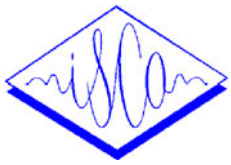


7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

A Study of Lexical 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

In this paper we describe a method that detects and remedies 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. Note that the TTS system being studied typically does no prosody modification on selected units, unlike most concatenative TTS systems.

We trained several machine learning algorithms using acoustic measurements from natural utterances and corresponding stress patterns: *CART*, *Adaboost+CART*, *SVM* and *MaxEnt*. Our experimental results showed that *MaxEnt* achieves the highest accuracy on natural stress pattern classification (83.3% for 3-syllable words, 88.7% for 4-syllable words correctly classified). Though precision rates are good in the classification of natural stress patterns, a large number of false alarms are produced in the classification of synthesized stress patterns when models trained with natural utterances were applied.

Results from a preference test showed that signal modifications based on false positives do little harm to the speech output, but also that listeners don't find much difference between the raw TTS outputs and the post-processed ones.

Index Terms: speech synthesis, unit selection, lexical stress

1. Introduction

The synthesis of human-like speech has been the goal of many researchers over decades. The introduction of a unit selection approach in speech synthesis brought dramatic improvement in segmental quality [1]. The new approach was assumed to produce human-like speech when it has enough speech audio. Soon another problem arose: how to arrange natural speech segments in order to sound natural as well as smooth. In this paper, we describe a method capable of synthesizing human-like rhythms, i.e. lexical stress patterns, in American English synthesis, especially using a unit selection approach.

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 the 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 [2]. 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 focused on stress assignment to predict the correct stress patterns from input text [3] [4] [5]. 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.

By contrast, 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 wrong lexical stress pattern even though the correct pattern was predicted by the TTS front-end.

Ironically, as segmental problems due to allophones and unexpected segments in the unit database are reduced [6] [7], 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.

In this study, we aim to intervene in unit selection synthesis only when an undesirable sequence of units was chosen to maximize natural prosodic variations, instead of molding the selected units into the stylized stress pattern. The TTS system being studied typically does no prosody modification on selected units, unlike most concatenative TTS systems. The intention is to use prosody modification only when it is necessary to avoid perceptual errors.

First, we define possible stress patterns for American English words and measure acoustic parameters from units in a recorded corpus. Human perception related to lexical stress patterns and acoustic parameters is modeled using several machine learning algorithms. When abnormal stress patterns in the selected unit sequence are detected by the trained algorithm, the duration and amplitude of the units are modified to match the desired measures, aiming to mimic human stress patterns without synthesis quality degradation.

2. System Overview

A unit selection TTS system typically consists of three modules: front-end, unit selection, and signal processing. First, the front-end module produces phonetic and prosodic specifications of units to be concatenated from the given text via text normalization, linguistic analysis and prosody generation, etc.

In the unit selection module, the specifications predicted from the front-end module are compared with the ones in the unit database. Unit selection is generally implemented as a Viterbi search to minimize not only the target cost but also perceptible acoustic mismatch between pairs of units to be concatenated. Unfortunately, all these decisions are strictly local, which does not allow for any higher-level view. It sometimes leads to an unexpected sequence of units which irritate the human listener. In this paper, we introduce a post-processing module to detect the undesirable stress patterns chosen by the unit selection module and remedy them in the signal processing stage as shown in Figure 1.

