE2011)
Venice, Italy
August 24-26, 2011

Automatic Assessment of American English Lexical Stress using Machine Learning Algorithms

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 assess lexical stress patterns in American English words automatically using machine learning algorithms, which could be used on the computer assisted language learning (CALL) system. We aim to model human production concerning lexical stress patterns by training stress patterns in a native speaker's utterances and making use of it to detect erroneous stress patterns from a trainee.

In this paper, all the possible lexical stress patterns in 3- and 4-syllable American English words are presented and four machine learning algorithms, *CART*, *AdaBoost+CART*, *SVM* and *MaxEnt*, are trained with acoustic measurements from a native speaker's utterances and corresponding stress patterns. Our experimental results show that *MaxEnt* correctly classified the best, 83.3% stress patterns of 3-syllable words and 88.7% of 4-syllable words.

Index Terms: automatic assessment, lexical stress, machine learning

1. Introduction

Learning a new language is a challenging task which involves several components of a language, such as vocabulary, grammar, and intonation, etc. This paper focuses on lexical stress patterns. Even if a foreign speaker pronounces a word with the correct sequence of phones, it may still be difficult to understand if the syllable stress is incorrect. Non-native speakers might not be aware of specific stress patterns in English words and put stress on the wrong syllables, especially those speakers whose native language doesn't have lexical stress.

English has strong-weak alternating rhythm and each word has its own specific stress pattern. While many other 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, an incorrect stress pattern is not only disruptive by itself, but can also degrade the intelligibility of of the speech.

Tutoring lexical stress is a very important area in English education and requires native-speakers' monitoring, therefore it should be introduced in computer assisted language learning (CALL). Thanks to automatic speech recognition (ASR) technology, English word pronunciation assessment has become popular in *CALL* [2] [3]. On the other hand, automatic stress assessment has received less attention from researchers.

A previous work related to automatic word assessment in *CALL* [4] enforces copying a native-speaker's intonation by comparing both pitch and energy based on *dynamic time warping* (DTW). Instead, our work introduces *machine learning* algorithms, allowing a greater degree of prosodic variation by modeling human production. A similar approach is proposed in [5], which trains SVMs to distinguish phoneme pairs, then applies them to detect pronunciation errors in L2 learners' speech.

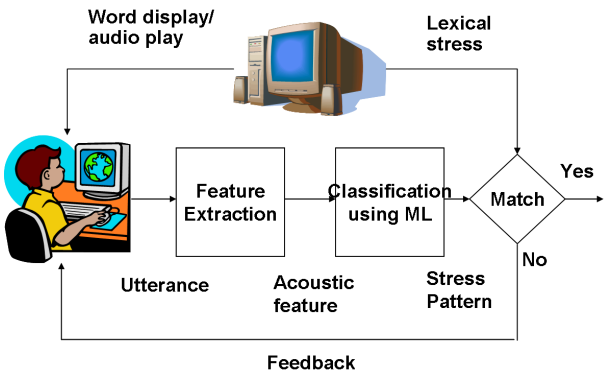


Figure 1: Stress Assessment for CALL

In this work, we propose a new stress assessment method using machine learning algorithms, aiming to allow natural intonation variations while monitoring incorrect lexical stress patterns as shown in Fig 1. First, we describe possible stress patterns for American English words and measure acoustic parameters of units in a recorded corpus. Human production related to acoustic parameters is modeled using several machine learning algorithms, *CART*, *AdaBoost+CART*, *SVM* and *MaxEnt*.

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* [6].

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 [7]. 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 multi-syllabic word in American English, there are a number of stress patterns possible to appear in native-speakers' utterances. Table 1 shows the stress patterns of 3- / 4-syllable words and the numbers of each patterns found in 20 hours of one female speaker's

Table 1: *Lexical stress patterns in 3- / 4-syllable words in the target speaker's database (around 15 hours of speech). The primary stress is written in bold and upper cases, and the secondary stress in upper cases only in the examples.*

	Stress pattern	No. of instances	Example
3-syllable words	010	3032	de**PART**ment
	100	3489	**CI**tizen
	102	2988	**JACK**sonVILLE
	120	895	**WEST**MINster
	201	515	ILLi**NOIS**
	210	1099	MON**TA**na
4-syllable words	0100	1015	a**ME**rican
	0102	74	re**LA**tionSHIP
	1000	71	**TEM**perature
	1002	32	**LI**bertyTOWN
	1020	361	**O**perAtor
	1200	29	**PAINS**TAkingly
	2010	1953	PENNsyl**VA**nia
	2100	283	MON**GO**lia

recording which includes many American street and city names. Stress patterns consist of primary ('1'), secondary ('2') stressed, or unstressed ('0') syllables. This stress patterns are trained as target classes by machine learning algorithms.

