AUTOMATIC SYLLABLE STRESS CLASSIFICATION METHODS

K.L. Jenkin and M.S. Scordilis*

Telstra Research Laboratories

*Wire Communications Laboratory
University of Patras, Greece

ABSTRACT - Several approaches to the task of automatically classifying syllable stress in continuous speech are detailed. The neural network and Markov chain techniques are shown to achieve good performance rates of 81-84% and 78-80% respectively. Preliminary findings concerning the utilisation of the classified stress labels for phoneme recognition are provided and enhance the cause for prosodic information to be included within continuous speech understanding systems.

INTRODUCTION

Prosody comprises the suprasegmental phenomena of speech that extend over regions larger than the consonants and vowels, such as intonation and rhythm, and which are generally linked to the higher levels of linguistic processing, like meaning, speaker intent and speaker attitude. Prosodic information has long been recognised as having useful properties that could be exploited in a continuous speech understanding system (Lea, 1980). Specifically, as part of the prosodic structure of speech, syllable stress or prominence can be of great benefit by providing assistance in lexical access (Cutler & Norris, 1988), locating islands of reliability for phoneme recognition (Lea, 1980), and marking the likelihood of new versus old information. However, the lack of effective methods for the exploitation of stress information can be attributed to the many different definitions of stress and other related terms that can be found, plus the difficulty in the identification of this component of speech. Also, the variety of related research work focusing upon different target phenomena for recognition has made the comparison of results difficult. In this work, from the use of a common speech data set definite comparisons can be confidently made between the three classification methods presented.

Although there is dispute over the number of identifiable stress levels, three levels of stress were used in this work: primary stress, secondary stress and zero (or no) stress. Primary stressed syllables were those deemed to be prominent and carrying the utterance's rhythmic beat, whilst any remaining strong syllables were considered to be secondary stressed. All other syllables were unstressed, by default.

Stress is usually realised by a combination of increases in fundamental frequency, intensity and duration, and these changes mostly occur over the sonorant portion of a syllable (Morton & Jassem, 1965). Given these acoustic correlates of stress, the six features shown in Table 1 were chosen for this work. All features were normalised to lie in the range [-1, 1] by using the corresponding feature maximum and minimum from each sentence concerned.

Feature	Number
peak-to-peak amplitude integral over nucleus	f_1
energy mean over nucleus	f_2
nucleus duration	f_3
syllable duration	f_4
maximum pitch over nucleus	f_5
mean pitch over nucleus	f_6

Table 1. Acoustic features for syllable stress classification.

PROCESSING OF SPEECH DATA

Dialect one of the DARPA TIMIT continuous speech database was used, with fourteen female and twenty-four male speakers reading aloud eight sentences each. No constraints were placed upon the number of syllables allowed per word in the selection of this data.

Perceptual stress assignments were conducted with two listeners independently listening to the sentences and labelling all of the syllables with one of the three stress types. The cross-listener consistencies were found to be about 82% in comparing three levels of stress for both the female and male speaker sentences. In grouping the primary and secondary stress classes together, the cross-listener consistencies rose to 91.5% and 93.7% for the female and male speaker sets, respectively. A further joint session was held and resulted in 99.8% agreement between the listeners' stress labels. All of the sentences were then hand-segmented to mark the boundaries of each syllable's vowel nucleus, which gave a total of 3796 nuclei.

Two female and four male speakers' sentences were separated from the corpus to form an unbiased testing set. This was done to represent a collection of new speakers and (generally) new sentences that the classifiers would not have "seen" during the training phase. The rest of the corpus was divided into two groups by extracting every sixth syllable from the entire set (i.e., ignoring sentence boundaries) and forming a biased testing set, leaving the remainder as a training set. The biased testing set represented a collection of new sentences for the classifiers, but the same speakers and background environments as in the training set. Table 2 shows the resultant number of syllables in each of the stress classes for the all-speaker group.

Data Set	Primary Stress	Secondary Stress	Zero Stress
Training	725	456	1490
Biased Testing	145	99	291
Unbiased Testing	163	105	322

Table 2. Number of stress class syllables in each set for the all-speaker group.