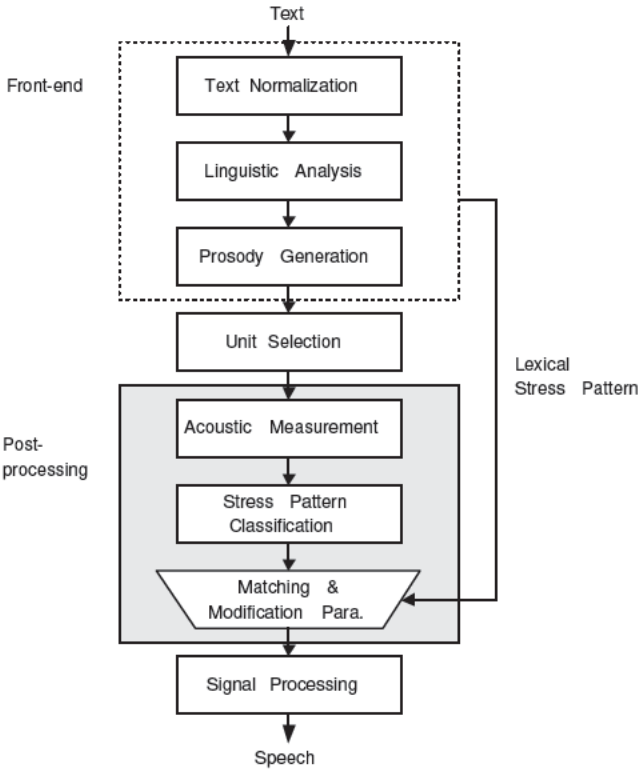


Figure 1: Flow of the proposed post-processing in the unit selection TTS system

3. Stress Pattern Classification

3.1. 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* [8]. 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 multi-syllabic word in English, there are a number of possible stress patterns. In previous work by Clopper [9], 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 [10] [11]). This allows for 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. In the examples, the primary stress is written in bold and upper case, and the secondary stress in upper case only.

	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

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 a TTS voice database produced from 20 hours of speech from one female voice talent which includes many street and city names. Stress patterns consist of primary stressed ('1'), secondary ('2') stressed, or unstressed ('0') syllables. These stress patterns are used as target classes for our machine learning algorithms.

Even though any stress value can be assigned to any syllable in a 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 that has the primary stress in the final syllable. Also, note that there are more 4-syllable words that have the primary stress in the second or the third syllable than ones that have the primary stress in the first syllable.

3.2. Acoustic Measures for Stress

It is widely agreed that a stressed syllable is uttered with a greater amount of energy than an unstressed syllable [2]. The greater energy is realized in various acoustic forms in speech; increase in *pitch* (fundamental frequency), in *amplitude* or in *duration*.

To learn how acoustic parameters are used to deliver lexical stress patterns in human speech, *pitch*, *amplitude* and *duration* were measured *quantitatively* from a female TTS voice talent's

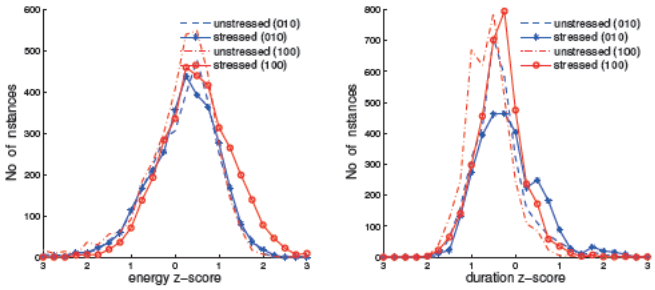
natural utterances. Prior to acoustic measurement, audio files in the unit database were energy-normalized by sentence in order to reduce unwanted variations from a series of recording sessions. Even though the TTS voice talent was asked to utter sentences in a consistent manner, some amount of variation cannot be avoided. Meanwhile, pitches and durations in the audio files were not modified.

Pitch and amplitude were both measured from the audio files at 10 ms intervals and then averaged at the nucleus of the syllable. For amplitude measurement, log values were used rather than raw values. Durations of phone segments were computed from automatically identified phone boundaries [12]. Another indication of stress is the rise in pitch that usually occurs, caused by additional muscular activity. We modeled this phenomena by measure pitch *slope* ($\Delta f0$), which was also computed in every half-phone.

In addition to the 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 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 measuring *Z-score* at each syllable, $Z_i(n)$, in Eq. (1), we can use stylized stress patterns independent of the phone's intrinsic variations.

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

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.



(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'

It is well known that the amplitude and 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 stressed and unstressed in actual data. 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 of stressed syllables are slightly larger than those of unstressed syllables, but it is not a distinct bimodal distribution. We believe this is due to variation in context and in syllable position within phrases. In this study, additional 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 for this intrinsic bias is helpful. In further study, we plan to use either isolated words or words from similar contexts.

