AUTOMATIC METHODS OF SYLLABLE STRESS CLASSIFICATION IN CONTINUOUS SPEECH

K.L. Jenkin and M.S. Scordilis

Department of Electrical and Electronic Engineering
The University of Melbourne

ABSTRACT - This paper addresses the task of classifying syllables from continuous speech into primary stress, secondary stress and zero stress categories using artificial neural networks. The results compare favourably with other studies performed in this area, while highlighting the problem of classifying the secondary stress class.

INTRODUCTION

Syllable stress or prominence is often used to communicate important discourse functions such as word meaning, sentence meaning and intent. All levels within a continuous speech recognition system can benefit from a syllable stress parser by allowing it to control lexical access, constrain syntactic and semantic parsing, mark new versus old information and discern emphasis or contrast (Kouper-Kuhlen, 1986, Lea, 1980).

Our previous work on syllable stress (Jenkin & Scordilis, 1993) examined the suitability of artificial neural networks (ANNs) for the task of classifying syllables into either a primary stress category or a combined secondary and zero stress category. This work also studied the effect of syllable context on recognition. The results of 83-89% for two speakers from the DARPA TIMIT continuous speech database warranted further investigation of the novel use of ANNs. Other studies have been dominated by statistical methods, such as hidden Markov models (Freij & Fallside, 1988, Wagner et al, 1990) and Bayesian classifiers (Waibel, 1986), or by rule-based systems (Hieronymus, 1989).

This paper addresses the task of classifying syllables from continuous speech into primary stress, secondary stress and zero stress categories. A comparative analysis of the different ANN architectures and learning methods employed for performing this task is given, and the results provide a solid grounding for future decisions made in the area of ANN syllable stress classification.

SPEECH CORPORA

Dialect one of the TIMIT database was chosen to provide the continuous speech sentences. Only the "si" and "sx" sentences of the fourteen female and twenty-four male speakers were used in this study. Each sentence was hand-segmented into syllable nuclei and then assigned stress values, as follows.

Segmentation of syllables

For the purposes of stress or prominence detection, a syllable boundary is defined to be at the onset of its vowel nucleus because this point is believed to coincide with the beat of perceived sentence rhythm. The start and end points of the nuclei were subsequently hand-labelled to provide a region of feature extraction for each syllable.

During segmentation certain decisions were made to resolve some of the ambiguities involved with the syllabification process. For the syllabic consonants /el/, /en/, /em/, and /eng/ the nucleus was included as schwa. For words spoken with a "missing" syllable the nucleus was included as schwa only if the phonetic context required a nucleus to be present. For example, the word "every" with the phonetic representation /eh/v/r/iy/ did not have the second possible nucleus included, whereas the word "has" with the phonetic representation /hv/z/ did have the nucleus included. For words that have ambiguous syllable counts due to variations in speaker pronunciation, such as "hire" and "higher", the inclusion of each nucleus depended upon the particular phonetic representation of the word.

Labelling of stress values

The TIMIT database contains no specific sentence stress information, as opposed to lexical stress information in the dictionary, so perceptual stress assignments had to be carried out. The eight sentences for each speaker were printed out on sheets in plain text format, thus having no segmentation information. The stress assignment sessions required each listener to mark a 1 above those syllables perceived as having primary stress and a 2 above those syllables perceived as having secondary stress. No mark (the default) above a syllable denoted zero stress.The stress values were first labelled independently by two listeners who followed the procedure outlined in Veatch (1991). Primary stress is defined to occur on the most prominent syllables which appear to carry the rhythmic beats of an utterance, while secondary stress occurs on syllables which appear strong in addition to these primary beats. This method of assigning stress is recognised as being top-down rather than left-to-right, which highlights the importance of context in the stress classification process, and also results in primary- and zero-stressed syllables being more common than secondary-stressed syllables.

Analysis of the two independent stress assignment sessions revealed a 60% agreement for the primary and secondary stress labels across both the female and male speaker sets. The two listeners then resolved 99% of the conflicting labels in a joint stress assignment session. Those speakers with an unresolved stress-labelled sentence were designated to form part of the test set.

SYLLABLE PROCESSING

The sentences were downsampled from 16 kHz to 8 kHz by prefiltering the signal with a 9th-order elliptic low pass filter and then directly decimating the filter output. This was done to minimise the processing requirements of the entire stress parser.