STRESS CLASSIFICATION METHODS

Three different techniques were developed and investigated for the task of stress classification. Due to stress being relative in nature context needed to be appropriately included in these processes. Previous research has generally achieved contextual knowledge through the use of relative features, however because the features used here were absolute the inclusion of the context features were necessary. Feedforward neural networks accommodated the context spatio-temporally, the Markov chains incorporated context into its state transitions, whilst the rule-based recognisers explicitly provisioned for context in its set of rules. The three contexts investigated were the syllable preceding the one to be classified (context 1-0), the two syllables preceding the one to be classified (context 2-0) and the three syllables preceding the one to be classified (context 3-0). Undefined context feature vectors were set to all zeros.

All of the classifiers exhibited performance improvements of approximately 10% by grouping the primary and secondary stress classes into one "stressed" class. This was due mostly to the 50% improvement in the secondary stress class recognition rates, and thus only the results from this two stress class case will be discussed in the rest of the paper.

Neural Network Classifiers

Feedforward neural networks with two hidden layers were trained using two adaptive backpropagation algorithms that dictated when the weights were adjusted: *adap33* initially updated the weights every 3 training examples and then increased this update size by 3 every 20 epochs, whereas *adap55* initially updated the weights every 5 training examples and then increased this update size by 5 every 20 epochs. Sigmoid activation functions were used in the networks, with the learning rate set to 0.05 and the momentum rate to 0.2. The weights and biases were pseudo-randomly initialised to lie in the

range [-1, 1]. Two recurrent neural network architectures had been examined in previous work (Jenkin & Scordilis, 1994) but were abandoned due to unsatisfactory classification results of around 56%.

Six input neurons for each context and syllable vector were required, so there were 12 input neurons for context 1-0, 18 for context 2-0 and 24 for context 3-0. These input vectors were ordered according to their relative positions in the original sentence. The number of output neurons used matched the number of distinct classes to be categorised. The arbitrarily chosen hidden layer dimensions were directly dependent upon each context type in order to handle the relative complexity of the different number of inputs. In terms of the first and second hidden layers sizes, 10 and 7 neurons were used for contexts 1-0 and 2-0, 13 and 10 neurons were also used for context 2-0, and 16 and 13 neurons were used for context 3-0.

The best results for the superior classification configurations are presented in Table 3 and usually occurred within 500 epochs of training. This is because overlearning of the training data occurred over time causing gradual recognition improvement for the biased test set versus recognition degradation for the unbiased test set. Context 3-0 did not provide a sufficient performance advantage over the smaller contexts to warrant its increased computational requirements.

Context, Method, Dimensions	Data Set	Overall Rate	Primary Stress	Secondary Stress	Zero Stress
1-0, adap33, 10/7	training	86.48	89.38	74.56	88.72
	biased	82.06	84.83	68.69	85.22
	unbiased	84.24	88.96	70.48	86.34
1-0, adap55, 10/7	training	86.37	89.24	74.78	88.52
	biased	82.06	84.83	68.69	85.22
	unbiased	84.58	88.34	71.43	86.96
2-0, adap33, 13/10	training	86.41	89.93	76.97	87.58
	biased	81.68	86.90	69.70	83.16
	unbiased	84.24	87.73	72.38	86.34
2-0, adap55, 13/10	training	86.48	90.48	77.85	87.18
	biased	81.87	87.59	70.71	82.82
	unbiased	84.41	87.73	73.33	86.34

Table 3. Best NN classification rates (%) for the all-speaker group with two stress classes.

Probabilistic Classifiers

To reduce the problem of data classification from a set of continuous-valued feature vectors into a set of discrete representative tokens, the training set was clustered into so-called "supervised" and "unsupervised" codebooks of size 32 and 64 following a method based upon the basic Isodata algorithm (Duda & Hart, 1973). Firstly, for both types of codebooks, the initial choice of cluster centroid vectors was made by selecting training examples to adhere to the 3:2:6 ratio of primary stress, secondary stress and zero stress classes in the set. This was done to minimise the possibility of zero membership for any of the final cluster centroids. To generate three different codebooks of each size, every 25th, 33rd and *adap*th training example was inspected and used if its particular stress class quota had not been exhausted, where *adap* was dependent upon the codebook and training set sizes.

Knowledge of each of the training example's stress class was used during the supervised clustering process. Their membership was constrained to be assigned only to the closest centroid of the same stress class, whereas for the two-stress class distinction primary and secondary stressed training examples could belong to either primary or secondary stress class centroids. The unsupervised clustering process did not use this stress class information and was the method finally adopted.