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, another level of stress, *secondary stress* is found in American English dictionary and stress assignment rules [9] [10], which is also considered in this work. Considering both primary and secondary stress provides a more accurate stress pattern, but also could bring a wider range of variations.

Even though any stress value can be assigned to any syllable in a English word, stress patterns in our recording are not evenly distributed, as shown in Table 1. For example, 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 Lexical Stress

It is well known that a stressed syllable is uttered with a greater amount of energy than an unstressed syllable [1]. The perceived syllable 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 native speaker's utterances. Speech signals are sampled at 16 bit, 16 kHz linear PCM and segmented by sentence. Prior to acoustic measurement, audio files used in this work were energy-normalized by sentence in order to reduce unwanted variations from a series of recording sessions. Even though the native speaker was asked to utter sentences in the consistent manner, some amount of variation cannot be avoided.

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 ($\Delta f0$), 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 /aa/ in 'Tom' in comparable contexts. By introducing *Z-score* at the n-th syllable, $Z_i(n)$, in Eq. (1), we can use stylized stress patterns independent of the phone's intrinsic variations [12].

$$Z_i(n) = \frac{(X_i(n) - \mu_i)}{\delta_i} \quad (1)$$

where μ_i and δ_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.

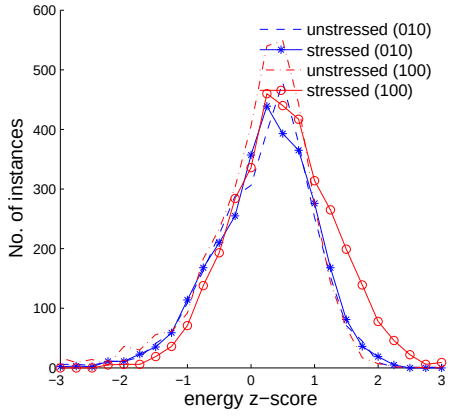
With the features described above an attribute selection test, *CFS* (Correlation-based Feature Subset Selection) [13] in *WEKA*, was performed [14]. This method provides high scores to the subsets that include features that are highly correlated to the class attribute, but have low correlation to each other. As shown in Table 2, the amplitude and duration of syllables are more highly correlated to the lexical stress pattern class than other features.

Table 2: Result of attribute subset evaluation for lexical stress pattern classification in the case of 4-syllable words

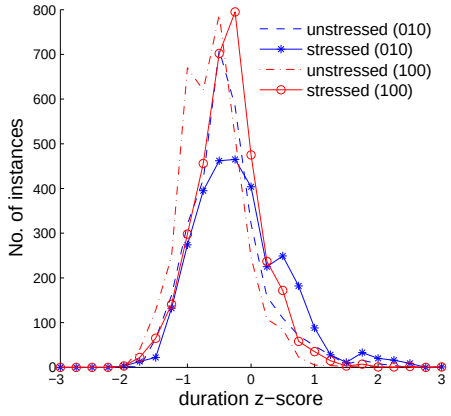
```
Attribute Subset Evaluator
  (supervised, Class:  37 class):
CFS Subset Evaluator
Including locally predictive attributes

Selected attributes:
  2,6,8,9,12,15,17,18,24,26,27 :  11
        eng[1]
        dur_z[1]
        eng[2]
        dur[2]
        dur_z[2]
        dur[3]
        eng_z[3]
        dur_z[3]
        dur_z[4]
        f0_norm[2]
        f0_norm[3]
```

Though the amplitude and the duration of a syllable are most influencing features for realization of stressed syllables shown in Table 2, however, it is still difficult to draw a clear line between the stressed and the unstressed in actual data, as shown in Figure 2. Each plot shows the distributions of energy (a) or duration (b) at both the stressed syllable and the unstressed syllable for each of two stress patterns of 3-syllable words ('100' in red and '010' in blue).

The average amplitude and duration in stressed syllables are slightly larger than those at unstressed syllables, but it is not



(a)



(b)

Figure 2: *Distribution of Z-score energy (a) and duration (b) of the stressed syllable and the unstressed syllable in the stress pattern, '010' and '100'*

a distinct bimodal distribution. We believe that lexical stresses are not related any acoustic attribute alone, but related with the combination of several acoustic attributes.

In this study, machine learning algorithms are introduced to distinguish lexical stress patterns in the acoustic feature space. Additional duration normalization was done within words considering phrase position and speaking rate. For example, the final syllable in 3-syllable words tends to be longer regardless of stress, so compensation of this intrinsic bias is helpful.

4. Classification using Machine Learning

Our goal in this work is to model human production concerning lexical stress patterns and make use of it to detect erroneous stress patterns. To model human production, 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.

The machine learning algorithms used in this work came from *WEKA* which is a collection of machine learning algorithms for data mining tasks [14]. 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 [15]. Implements John Palate'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.