Feature extraction

The acoustic correlates of stress are consistently stated to be derived from pitch, energy and duration measurements (Fry, 1955, O'Shaughnessy, 1979). The six features used in Jenkin & Scordilis (1993) were again used in this study. They are: peak-to-peak amplitude integral over nucleus, energy mean over nucleus, nucleus duration, syllable duration, maximum pitch over nucleus and mean pitch over nucleus. These features were chosen also with an emphasis on minimal processing by using only the vowel nucleus of the syllable. Note that for the schwa nuclei only the syllable duration feature was non-zero.

The amplitude and energy features are calculated using a 16 msec rectangular window with 50% overlap. The nucleus and syllable duration measurements are derived from the nuclei hand-labels.

The pitch tracker used to find the mean and minimum pitch features is based on the algorithm described in Medan et al (1991), with the maximum possible pitch range limited from 40 Hz to 500 Hz. The signal data is first passed through a 4-4-5 FIR filter, approximating a lowpass filter, to remove unnecessary harmonics. Then the pitch value is found by choosing the first cross-correlation peak that exceeds a certain threshold. This threshold is adapted after each pitch value is found, and the possible pitch range is adapted after 3 successive voiced segments to reduce computational time and improve pitch tracking. Pitch is calculated every 10 msecs, which is five times the minimum pitch period allowed.

Normalisation

The feature vectors are normalised before being presented to each ANN. For each sentence the maximum and minimum values of each feature are determined. These values are then used to linearly normalise the associated sentence features onto the range [-1.0, 1.0].

Context

Due to the relative nature of stress knowledge of surrounding syllables, or context, needs to be incorporated into a syllable stress classifier (Jenkin & Scordilis, 1993, Moore & Roach, 1993). Our previous work showed that limited pre-syllable context outperformed limited post-syllable context, therefore only pre-syllable contexts are examined here: the two syllables previous to the syllable being classified (2-0) and the three syllables previous to the syllable being classified (3-0). Larger contexts would tend to increase stress recognition, however there is a trade-off between the inclusion of context and the training and classification time required for a more complex data space. For those syllables that had part or all of their context undefined the context feature vectors were set to zero.

NEURAL NETWORK ARCHITECTURES

The multi-layer feedforward network had three output neurons, one for each class, and two hidden layers whose size depended upon the data set being learned. Syllable context is represented spatially in the network by simultaneously presenting the feature vectors for the syllable context and the syllable to be classified to the input. Thus there were eighteen input neurons for context 2-0 and twenty-four input neurons for context 3-0.

The two recurrent networks used had three output neurons, one for each class, and one hidden layer. Syllable context is represented temporally by feeding back, for each successive forward pass, the hidden layer activations in the Elman network, while the Jordan network feeds back the output layer activations with an accumulative decaying history. The recurrent networks had six input neurons since only one syllable's feature vector was presented to the network for each pass.

The weights and bias terms for all architectures were initialised to lie in [-1.0, 1.0], using the same pseudo-random sequence for each training run. The learning rate was set to 0.02, the momentum rate to 0.3. For the Jordan network, the decay term was set to 0.7. The sigmoidal activation function was used, with the output neurons being set at >=0.9 for the desired class, <=0.1 for the other classes. Standard backpropagation was used to train the feedforward network, while backpropagation through time (Werbos, 1990) was used to train the recurrent networks (with the output error set to 0 for intermediate passes).

Each training configuration was set to run for at least 3000 epochs to give sufficient time to monitor the general behaviour of the network.

CLASSIFICATION EXPERIMENTS

For the purposes of investigating the ability of ANNs to classify stress, three data sets were generated from the female speaker group, the male speaker group and the entire speaker group. First, two female speakers and four male speakers were chosen to form an unbiased test set. No syllables from these speakers' sentences were seen by the network during training. The remaining speakers were then used to form the training set, with every sixth syllable in the training set being extracted to form a biased test set. The biased test set contains syllables that weren't presented to the network for training, but have come from the same sentences (same environment and speaker) as those syllables used in the training set. The resultant counts of syllables in each set is given below in Table 1.

	Primary Stress	Secondary Stress	Zero Stress
Female training set	270	172	544
Female biased test set	56	30	112
Female unbiased test set	55	37	120
Male training set	464	302	919
Male biased test set	80	51	206
Male unbiased test set	108	68	202
All training set	725	456	1490
All biased test set	145	99	291
All unbiased test set	163	105	322

Table 1. Counts of syllables in each stress class for the different speaker groups and data sets

The female speaker group was used as the basis for comparing the different configurations examined, while the male and entire speaker groups were employed to give the performance for the best configurations found. Initially, the networks were trained using batch learning to provide a standard for the other possible configurations. For the female set with context 2-0 the best overall performance was 72.2% for the biased test set and 72.3% for the unbiased test set, however the secondary stress performance was only 0% and 40% for the two test sets while the primary stress performance was 66% and 63%. For the two different recurrent network architectures the unsatisfactory results of 56% and 56% were obtained for the two test sets. Even the inclusion of a second hidden layer in the Jordan network did not improve these results. Other configurations yielded a similar performance for the recurrent networks so all results discussed hereafter will concern the feedforward networks only.