Since duration was already inherent in the feature representations of the syllables, it was unnecessary to follow the trend of using hidden Markov models. Instead, a simple left-to-right Markov chain with no state feedback or recursions seemed sufficient for the task. The Markov chain was

comprised of a set of states representing the codebook centroid indices plus the corresponding transition probabilities between these states. No initial state probabilities were necessary since the starting state was determined by the first context syllable's centroid index number. Also, since there were times when contexts were undefined, a special additional centroid index was included to represent this absent syllable in the transition process of the classifier.

For context 1-0, the first state S_1 represented the context and the second state S_2 represented the syllable for classification. The probability of the state sequence being identified as stress class j, where j is 0 for unstressed and 1 for stressed, can then be given by:

$$\Pr(S_2 = (k_n, j)) = \Pr(S_2 = (k_n, j) | S_1 = (k_{n-1}))$$

Also, for context 2-0, by combining the contexts into the first state, the probability can similarly be:

$$\Pr(S_2 = (k_n, j)) = \Pr(S_2 = (k_n, j) | S_1 = (k_{n-2}, k_{n-1}))$$

During training, the state transition and stress class representing each example were counted. Then for recognition the stress class with the highest count for each testing example's state transition was chosen. Two problems arose out of this method of classification: tied stress class counts, and the existence of unseen state transitions in the training set (resulting in zero counts for the stress classes). The resolution of the tied counts was performed by using either a default or an enhanced method. The default method simply chose the unstressed over the stressed class because of its higher frequency of occurrence, whilst the enhanced method selected the stress type according to the highest class membership of the syllable's codebook centroid from the original clustering process. A technique was implemented for finding alternative state transitions when zero stress class counts were encountered. This involved changing the states in the transition by successively finding those centroids that were closest to the original ones until a non-zero stress class count was found.

The best Markov chain classification rates are given in Table 4. No significant advantage was seen from using either of the two tie resolution methods. Context 1-0 achieved slightly better recognition performance over context 2-0.

Codebook, Tie Method, Context	Data Set	Overall Rate	Primary Stress	Secondary Stress	Zero Stress
25, default, 1-0	training	88.73	88.28	76.32	92.75
	biased	78.13	77.93	61.62	83.85
	unbiased	80.34	78.53	64.76	86.34
25, enhanced, 1-0	training	88.66	88.28	76.32	92.62
	biased	78.13	77.93	61.62	83.85
	unbiased	80.34	79.14	64.76	86.02
33, default, 1-0	training	88.02	86.62	76.32	92.28
	biased	77.76	77.93	56.57	84.88
	unbiased	77.12	77.30	57.14	83.54
33, enhanced, 1-0	training	88.02	86.62	76.54	92.21
	biased	77.57	77.93	56.57	84.54
	unbiased	77.29	77.91	58.10	83.23
25, default, 2-0	training	97.34	96.69	93.86	98.72
	biased	77.20	78.62	64.64	80.76
	unbiased	77.46	79.14	60.95	81.99
25, enhanced, 2-0	training	97.30	96.83	93.86	98.59
	biased	77.20	78.62	64.64	80.76
	unbiased	77.29	79.14	60.95	81.68

Table 4. Best Markov chain classification rates (%) for the all-speaker group with two stress classes using unsupervised codebooks.

Rule-based Classifiers

About half of the work conducted in this area has utilised rules to determine stress, therefore an investigation into this type of method was also carried out. To begin with, syllables were automatically classified as stressed if they had at least three sentence maximum features, whilst syllables with at least three sentence minimum features were automatically classified as unstressed. This was done regardless of the contextual information, a procedure similar to that done by Hieronymus (1989).

Then, in order to introduce context into the rules, each of the context feature vectors were altered to be the difference between their corresponding syllable feature vector and the original context feature vector:

$$\mathrm{cont}_{\mathrm{new}}(\mathrm{f_i}) = \mathrm{syll}(\mathrm{f_i}) - \mathrm{cont}_{\mathrm{old}}(\mathrm{f_i})$$

to take into account the relative nature of stress. Statistical analysis of these new context feature vectors revealed good separability between the stress classes and thus thresholds were established for each feature according to its likelihood of belonging to a particular stress class given its new value.

Using the all-speaker group's training set, the primary and secondary stress means for each feature and context type were averaged to generate a primary stress class threshold. Similarly, this was done with the secondary and zero stress means to generate a secondary stress class threshold. Then, two different stressed class thresholds were generated: the first was simply a copy of the secondary stress class threshold, whereas the second was an average of the primary and secondary stress class thresholds. The latter threshold proved to be the most successful in the rule-based system.

