VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

School of Mathematical and Computing Sciences
Computer Science

Detecting Stress in Spoken English using Decision Trees and Support Vector Machines

Huayang Xie, Peter Andreae, Mengjie Zhang, Paul Warren

Technical Report CS-TR-03/14
October 2003

School of Mathematical and Computing Sciences
Victoria University
PO Box 600, Wellington
New Zealand

Tel: +64 4 463 5341
Fax: +64 4 463 5045
Email: Tech.Reports@mcs.vuw.ac.nz
http://www.mcs.vuw.ac.nz/research

VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

School of Mathematical and Computing Sciences
Computer Science

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341, Fax: +64 4 463 5045
Email: Tech.Reports@mcs.vuw.ac.nz
http://www.mcs.vuw.ac.nz/research

Detecting Stress in Spoken English using Decision Trees and Support Vector Machines

Huayang Xie, Peter Andreae, Mengjie Zhang, Paul Warren

Technical Report CS-TR-03/14
October 2003

Abstract

This paper describes an approach to the detection of stress in spoken New Zealand English. After identifying the vowel segments of the speech signal, the approach extracts two different sets of features — prosodic features and vowel quality features — from the vowel segments. These features are then normalised and scaled to obtain speaker independent feature values that can be used to classify each vowel segment as stressed or unstressed. We used Decision Trees (C4.5) and Support Vector Machines (LIBSVM) to learn stress-detecting classifiers with various combinations of the features. The approach was evaluated on 60 adult female utterances with 703 vowels and a maximum accuracy of 84.72% was achieved. The results showed that a combination of features derived from duration and amplitude achieved the best performance but the vowel quality features also achieved quite reasonable results.

Keywords Machine learning, feature extraction, speech recognition, stress detection, decision tree, support vector machine.

Author Information

All the authors are at Victoria University of Wellington, New Zealand. Huayang Xie is a Masters student in Computer Science, Peter Andreae and Mengjie Zhang are academic staff members in computer science in the School of Mathematical and Computing Sciences, and Paul Warren is an academic staff member in the School of Linguistics and Applied Language Study.

Detecting Stress in Spoken English using Decision Trees and Support Vector Machines

Huayang Xie* **Peter Andreae*** **Mengjie Zhang*** **Paul Warren+**

*School of Mathematical and Computing Sciences
+School of Linguistics and Applied Language Studies
Victoria University of Wellington,
P. O. Box 600, Wellington, New Zealand,
Email: {`Huayang.Xie, Peter.Andreae, Mengjie.Zhang,Paul.Warren`}`@vuw.ac.nz`

Abstract

This paper describes an approach to the detection of stress in spoken New Zealand English. After identifying the vowel segments of the speech signal, the approach extracts two different sets of features — prosodic features and vowel quality features — from the vowel segments. These features are then normalised and scaled to obtain speaker independent feature values that can be used to classify each vowel segment as stressed or unstressed. We used Decision Trees (C4.5) and Support Vector Machines (LIBSVM) to learn stress-detecting classifiers with various combinations of the features. The approach was evaluated on 60 adult female utterances with 703 vowels and a maximum accuracy of 84.72% was achieved. The results showed that a combination of features derived from duration and amplitude achieved the best performance but the vowel quality features also achieved quite reasonable results.

Keywords: Machine learning, feature extraction, speech recognition, stress detection, decision tree, support vector machine.

1 Introduction

As English becomes more and more important as a communication tool for people from all countries, there is an ever increasing demand for good quality teaching of English as a Second Language (ESL). New Zealand is one of the destinations for foreign students wanting to learn English from English speaking teachers, and for political reasons is often perceived as a desirable destination. Learning English well requires lots of practice and a great deal of individualised feedback to identify and correct errors. Providing this individualised feedback from ESL teachers is very expensive, and the shortage of ESL teachers means that there is increasing demand for computer software that can provide useful individualised feedback to students on all aspects of their English.

The ESL Software Tools research group at Victoria University of Wellington is developing a software system to provide individualised feedback to ESL students on prosodic aspects of their speech production, focusing particularly on the stress and rhythm of the speech. The overall design of the system involves a pedagogic component that engages in simple dialogues with the student, and a speech analyser that analyses the student's speech, identifying the stress pattern in the speech and comparing it with a target pattern in order to provide useful feedback on stress and rhythm errors.