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]
```

3.3. Model using Machine Learning Algorithms

Our goal in this work is to model human perception of lexical stress patterns and use these models to detect abnormal synthesized stress patterns. 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 using the acoustic parameters from each syllable in a word as features, and using 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 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.

3.4. Classification of Natural Utterance

We ran a 10-fold cross-validation experiment using the 20 hours of data in our TTS voice database.

Table 3: Experimental results of natural stress patterns classification

	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**

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)

For both 3- and 4-syllable word stress pattern classifications, *MaxEnt* outperformed the other algorithms. It correctly classified 83.3% of stress patterns for 3-syllable words and 88.7% of stress patterns 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 of 4-syllable words in two categories ('0100' and '2010'). The distribution of stress pattern is more uniform for 3-syllable words.

Table 4 shows the confusion matrix when the *MaxEnt* algorithm was used to classify stress patterns. Secondary stress gave rise to most of the classification errors, 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).

3.5. Classification of Synthesized Utterance

When we apply models trained with natural utterances to predict the stress pattern of a synthesized word, the models' performance dropped. They produced a huge number of *false negatives* which sound reasonable to a native listener, but disagree with the given lexical stress patterns.

In our experiment, we played misclassified synthesized 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 the stress pattern's 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* is reduced.

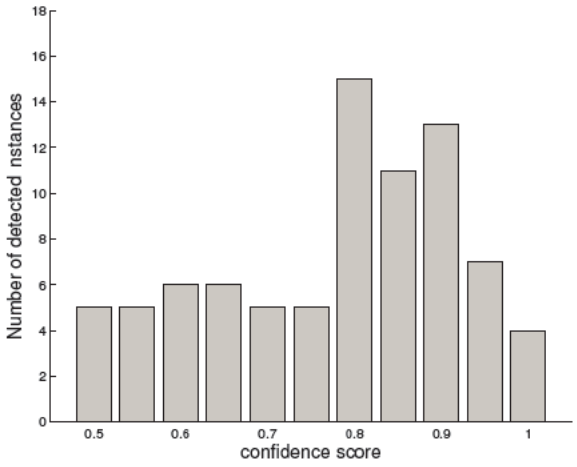


Figure 3: Number of abnormal stress patterns perceived by a native listener in synthesized 4-syllable words

4. Lexical Stress Pattern Synthesis

4.1. Prosody Modification

Once the trained model finds the mismatch between the desired stress pattern and the classified pattern, the waveform signal of the selected units are modified to realize human-like prosody.

In an earlier study, it has seen that listeners significantly prefer synthesis quality when units from a large speech inventory were selected and concatenated with the minimal signal modification, than when the pitch and duration of these units

were modified with either *TD-PSOLA* or *HNM* (harmonic noise model) techniques to match the prosody predicted by the TTS system [16]. Our interpretation of the previous study is that drastic signal modification could disrupt human perception only even if it matches human prosody patterns.

In this study, only the duration and amplitude of the selected units, which are shown to be highly correlates with stress pattern class in Table. 2, are modified to match the target stress pattern using *WSOLA* technique. It is also believed that the duration and amplitude modification may introduce minimal quality degradation only.

After the decision which synthesis unit should be stretched or amplified to mimic the natural stress pattern, the target duration and amplitude, $\overline{X}_i(n)$, are computed as in Eq. (2), which is the inverse form of Eq. (1).

$$\overline{X}_i(n) = \mu_i + \delta_i' \cdot Z_i(n) \quad (2)$$

where δ_i' is the standard deviation of duration or amplitude, which will be positive for the unit to be stretched and negative for the one to be shrunk.

4.2. Listening Test

A listening test was conducted to evaluate whether the proposed signal modification can alter synthesized speech effectively to match the desired stress patterns and whether the signal modification brings any degradation noticeable by listeners.

20 pairs of the test stimuli were chosen mainly from rare proper names which cause many unit concatenations in the unit selection TTS system and have potential problems in stress pattern realization. Meanwhile, common names, such as 'California' or 'Arizona', have dozens of instances in the unit selection database, and the unit selection TTS system is less likely to produce strange stress patterns for those words.