To facilitate classification, the number of context differenced features that were above their particular threshold were accumulated and then tested against a lower limit L_{str}. If the total was above limit the syllable was classified as stressed, otherwise unstressed. The best rates from the different limit values using the second type of stressed class threshold are given in Table 5. It shows the trade-off that occurred between the two classes when L_{str} was adjusted, plus the inferior recognition rates of 66-74% compared with the previous two methods.

Context, L_{str}	Data Set	Overall Rate	Primary Stress	Secondary Stress	Zero Stress
1-0, 3	training	75.70	67.45	59.21	84.77
	biased	74.02	64.83	56.57	84.54
	unbiased	75.93	69.94	54.29	86.02
1-0, 2	training	73.46	81.66	75.44	68.86
	biased	71.59	81.38	71.72	66.67
	unbiased	73.56	85.28	69.52	68.94
2-0, 2	training	72.74	65.24	52.85	82.48
	biased	72.15	67.59	50.51	81.79
	unbiased	69.66	63.19	44.76	81.06
2-0, 1	training	69.94	85.93	75.66	60.40
	biased	66.92	80.00	72.73	58.42
	unbiased	69.66	85.28	65.71	63.04

Table 5. Best rule-based classification rates (%) for the all-speaker group with two stress classes using type 2 thresholds.

DISCUSSION

A comparison of the performances for the different classification methods showed feedforward neural networks to be slightly more suited to the task of stress classification than Markov chains, whilst the rule-based system did not have an adequate set of rules to handle the complexity associated with the task. From the relatively poor performances achieved for the secondary stress class, it could be said that secondary stress requires more context to be recognised than the primary and zero stress

classes. This is supported by Lieberman's (1965) discovery that linguists could not distinguish between primary and secondary stress without segmental information.

Following on from the previous assertion that syllable stress may be used to locate reliable segments of speech, the outputs of a phoneme recogniser (Grayden & Scordilis, 1994) were correlated with the stress output labels from the best performing neural network classifier for the unbiased testing set. It was found that 66.7% of the stressed vowel nuclei were correctly identified by the phoneme recogniser, compared with only 39.8% of the unstressed vowel nuclei. When compared with the original perceptual stress labels these figures were 61.3% and 43.3%, suggesting that perhaps the classifier can locate reliable segments of speech more effectively than a human can.

SUMMARY

The neural network and Markov chain techniques were shown to achieve recognition rates of 81-84% and 78-80%, respectively, which is extremely comparable with other work conducted in the area. The rule-based classifier did not perform as well, with recognition rates of around 66-74%. Preliminary findings concerning the utilisation of the classified stress labels for phoneme recognition are provided and enhance the cause for prosodic information to be included within continuous speech understanding systems.

ACKNOWLEDGMENTS

This work was done at the Department of Electrical and Electronic Engineering, the University of Melbourne.

REFERENCES

Cutler, A. & Norris, D. (1988) *The role of strong syllables in segmentation for lexical access*, J. Experimental Psychology: Human Perception and Performance, vol.14, pp.113-121.

Duda, R.O. & Hart, P.E. (1973) *Pattern classification and scene analysis*, John Wiley and Sons: New York.

Grayden, D.B. & Scordilis, M.S. (1994) *A hierarchical approach to phoneme recognition of fluent speech*, Proc. 5th Australian International Conference on Speech Science and Technology, vol.2, pp.473-478.

Hieronymus, J.L. (1989) *Automatic sentential vowel stress labelling*, Proc. Eurospeech, pp.226-229.

Jenkin, K.L. & Scordilis, M.S. (1994) *Automatic methods of syllable stress classification in continuous speech*, Proc. 5th Australian International Conference on Speech Science and Technology, vol.2, pp.731-736.

Lea, W.A. (1980) *Prosodic aids to speech recognition*, In Trends in Speech Recognition, edited by W.A. Lea, Prentice-Hall: New Jersey, ch.8, pp.166-205.

Lieberman, P. (1965) *On the acoustic basis of the perception of intonation by linguists*, Word, vol.21, pp.40-54.

Morton, J. & Jassem, W. (1965) *Acoustic correlates of stress*, Language and Speech, vol.8, pp.159-181.