The first stage of the speech analyser performs phoneme level speech recognition on the student's speech to identify the start and end times of all the phonetic elements of the speech. The second stage analyses the vowel elements of the speech to identify which elements would be perceived as stressed. The final stage matches the pattern of stresses to the target pattern to identify any stress or rhythm errors in the student's speech.

The first stage of the speech analyser was described in (Xie, Andreae, Zhang & Warren 2004). The current paper focuses on the second stage, which involves first selecting, extracting and normalising a set of features of each vowel segment identified in the first stage, and then using these features to classify the vowels as stressed or unstressed. We use machine learning methods, namely decision trees and support vector machines, to construct the classifier. It is not clear from the literature exactly which features are most effective for the automatic determination of stress; an important goal of our research is therefore to identify which features are most helpful in detecting stress in complete sentences. One of the features that we investigate is vowel quality. Mea-

This research is supported by the New Economy Research Fund, New Zealand.

suring vowel quality is not as straightforward as the other features we investigate, and the paper describes a novel method for extracting vowel quality features.

The remainder of the paper is organised as follows: section 2 presents some essential background; section 3 describes feature extraction and normalisation; section 4 describes our experimental design and methods, and section 5 presents results. Section 6 gives conclusions and future work.

2 Background

2.1 Stress

Stress is a form of prominence in spoken language. Usually, stress is seen as a property of a syllable or of the vowel nucleus of that syllable. In English, we can consider two types of stress. Lexical stress refers to the relative prominences of syllables in individual words. Normally, words of more than one syllable will have one syllable that carries primary lexical stress; other syllables may carry secondary lexical stress, and the remaining syllables will be unstressed. Rhythmic stress can be used to refer to the relative prominences of syllables in longer stretches of speech than the isolated word. When words are used in utterances, their prominences may be altered to reflect the rhythmic (as well as semantic) structure of the utterance.

Although much of the work in the literature on detecting stress has focused on lexical stress, the practical application that is the ultimate goal of our work requires that we focus on rhythmic stress. Therefore the experiments and results below deal with recognising stress in whole sentences.

2.2 Prosodic features

There are a number of prosodic (sometimes referred to as 'suprasegmental') features that relate to stress. Thus the perception of a syllable as stressed or unstressed may depend on its relative duration, its amplitude and its pitch. The first of these is straightforwardly how long the syllable lasts; the second relates to the perceived loudness of the syllable, and is a measure of its energy; pitch is the perceptual correlate of the fundamental frequency of the sound signal (i.e. the rate of vibration of the vocal folds during voiced segments).

All of these prosodic features may vary for reasons other than signalling stress. For instance, there will be intrinsic differences in the duration and amplitude of different speech sounds, and different speakers will, because of their different physiologies, produce different ranges of amplitude and pitch. In addition, prosodic features indicate other important aspects of the meaning of utterances — for instance pitch variation is part of the intonation or melody of speech, which can signal differences in the meanings intended by the speaker or their emotional state.

2.3 Vowel Quality

A further correlate of stress is the quality of the vowel in a syllable. Vowel quality is determined by the configuration of the tongue, jaw, and lips (Ladefoged 1967, Bernthal & Bankson 1988, Ladefoged & Maddieson 1990, Pennington 1996). Since there is some flexibility in the formation of a vowel, there will in fact be a range of articulator parameter values that correspond to the same vowel. Of particular relevance to the detection of stress is the range from the "full" form of a vowel to the "reduced" form of the same vowel. Reduced vowel forms tend to have a more central articulation (i.e. the tongue is nearer its "rest" position) (Cruttenden 1997, Ladefoged 1993). In English, the central vowel /ə/ or "schwa" is always reduced. In NZ English, the short "I" or /ɪ/ tends to also have a schwa-like quality, and vocalised forms of /l/ can also be considered reduced. When unstressed, other vowels may also be pronounced in reduced forms, making them more schwa-like.

In general, the vowels of unstressed syllables tend to be reduced, and the vowels of stressed syllables tend to be full. However, the correlation is not perfect, since although reduced vowels only occur in unstressed syllables, all English vowels — including full vowels — can occur in unstressed syllables, so that vowel quality is not a completely reliable indicator of stress (Ladefoged 1993).

2.4 Related Work

There have been a number of reports on stress detection. Most reports focused on lexical stress detection based on isolated words, but a few have addressed rhythmic stress detection in complete utterances.