5. Experimental Results

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

	Machine Learning Algorithm	Correctly Classified (%)
3-syllable words	CART	74.8
	AdaBoost+CART	81.3
	SVM	81.6
	MaxEnt	**83.3**
4-syllable words	CART	77.8
	AdaBoost+CART	83.6
	SVM	85.3
	MaxEnt	**88.7**

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

	classified as					
	010	100	102	120	201	210
010	2867	41	9	45	6	64
100	33	3063	269	72	8	44
102	13	322	2539	28	68	18
120	56	202	42	457	2	136
201	3	66	180	6	252	8
210	72	78	18	90	4	837

(a)

	classified as							
	0100	0102	1000	1002	1020	1200	2010	2100
0100	967	6	1	1	2	4	5	29
0102	20	45	0	1	2	0	0	6
1000	1	0	47	5	6	3	5	4
1002	2	2	2	19	2	2	0	3
1020	1	0	3	2	214	0	136	5
1200	2	0	8	2	1	13	0	3
2010	8	1	1	0	67	0	1870	6
2100	41	4	1	4	4	5	12	212

(b)

Table 4 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 in Table 4 (b).

In the preliminary experiment, we randomly pick 100 utterances from 1996 CSR Hub-4 corpus, extract the acoustic features except f0 considering speaker variability to see how foreign speakers actually generate lexical stress patterns. The experimental result shows that there are some differences between native and foreign speakers' lexical stress realization as shown in Table 5. The further study should be performed with the actual L2 learners' speech since the acoustic conditions of foreign speakers in the broadcast news audio may not be equivalent to the ones of natives.

Table 5: Experimental results of stress patterns classification from broadcast news audio

	Number of Syllable in a word	Correctly Classified (%)
Native, Planned (F0)	3-syllable	71.6
	4-syllable	75.3
Foreign, Spontaneous (F5)	3-syllable	49.8
	4-syllable	56.1

6. Conclusions

Several machine learning techniques were used to model human production of stress patterns, aiming to detect erroneous stress patterns from a trainee. 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 a native speaker's stress patterns. In the preliminary experiment using broadcast news audio, it is observed that there are differences between native and foreign speakers' lexical stress realization.

For further work, this work to detect erroneous lexical stress patterns could be extended to correct a speaker's mistake using signal processing technologies. Signal processing enables to embed natives' intonation on a trainee's utterance by modifying pitch, energy, and duration of signal. Instead hearing other native speakers' utterances, hearing his/her own utterances corrected would be more effective for computer assisted language learning.

7. References

[1] P. Ladefoged, *A Course in Phonetics*. Harcourt Brace Jovanovich College Publishers, 1993.

[2] M. Peabody and S. Seneff, "A Simple Normalization Scheme for Non-native Vowel Assessment," in *Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan, 2010.

[3] M. Suzuki, Y. Qiao, N. Mnematsu, and K. Hirose, "Pronunciation proficiency estimation based on multilayer regression analysis using speaker-independent structural features," in *Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan, 2010.

[4] J. P. Arias, N. B. Yoma, and H. Vivanco, "Automatic intonation assessment for computer aided language learning," *Speech Commmunication*, vol. 52, pp. 254–267, 2010.

[5] S. Y. Yoon, M. Hasegawa-Johnson, and R. Sproat, "Landmark-based automated pronunciation error detection," in *Proc. Interspeech*, Tokyo, Japan, 2010.

[6] A. Cutler, *Errors of Stress and Intonation*. Academic Press, 1980, ch. 4, pp. 67–80.

[7] T. Visceglia, C. yu Tseng, Z. yu Su, and C.-F. Huang, "Interaction of Lexical and Sentence Prosody in Taiwan L2 English," in *Satellite Workshop on Second Language Studies: Acquisition, Learning, Education and Technology*, Tokyo, Japan, 2010.

[8] C. G. Clopper, "Frequency of stress patterns in english: A computational analysis," in *IULC Working Papers Online*, 2002.

[9] C. Coker, K. Church, and M. Liberman, "Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis," in *Proc. of ESCA Workshop on Speech Synthesis*. Autrans, France, 1990, pp. 83–86.

[10] K. Church, "Stress Assignment in Letter-to-Sound Rules for Speech Synthesis," in *ACL*, 1985, pp. 246–253.

[11] Y.-J. Kim and A. Conkie, "Automatic Segmentation combining an HMM-based Approach and Spectral Boundary Correction," in *Proc. ICSLP*. Denver, USA, 2002.

[12] A. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, Atlanta, 1996, pp. 373–376.

[13] M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning," Ph.D. dissertation, University of Waikato, 1998.

[14] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann; 2 edition, 2005.

[15] J. A. Bilmes and P. Haffner, "Machine Learning in Speech and Language Processing," in *Proc. ICASSP*. Philadelphia, PA, 2005.