To avoid listeners being distracted by other factors, we played only isolated pairs of words were played to listeners. All test files were renamed through symbolic links to prevent identification of test conditions. The listening test was web-based and interactive, and 21 AT&T employees participated in it.

Table 5: Experimental result of the paired preference test between baseline TTS synthesis and synthesis modified using lexical stress pattern prediction

No preference	baseline TTS	Modified
63.6%	17.6%	18.8%

As the experimental result in Table 5 show, many listeners couldn't recognize the difference between the raw TTS synthesis and the signal modified to match the target stress pattern. We ran a *t-test*, which showed that there is insufficient evidence that either the baseline TTS synthesis or the modified synthesis is preferred by the listeners ($t = 0.33 < T_{0.05} = 1.645$).

Listeners seemed not to be bothered by signal modification, but also they didn't perceive significant improvement from duration and amplitude modification to match the target stress pattern. It may be difficult for listeners to judge the quality of rare words even though some listeners preferred the modified versions to match the target stress pattern. If they heard common words or words in a particular context, awkward lexical stress patterns would be more obvious.

Future work will study the degree of duration and amplitude perceptible to listeners and that could be used to implement lexical stress patterns effectively.

5. Conclusions

In this paper, we experimented with using several machine learning techniques 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. *MaxEnt* models produced the best results in classification of natural stress patterns.

Sample words were synthesized with 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 words that violated 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 a target stress pattern.

High numbers of false alarms were noted in classification of synthesized stress patterns. The result of the listening test showed that unnecessary signal modifications caused by false alarms are not especially harmful to the speech output. Although our listening tests results are inconclusive, we are encouraged by the facts that (1) listeners were not bothered by stress pattern modifications and (2) our MaxEnt classifier was able to predict target stress patterns with high accuracy. In future work, we plan to do additional evaluation using more common words, and words in larger realization contexts.

6. Acknowledgments

The authors would like to thank Alistair D. Conkie for his help and for kindly providing his web-based evaluation package.

7. References

[1] Alan W Black and Andrew Hunt, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of ICASSP*. Atlanta, USA, 1996, pp. 373–376.

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

[3] A. I. C. Monaghan, "Rhythm and stress-shift in speech synthesis," *Computer Speech and Language*, vol. 4, pp. 71–78, 1990.

[4] Alan W Black, Kevin Lenzo, and Vincent Pagel, "Issus in building general letter-to-sound rules," in *The 3rd ESCA Workshop on Speech Synthesis*, 1998, pp. 77–80.

[5] Q. Dou, S. Bergsma, S. Jiampojamarn, and G. Kondrak, "A Ranking Approach to Stress Prediction for Letter-to-Phoneme Conversion," in *Proc. of the 47th Annual Meeting of the ACL and the 4th IJCNLP*, 2009, pp. 118–126.

[6] Yeon-Jun Kim, Ann K. Syrdal, Alistair D. Conkie, and Mark C. Beutnagel, "Phonetic Enrichment Labeling in Unit Selection Synthesis," in *Proc. Interspeech*. Pittsburgh, USA, 2006.

[7] Alistair Conkie, Ann Syrdal, Yeon-Jun Kim, and Mark Beutnagel, "Improving Preselection in Unit Selection

Synthesis," in *Proc. Interspeech*. Brisbane, Australia, 2008, pp. 585–588.

[8] Anne Cutler, *Errors of Stress and Intonation*, chapter 4, pp. 67–80, Academic Press, 1980.

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

[10] 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.

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

[12] Yeon-Jun Kim and Alistair Conkie, "Automatic Segmentation combining an HMM-based Approach and Spectral Boundary Correction," in *Proc. ICSLP*. Denver, USA, 2002.

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

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

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

[16] Matthias Jilka, Ann K. Syrdal, Alistair Conkie, and David A. Kapilow, "Effects of TTS quality of methods of realizing natural prosodic variations," in *Proc. ICPhS*, 2003.