Lieberman (1960) used duration, energy and pitch to identify lexical stress in bisyllabic noun-verb stress pairs (e.g. PREsent vs preSENT). These features were extracted from the hand-labelled syllables. The database consisted of isolated words pronounced individually by 16 native speakers and was used for both training and testing. A Decision Tree approach was used to build the stress detector and 99.2% accuracy was achieved.

Aull and Zue (1985) used duration, pitch, energy, and changes in spectral envelope (a measure of vowel quality) to identify lexical stress in polysyllabic words. The features were extracted from the sonorant portion of automatically labelled syllables. The pitch parameter used was the maximum value in the syllable. Energy was normalised using the logarithm of the average energy value. The database consisted of isolated words extracted from continuous speech pronounced by 11 speakers. A template-based algorithm was used to build the stress detector and 87% accuracy was achieved.

Ferij et al. (1990) used pitch, energy and spectral envelope to identify lexical stress in bi-syllabic words pairs. The first and second derivatives of pitch and the first derivative of energy were also used. The spectral envelope was represented by 4 LPC features. These 9 features were extracted from the hand-labelled syllables at 10ms intervals. The database consisted of isolated words extracted from continuous speech pronounced by three male speakers. Hidden Markov Models (HMMs) were used to build the stress classifier and a overall accuracy of 94% was achieved. Vowel quality, especially the distinction between reduced and full unstressed syllables was suggested as a direction for future work.

Ying et al. (1996) used energy and duration to identify stress in bi-syllabic words pairs. The energy and duration features were extracted from automatically labelled syllables and were normalised using several methods. The database consisted of isolated words extracted from continuous speech pronounced by five speakers. A Bayesian classifier assuming multivariate Gaussian distributions was adopted and the highest performance was 97.7% accuracy.

A few studies have investigated stress detection in longer utterances. Waibel (1986) used amplitude, duration, pitch, and spectral change to identify rhythmically stressed syllables. The features were extracted from automatically labelled syllables. Peak-to-peak amplitude was normalised over the sonorant portion of the syllable. Duration was calculated as the interval between the onsets of the nuclei of adjacent syllables. The pitch parameter used was the maximum value of each syllable nucleus. Spectral change was normalised over the sonorant portion of the syllable. The database consisted of 50 sentences read by 10 speakers. A Bayesian classifier assuming multivariate Gaussian distributions was adopted and 85.6% accuracy was reached.

Jenkin and Scordilis (1996) used duration, energy, amplitude, and pitch to classify vowels into three levels of stress — primary, secondary, and unstressed. The features were extracted from hand-labelled vowel segments. Peak-to-peak amplitude, energy and pitch were normalised over the vowel segments of the syllable. In addition syllable duration, vowel duration and the maximum pitch in the vowel were used without normalisation. The database consisted of 288 utterances (8 sentences spoken by 12 female and 24 male speakers) from dialect 1 of the TIMIT speech database. Neural networks, Markov chains, and rule-based approaches were adopted. The best overall performances ranged from 81% to 84% by using Neural networks. Rule-based systems performed more poorly, with scores from 67% to 75%.

van Kuijk and Boves (1999) used duration, energy, and spectral tilt to identify rhythmically stressed vowels in Dutch — a language with similar stress patterns to those of English. The features were extracted from manually checked automatically labelled vowel segments. Duration was normalised using the average phoneme duration in the utterance, to reduce speaking rate effects. Also a complex duration normalisation method introduced in (Wightman 1992) was adopted. Energy was normalised using several procedures, such as the comparison of the energy of a vowel to its left neighbour and its right neighbour, to the average energy of all vowels to its left and to the average energy of all vowels in the utterance. Spectral tilt was calculated using spectral energy in various frequency sub-bands. The database consisted of 5000 training utterances and 5000 test utterances from the Dutch POLYPHONE corpus. A simple Bayesian classifier was adopted, on the argument that the features can be jointly modelled by a N-dimensional normal distribution. The best overall performance achieved was 68%.

The summary above shows that stress classification is most accurate for a limited task, such as identifying the stressed syllable in isolated bi- or polysyllabic words, with performance levels noticeably lower in the few studies using longer utterances. The studies cited do not seem to indicate that a particular classification procedure is any more successful than any other.

3 Feature Extraction and Normalisation

To detect the stressed syllables in an utterance, our system first performs forced alignment speech recognition using an (HMM) recogniser (Xie et al. 2004). In forced alignment, the target sentence is known, so the recogniser only needs to perform a mapping between segments in the utterance and phonemes in the target sentence.[1] Using the phonemic transcription of the utterance, the system identifies those segments that correspond to each vowel in the target. Although stress is generally argued to be a property of the syllables in an utterance rather than of just the vowels, the prosodic and vowel quality features we use in our study are largely carried by the vowel as the nucleus of the syllable.