As can be seen from the examples in Table 2, the increase in performance of one class usually results in a decrease in the performance of one other class, or sometimes both of the remaining classes. The most common effect noticed during the training of the network was the gradual increase in performance of the secondary stress class with a degradation in performance of the primary and zero stress classes.

Epoch Number	Primary Stress Rate	Secondary Stress Rate	Zero Stress Rate	Overall Rate
280	83.63	0.00	87.50	70.89
460	61.81	5.40	95.83	70.89
920	56.36	40.54	87.50	70.89
1340	63.63	10.81	94.16	71.36
1720	63.63	43.24	84.16	71.36
6400	58.18	27.02	90.00	70.42
9280	70.90	27.02	84.16	70.42

Table 2. Recognition rates (%) of each stress class for the female 2-0 standard unbiased test set

To attempt to compensate for the lack of training examples in the primary and secondary stress classes the primary training data was duplicated and the secondary training data triplicated. This resulted in a worse performance of 71.2% and 68.9% for the two test sets, with the primary class still classified poorly and the secondary class not at all. This method was insufficient to combat the focus of learning on the unstressed class in the network. The two other methods of training concerned when the weights were to be updated. The first method (update 33) initially updated the weights every three examples and then increased this update size by three every twenty epochs, whereas the second method (update 55) initially updated the weights every five examples and then increased the update size by five every twenty epochs.

Due to the performance of the secondary class being less than 10% for three of the results and 0% for the rest of the results, Table 3 below focuses on the primary and zero stress classes while still retaining the overall rate of the three classes in the final column. Most of the results were found within the first one hundred epochs of training, before each network started to learn the secondary stress class.

Context and Learning Method	Test Set	Primary Stress Rate	Zero Stress Rate	Two Class Stress Rate	Overall Rate
female 2-0 standard	biased	66.07	94.64	85.11	72.22
	unbiased	63.63	94.16	84.57	71.69
female 2-0 compensate	biased	69.64	91.07	83.92	71.21
	unbiased	61.81	93.33	83.42	68.86
female 2-0 update 33	biased	75.00	90.17	85.11	72.22
	unbiased	74.54	92.50	86.85	71.69
female 2-0 update 55	biased	76.78	90.17	85.71	72.72
	unbiased	74.54	92.50	86.85	71.69
female 3-0 update 55	biased	76.78	91.96	86.90	73.73
	unbiased	72.72	92.50	86.28	71.22
male 2-0 standard	biased	78.75	94.17	89.86	76.26
	unbiased	82.40	92.57	89.03	73.01
male 2-0 update 55	biased	86.25	90.77	89.51	75.96
	unbiased	87.03	90.59	89.35	73.28
male 3-0 update 55	biased	87.50	91.74	90.55	76.85
	unbiased	87.03	90.59	89.35	73.28
all 2-0 update 55	biased	75.17	87.62	83.48	68.03
	unbiased	82.20	91.61	88.45	72.88
all 3-0 update 55	biased	74.48	87.97	83.48	68.03
	unbiased	80.98	92.54	88.65	73.89

Table 3. Best feedforward ANN recognition rates (%) of each stress class for different configurations

DISCUSSION

Interestingly, as the feedforward network progresses through the learning epochs, it classifies more and more secondary stresses correctly (while still having a high primary and zero stress training rate) so it does appear that the network can "learn" all three classes. Initially more than half of the misclassified secondary stress examples are seen as primary stress by the network, then the misclassifications divide evenly between the primary and zero stress classes as training continues.

The results for the primary and zero stress classes are good compared to related work in the stress recognition area, but the issue of concern here is the performance of the secondary class. The number of stress levels is constantly disputed within the linguistic community and most of the other studies have concentrated on the stressed/unstressed classification task (or in some literature strong versus weak classification). This begs the question of whether secondary stress can ever be classified by any type of stress parser, and whether it is necessary to distinguish primary stress from secondary stress with respect to the pragmatic use of the parser. Lieberman (1965) found that if segmental information was removed from the speech signal a listener could only distinguish between stressed and unstressed syllables, not between degrees of stress, which suggests that only two degrees of stress may have acoustic correlates independent of vowel quality. If the secondary stress class is combined with the primary stress class, to form a binary stressed-unstressed classification of syllables, the initial perceptual stress assignment sessions described earlier has the agreement result increased to 80%, whereas the increase is not as significant if the secondary stress class is combined with the zero stress class.