[1] There are generally multiple possible pronunciations of the target sentence, so that the recogniser still has to identify which possible pronunciation is present in the utterance, but this is still very much more constrained than recognising an utterance when the target sentence is not known.

The next task for our system is to determine which of the vowels are stressed. Each vowel is analysed in several different ways to extract a set of features that can be passed to the stress classifier. Since duration, amplitude, pitch and vowel quality are the parameters that have been shown to cue the perception of stress differences in English, the features we need to extract are related to these parameters.

For each of the prosodic parameters (duration, amplitude, pitch), there are many alternative measurements that can be extracted, and also many ways of normalising the features in order to reduce variation due to differences between speakers, recording situations or utterance contexts, etc. The subsections below describe these alternatives. Vowel quality features are more difficult to extract. We describe a method that uses the HMM phoneme models to re-recognise the segments of the utterance and extract measures of vowel quality.

3.1 Duration Features

Vowel durations can be directly measured from the output of the forced alignment recogniser since the recogniser identifies the start and end points of the vowels. The measurements are not completely reliable since it is hard for the recogniser to precisely determine the transition point between two phonemes that flow smoothly into each other. Furthermore, some short vowels may be inaccurately reported if they are shorter than the minimum number of frames specified for a phoneme in the system.

The absolute value of the duration of a vowel segment is influenced by many factors other than stress, such as the intrinsic durational properties of the vowel, the speech rate of the speaker, and local fluctuations in speech rate within the utterance. Therefore the absolute duration of the vowel segment is not a useful feature. What is required is a normalised duration that measures how much longer or shorter this vowel segment is than that vowel would "normally" be spoken by an "average" speaker. To reduce the impact of these contextual properties, we applied three different levels of normalisation to the raw duration values.

The first level normalisation reduces the effect of speech rate variation between speakers. To normalise, we need to compare the length of an utterance to the "expected" length of that utterance. To compute the latter, we first use the training speech data set to calculate the average duration of each of the 20 vowel phonemes of NZ English. We then compute the expected utterance length by summing the average durations of the phonemes in the utterance, and the actual utterance length by summing the actual durations of the vowel segments in the utterance. We can then normalise the durations of each vowel segment by multiplying by the expected utterance length divided by the actual utterance length.

The second level normalisation removes effects of variation in the durations of the different vowel phonemes. Each phoneme has an intrinsic duration — long vowels and diphthongs normally have longer durations than short vowels. There are several possible ways to normalise for intrinsic vowel duration. One method is to normalise the vowel segment duration by the average duration for that vowel phoneme, as measured in the training data set. Another method is to cluster the 20 vowel phonemes into three categories (short vowel, long vowel and diphthong) and normalise vowel segment durations by the average duration of all vowels in the relevant category. We consider both methods.

The third level normalisation removes the effect of the variation in speech rate at different parts of a single utterance. To remove this influence, the result of the second level normalisation is normalised by a weighted average duration of the immediately surrounding vowel segments.

Based on the three levels of normalisation, we computed five duration features for each vowel segment:

- *Utterance normalised duration: the absolute duration normalised by the length of the utterance;*
- Phoneme normalised duration*: the duration normalised by the length of the utterance and the average duration of the phoneme;*
- Category normalised duration*: the duration normalised by the length of the utterance and the average duration of the vowel category;*
- Phoneme neighbourhood normalised duration*: the phoneme normalised duration further normalised by the durations of neighbouring vowels;*
- Category neighbourhood normalised duration*: the vowel category normalised duration further normalised by the durations of neighbouring vowels.*

3.2 Amplitude Features

The amplitude of a vowel segment can be measured from the speech signal, but since amplitude changes during the vowel, there are a number of possible measurements that could be made — maximum amplitude, initial amplitude, change in amplitude, *etc.* A measure commonly understood to be a close correlate to the perception of amplitude differences between vowels is the root mean square (RMS) of the amplitude values across the entire vowel. This is the measure chosen as the basis of our amplitude features. As with the duration features, amplitude is influenced by a variety of factors other than stress, including speaker differences and differences in recording conditions as well as changes in amplitude across the utterance. We therefore need to normalise measured amplitude to reduce variability introduced by these effects. We apply two levels of normalisation to obtain two amplitude features.