In examining the feature statistics for the three classes it is clear that the primary and secondary stress classes are closer than the secondary and zero stress classes. The mean of each feature is given in Table 4 below. However, there is still significant overlap between the secondary class and the other two classes. In our previous study, the result of ~90% for primary stress versus secondary and zero stress suggests that there is overlap even for the binary class decision, let alone for the ternary class decision required here.

	Primary Stress	Secondary Stress	Zero Stress
Peak-to-peak amplitude	51859	35788	13323
Energy mean	7359	6273	3905
Nucleus duration	1061	906	504
Syllable duration	2067	1893	1387
Maximum pitch	170	161	134
Mean pitch	115	116	104

Table 4. Means of features in each stress class for the female speakers

The results from Table 3 reveal that context 3-0 is not significantly advantageous over context 2-0, however it is possible that more surrounding knowledge is required for secondary stress to be learned and recognised. In examining the different neural network configurations we found that adaptive training improves performance over batch training by not allowing the larger zero stress class to dominate the weight changes. Further investigation into the difference in performance for the female and male speaker sets is still required.

CONCLUSION & FURTHER WORK

The discussion about the bad performance of secondary stress points to the possibility that perhaps secondary stress requires more context to be deduced compared to the primary stress and no stress classes, particularly since secondary stress is more subjective in nature. This seemed evident after discussions about the stress assignment sessions revealed that listener KLJ tended to use rhythm over other acoustic cues, whereas listener NSN tended to use the intonation contour to assign stress. This initially resulted in KLJ assigning more primary stresses than NSN. Perception of secondary stress may therefore be dependent upon the particular acoustic cues a speaker prefers to utilise. As Veatch (1991) points out, even consistency for one person assigning three levels of stress to the same utterance on separate occasions can be as low as 87%.

In order to fully benchmark the use of the feedforward ANNs for stress classification, a comparison of hidden Markov models and rule-based systems for the same data needs to be done.

To ultimately create a self-contained stress parser, an automatic segmentation procedure will be necessary. The pitch algorithm currently in use can be used to segment speech into voiced/unvoiced regions, then energy values can be compared for different frequency ranges to segment the voiced regions into vowel nuclei and voiced consonants.

ACKNOWLEDGMENTS

We would like to thank Nick Nicholas for sitting through many excruciating hours of stress assignment sessions. We would also like to thank David Grayden for providing ANN advice and support.

REFERENCES

Freij, G.J. & Fallside, F. (1988) *Lexical stress recognition using hidden Markov models*, Proc. IEEE ICASSP, 135-138.

Fry, D.B. (1955) *Duration and intensity as physical correlates of linguistic stress*, J. Acoust. Soc. Am., vol. 27, no.4, 765-768.

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

Jenkin, K.L. & Scordilis, M.S. (1993) *A continuous speech syllable stress classifier using neural networks*, Proc. ANZIIS-93, Perth, Western Australia, 249-253.

Kouper-Kuhlen, E. (1986) *An introduction to English prosody*, (Edward Arnold: Great Britain).

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

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

Medan, Y., Yair, E. & Chazan, D. (1991) *Super resolution pitch determination of speech signals*, IEEE Trans. on Signal Processing, vol.39, no.1, 40-48.

Moore, J. & Roach, P. (1993) The role of context in the automatic recognition of stressed syllables, Proc. IEEE ICASSP, 767-770.

O'Shaughnessy, D. (1979) *Linguistic features in fundamental frequency patterns*, J. Phonetics, vol.7, 119-145.

Veatch, T.C. (1991) *Impressionistic coding for phrasal stress*, In English Vowels: their Surface Phonology and Phonetic Implementation in Vernacular Dialects, PhD dissertation, ch.5, section 4, (Department of Linguistics: University of Pennsylvania).

Wagner, M., McKay, B., Sampath, S. & Slater, D. (1990) *Modelling the prosody of simple English sentences using hidden Markov models*, Proc. SST, 180-185.

Waibel, A. (1986) *Recognition of lexical stress in a continuous speech understanding system - a pattern recognition approach*, Proc. IEEE ICASSP, 2287-2290.

Werbos, P. (1990) *Backpropagation through time: what it does and how to do it*, Proc. IEEE, vol.78, no.10, 1550-1560.