Our first level normalisation of amplitude takes into account global influences such as speaker differences and recording situation, and normalises the RMS amplitude of each vowel segment against the overall RMS amplitude of the entire utterance.

Our second level normalisation considers local effects at different parts of the utterance and normalises the vowel amplitude against a weighted average amplitude of the immediately surrounding vowel segments.

3.3 Pitch Features

The primary acoustic correlate of the pitch of a vowel segment is the fundamental frequency, F_0, of the speech signal. Like amplitude, pitch can vary over the course of the vowel segment and is influenced by a variety of different factors, including the basic pitch of the speaker's voice. To reduce the effects of speaker differences, we normalise the pitch measurement of a vowel segment by the average pitch of the entire utterance.

The change in pitch over the vowel segment is at least as important as the pitch level of the vowel, but it is not clear exactly which properties of pitch are most significant for determining stress. Therefore, we extracted 10 different pitch features, including not only the average normalised mean pitch value of a vowel segment, but other features intended to capture changes in pitch. The 10 pitch features of a vowel segment are calculated as follows:

- *Normalised mean pitch*: the mean pitch value of the vowel normalised by the mean pitch of the entire utterance.
- *Normalised pitch value at the start point*: the pitch value at the start point of the vowel divided by the mean pitch of the utterance.
- *Normalised pitch value at the end point*: the pitch value at the end point of the vowel divided by the mean pitch of the utterance.
- *Normalised maximum pitch value*: the maximum pitch value of the vowel divided by the mean pitch of the utterance.
- *Normalised minimum pitch value*: the minimum pitch value of the vowel divided by the mean pitch of the utterance.
- *Relative pitch difference*: the difference between the normalised maximum and minimum pitch values. A negative value indicates a falling pitch and a positive value indicates a rising pitch.
- *Absolute difference*: the magnitude of the *Relative difference*, which is always positive.
- *Pitch trend*: The sign of the *Relative difference* — 1 if the pitch "rises" over the vowel segment, -1 if it "falls", and 0 if it is "flat".
- *Boundary Problem*: a boolean attribute, true iff the pitch value at either the start point or the end point of the vowel segment cannot be detected.
- *Length Problem*: a boolean attribute. This attribute is true iff the vowel segment is too short to compute meaningful minimum, maximum, or difference values.

3.4 Vowel Quality Features

Unstressed syllables are associated with centralised vowels, particularly the /ə/ vowel which is the primary reduced vowel. In NZ English, /ɪ/ is also pronounced very centrally and often acts as a reduced vowel. Full vowels tend to be more peripheral, and are associated with stressed syllables. However, a vowel with the quality of a full vowel that is pronounced more centrally than usual may act as a reduced vowel, even if it is not central enough to be transcribed as /ə/.

To determine whether a vowel segment represents a reduced vowel, we need to recognise the intended vowel phoneme, and also to determine whether it is pronounced more centrally than the norm for that vowel. Since our speech recogniser uses forced alignment, it only identifies the segments of the utterance that match each expected vowel best and does not identify how the speaker pronounced the vowel. For the prosodic features above, this is all that is needed, and using a full recogniser on the entire sentence would reduce the accuracy of the recognition. However, for measuring vowel quality, we need to know what vowel the speaker actually said, and how they pronounced it.

To determine the actual vowel quality of the vowels, we apply a very constrained form of full recognition to each of the vowel segments previously identified by forced alignment, and use the probability scores of the individual HMM phoneme models to compute several features that indicate whether the vowel is reduced or not. The algorithm is illustrated in figure 1 and outlined below.

Step 1 Extract vowel segments from utterance using forced alignment. Label each segment with the expected vowel phoneme label, based on the target sentence and the pronunciation dictionary.

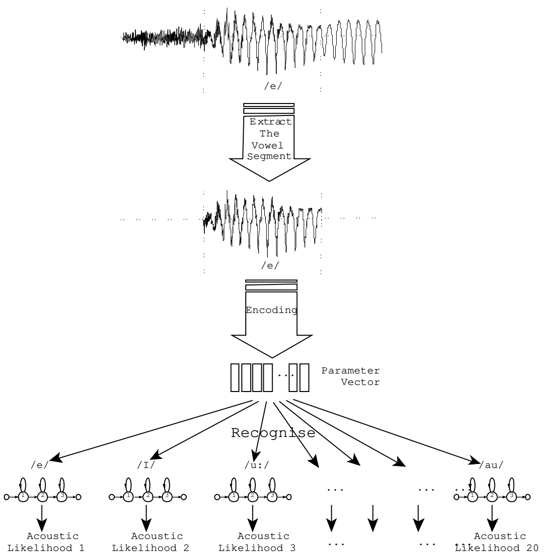

Figure 1: Vowel Quality Features Processing

Step 2 Encode each vowel into a sequence of acoustic parameter vectors, using a 15ms Hamming window with a step size (frame period) of 11ms. These parameters consist of 12 MFCC features and the 0'th cepstral coefficient with their first and second order derivatives. The values of these parameters were suggested by our previous study (Xie et al. 2004).

Step 3 Feed the parameter vector sequence into the 20 pre-trained HMM vowel recognisers to obtain 20 normalised acoustic likelihood scores. Each score is the geometric mean of the acoustic likelihoods of all frames in the segment, as computed by the HMM recogniser. The scores are likelihoods that reflect how well the segment matches the vowel type of the HMM.

Step 4 Find the score of the expected vowel type S_e, the maximum score of any full vowel phoneme S_f and the maximum score of any reduced vowel phoneme S_r from the above 20 scores.

Step 5 We then compare the scores of the best matching full vowel and the best matching reduced vowel to the score of the expected vowel. We compute four features, two of which measure the difference between the likelihoods, and two measure the ratio of the likelihoods. In each case, we take logarithms to reduce the spread of values:

$$R_d = \begin{cases} -log(S_r - S_e) & \text{if } S_e < S_r \\ 0 & \text{if } S_e = S_r \\ log(S_e - S_r) & \text{if } S_e > S_r \end{cases} \quad (1)$$

$$F_d = \begin{cases} -\log(S_f - S_e) & \text{if } S_e < S_f \\ 0 & \text{if } S_e = S_f \\ \log(S_e - S_f) & \text{if } S_e > S_f \end{cases} \quad (2)$$

$$R_r = \log(S_e/S_r) = \log S_e - \log S_r \quad (3)$$

$$F_r = \log(S_e/S_f) = \log S_e - \log S_f \quad (4)$$

Both difference and ratio measures have advantages and disadvantages. We explore which of the two approaches is better for the detection of rhythmic stress.

Step 6 We also compute a boolean vowel quality feature, T, to deal with cases where the vowel segment is so short that F or R can not be calculated. The attribute is true iff the vowel segment is less than 33ms (the minimum segment duration allowed by our HMM). If this attribute is true, we set F and R to 0.

4 Experiment Design and Methods

4.1 Experimental Goals

The goals of our experiments are to investigate whether it is feasible to build an effective automated stress detector for English utterances and to evaluate the different sets of features that we have extracted. Our approach is to use two examples of two standard machine learning tools — a decision tree constructor (C4.5) and a support vector machine (LIBSVM)(Chang & Lin 2003) — to construct stress detectors using these features, and to measure the performance of the resulting stress detectors. One reason for considering a decision tree constructor is that they can generate explicit rules that might help us identify which features were most significant for the stress detector.

4.2 Data Set

The experiments used a speech data set collected by the School of Linguistics and Applied Language Studies at Victoria University. This data set contains 60 utterances of ten distinct English sentences produced by six adult female NZ speakers, as part of the New Zealand Spoken English Database (www.vuw.ac.nz/lals/nzsed). The utterances were hand labelled at the phoneme level, and each vowel was labelled as *stressed* or *unstressed*. There are 703 vowels in the utterances; 340 are stressed and 363 unstressed. The prosodic and vowel quality features were extracted for each of these vowels.

4.3 Performance Evaluation

The task of our stress-detector is to classify vowels as stressed or unstressed. Neither stress category is weighted over the other, and so we use classification accuracy to measure the performance of each classifier.

Since the data set is relatively small, we applied the 10-fold cross validation method for training and testing the stress detectors. In addition, we repeated this training and testing process ten times. Our results below report the average results over the ten repetitions.

4.4 Experiment Design

As discussed earlier, we computed several sets of features and selected two learning algorithms for the construction of stress detectors. To build the stress detector, the training and test data needed to be as accurate as possible. We therefore used hand labelling to determine the vowel segments and phoneme labels in the input speech data. We designed three experiments to investigate a sequence of research questions.

To explore which subset of prosodic features is most useful for learning stress detectors for our data, the first experiment uses the two learning algorithms in conjunction with all seven different combinations of the prosodic features (D, A, P, D+A, D+P, A+P, and D+A+P, where D, A, and P are the sets of duration, amplitude, and pitch features, respectively).

To assess the contribution of vowel quality features to stress detection, the second experiment uses the two learning algorithms in conjunction with six different combinations of the vowel quality features ($F_d + T$, $R_d + T$, $R_d + F_d + T$, $F_r + T$, $R_r + T$, and $R_r + F_r + T$).

The third experiment investigates whether combining the prosodic features and the vowel quality features improves performance.

In these experiments, we also investigate whether scaling the feature values to the range [-1 ... 1] improves performance.

For the SVM, we used a RBF kernel and a C parameter of 1.0.

5 Results

5.1 Experiment 1: Prosodic Features

Features	C4.5		LIBSVM	
	Unscaled	Scaled	Unscaled	Scaled
D	80.66	80.22	81.00	82.55
A	68.18	68.26	70.18	69.08
P	55.12	56.00	57.82	58.45
$D + A$	81.34	81.06	83.88	84.72
$D + P$	80.84	80.10	79.27	81.55
$A + P$	66.96	66.36	70.00	70.28
$D + A + P$	80.40	80.58	79.72	83.23

Table 1: Results for prosodic features.

The results for the prosodic features in the first experiment are shown in table 1. Overall, the best results obtained by the LIBSVM are almost always better than those obtained by C4.5 for all feature combinations. Scaled data led to better performance than unscaled data for LIBSVM in most cases, but this is not true for C4.5. For both LIBSVM and C4.5 , the combination of duration and amplitude features (D + A), produced the best results: 84.72% and 81.34%, respectively. Adding the pitch features to this subset did not improve performance in any case. These results suggest that the subset of features (D+A) is the best combination for our data set. While both decision trees and support vector machines are supposed to be able to deal with the redundant features, neither of them performed well at ignoring the less useful features in this experiment.

5.2 Experiment 2: Vowel Quality Features

Features	C4.5		LIBSVM	
	Unscaled	Scaled	Unscaled	Scaled
$F_d + T$	65.50	66.17	66.57	68.27
$R_d + T$	80.74	80.87	81.36	81.51
$F_d + R_d + T$	79.88	79.73	79.12	81.51
$F_r + T$	67.80	68.38	62.56	63.44
$R_r + T$	82.14	82.15	82.50	78.37
$F_r + R_r + T$	80.64	80.48	81.29	78.37

Table 2: Results for Vowel Quality Features.

The second experiment investigated the performance of the vowel quality features. The results are shown in table 2. The vowel quality features alone achieved results that were very comparable to the performance of the prosodic features. The best result was 82.50%, which was achieved by LIBSVM using the features ($R_r + T$). This result was only 2.22% lower than the best result achieved by the seven prosodic features. In addition, the following points can be noted.

- The reduced vowel quality features R_d and R_r are more reliable than full vowel quality features F_d and F_r.
- In most cases, using the likelihood ratios is better than using the likelihood differences.
- For LIBSVM, if using likelihood differences for vowel quality features, scaling is recommended; if likelihood ratios are used, scaling is not needed.
- For C4.5, in most cases, scaling produced slightly better results than non-scaling, regardless of whether differences or ratios were used, but the difference in performance between scaling and non-scaling was always very small.

5.3 Experiment 3: All Features

The third experiment was performed using the combination of all the prosodic features (C) and the vowel quality features using either the difference ($V_d = F_d + R_d + T$) or the ratio measure of vowel quality

Features	C4.5		LIBSVM	
	Unscaled	Scaled	Unscaled	Scaled
$C + V_d$	80.26	81.38	81.04	82.23
$C + V_r$	80.30	80.42	81.14	82.40

Table 3: Results for Prosodic and Vowel Quality features.

($V_r = F_r + R_r + T$) As can be seen from table 3, combining all features from the two sets did not improve the best performance on our data set over using either prosodic or vowel quality features alone. However, the result did demonstrate that the SVM achieved better performance than the C4.5 on the data set, suggesting that SVM is more suitable for a relatively large data set with all numeric data.

For all the three experiments, C4.5 produced rules that were far more complex and much harder to interpret than expected. Given that most of the features are numeric, this should not have surprised us.

6 Conclusions

The goal of this paper was to develop an approach to rhythmic stress detection in NZ English. Vowel segments were identified from speech data and a range of prosodic and vowel quality features were extracted from the vowel segments. The vowel quality features were calculated using individual HMM vowel models. Different combinations of these features were normalised and/or scaled and then fed into the C4.5 and LIBSVM algorithms to learn the stress detectors. The approach was tested on 60 adult female utterances containing 703 vowels. The results show that a combination of duration and amplitude features achieved the best performance (84.72%) and that the vowel quality features also achieved good results (82.50%). On this data set, the support vector machine achieved better results than decision trees. It is interesting to note that the prosodic features and the vowel quality features are each equally effective at detecting stress, but that their combination did not appear to enhance performance.

We were surprised that the pitch features did not turn out to be particularly useful. We suspect that the algorithm used to identify the pitch may be error-prone, and that the pitch normalisation was inadequate. We will examine alternative pitch detection algorithms and better normalisation methods.

While the maximum accuracy is not good enough yet to be very useful for a commercial system, these results are quite comparable to (even slightly better than) similar systems in this area (Jenkin & Scordilis 1996, van Kuijk & Boves 1999), reflecting the fact that rhythmic stress detection from continuous speech remains a difficult problem in the current state of the art of speech recognition. We will need to explore a variety of techniques to improve the performance. We will need to explore a greater variety of machine learning tools — there are other decision tree constructors and other varieties of SVM, as well as other techniques such as neural nets and genetic programming (some preliminary experiments with genetic programming are very promising). We will also need to examine different combinations of the features and vowel quality and explore better normalisation methods.

Acknowledgement

We would like to thank other members in our group particularly David Crabbe, Irina Elgort and Mike Doig for a number of useful discussions.

References

Aull, A. M. & Zue, V. W. (1985), 'Lexical stress determination and its application to speech recognition', in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1549–1552.

Bernthal, J. E. & Bankson, N. W. (1988), *Articulation and phonological disorders*, Prentice Hall, New Jersey.

Chang, C.-C. & Lin, C.-J. (2003), 'Libsvm: a library for support vector machines', http://www.csie.ntu.edu.tw/˜cjlin/papers/libsvm.pdf.

Cortes, C. & Vapnik, V. (1995), 'Support-vector network', *Machine Learning* **20**, 273–297.

Cruttenden, A. (1997), *Intonation*, Second edition, Cambridge University Press, Cambridge.

Freij, G., Fallside, F., Hoequist, C. & Nolan, F. (1990), 'Lexical stress estimation and phonological knowledge', *Computer Speech and Language* **4**(1), 1–15.

Jenkin, K. L. & Scordilis, M. S. (1996), 'Development and comparison of three syllable stress classifiers', in *Proceedings of the International Conference on Spoken Language Processing*, Philadelphia, USA, pp. 733–736.

Ladefoged, P. (1967), *Three Areas of experimental phonetics*, Oxford University Press, London.

Ladefoged, P. (1993), *A Course in Phonetics*, Third edition, Harcourt Brace Jovanovich, New York.

Ladefoged, P. & Maddieson, I. (1990), 'Vowels of the world's languages', *Journal of Phonetics*, **18**, 93–122.

Lieberman, P. (1960), 'Some acoustic correlates of word stress in American English', *Journal of the Acoustical Society of America*, **32**, 451–454.

Mateescu, D. (2003), 'English phonetics and phonological theory', http://www.unibuc.ro/eBooks/filologie/mateescu.

Pennington, M. C. (1996), *Phonology in English language teaching: An international approach*, Longman, London.

van Kuijk, D. & Boves, L. (1999), 'Acoustic characteristics of lexical stress in continuous speech', in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **3**, Munich, Germany, pp. 1655–1658.

Waibel, A. (1986), 'Recognition of lexical stress in a continuous speech system – a pattern recognition approach', in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, pp. 2287–2290.

Wightman, C. W. (1992), *Automatic detection of prosodic constituents for parsing*, PhD thesis, Boston University.

Xie, H., Andreae, P., Zhang, M. & Warren, P. (2004), 'Learning models for English speech recognition', *Proceedings of the 27th Australasian Computer Science Conference*, Dunedin, New Zealand.

Ying, G. S., Jamieson, L. H., Chen, R., Michell, C. D. & Liu, H. (1996), 'Lexical stress detection on stress-minimal word pairs', *Proceedings of the 1996 International Conference on Spoken Language Processing* pp. 1612–1615.