Development and Comparison of Three Syllable Stress Classifiers

*Karen L. Jenkin & Michael S. Scordilis**

Telstra Research Laboratories, Clayton, VIC, Australia
*Wire Communications Laboratory, University of Patras, Greece

ABSTRACT

This paper describes the development of three alternative techniques for the classification of syllable stress in fluent speech. They are based on: (1) neural networks that use contextual syllabic information, (2) first and second order Markov chains that depend on a new dynamic vector quantization approach, and (3) a rule-based approach. Both the neural network and the statistical approach achieved performance above 80%, with the neural networks slightly outperforming the Markov models. Experimental results also show that stress classification could enhance speech recognition.

1. INTRODUCTION

As the demand for more advanced speech recognition applications increases, there is strong motivation for providing additional information, such as from the suprasegmental level, to speech processing systems. Extracting that information from continuous speech, utilizing it effectively and measuring its contributions could be important for new challenging applications currently under consideration, including the recognition of spontaneous speech, extracting the meaning of spoken messages and the processing of fluent dialogs.

Prosody is such information, usually associated with higher levels of linguistic processing and it provides significant additional details about the speech signal and the knowledge it conveys, such as word meaning, intonation, rhythm, emotional state, attitudinal state and speaking rate [1]. Prosody encompasses the suprasegmental phenomena of speech that extend over regions larger than the underlying consonants and vowels [2]. Prosody is perceived by changes in pitch, loudness and segment length [3]. A principal component of prosodic structure is stress or prominence. The underlying temporal structure of speech can be defined in terms of syllables, and therefore syllable stress or prominence conveys important prosodic information about a sentence.

Syllable stress results from a stronger force that gives syllables more prominence [4]. While there are several stress types or levels which have been considered elsewhere, in this work three types of stress were used: primary, secondary or zero stress. Primary stress (PS) is assigned to most prominent (stronger) syllables or syllables that carry the rhythmic beat of a spoken sentence. Secondary stress (SS) is assigned to any remaining strong syllables, but weaker than primary. Zero or no stress (NS) is assigned to all remaining syllables.

Syllable stress is extremely useful in speech processing. Most pioneering work in utilizing stress has been in the area of text-to-speech synthesis, where the need for producing intelligible and natural-sounding speech is paramount [5]. In the areas of speech recognition and understanding in spite of the potential benefits from prosodic information its involvement up until now there has been limited [1]. Syllable stress can be useful in facilitating lexical access in isolated word recognition systems. It is estimated that 75% of content words begin with a strong syllable and lack of accent is a reliable indication of a function or repeated word [6]. Stress also indicates new information in dialog sentences. In phoneme recognition stress classification can be useful for identifying vowel phonemes that are more accurately and consistently articulated, in other words stressed, and which are unlikely to be missed or substituted by the recognition system.

The acoustic correlates of stress that can be objectively measured are (1) the signal intensity, which corresponds to the effort exerted during speech production, (2) the fundamental frequency (F0) and its dynamics, and (3) the syllable duration. For the latter, primary and secondary-stressed syllables are longer than those with no stress. In a syllable, durational variations are carried by the vowels, while the consonants are more duration invariant [7].

Stress levels cannot be expressed in terms of fixed feature values but are realized by combinations of relative changes of different acoustic features of speech. In this work, the features used for stress classification are listed in Table 1, and nucleus denotes the vowel part of the syllable.

Features used for classification
peak-to-peak amplitude integral over syllable nucleus (f1)
energy mean over nucleus (f2)
nucleus duration (f3)
syllable duration (f4)
maximum pitch over nucleus (f5)
mean pitch over nucleus (f6)

Table 1: Features used for syllable stress classification

Examination of the means and variances of the used syllable features indicated that while some features had good spatial separation, others were closely located. Because syllable stress is a relative quantity, context was used to provide the additional information required for classification, by including syllables prior to the one under consideration.

The speech material used for this work came from the TIMIT speech database and it comprised twelve female and twenty four male speakers from dialect 1, each uttering eight different sentences. The formation of the syllables was facilitated by the phonemic transcription of the spoken text and by the markings provided in this fully annotated corpus. For the development of the reference material, two expert listeners independently

identified each syllable according to each of the PS, SS, NS classes. This subjective perceptual classification resulted in 82.5% overall agreement between the two evaluators, 92.5% agreement in Stress/No-stress classification, and 89% agreement in the PS/NS task. Conflicts were resolved in a common listening session.

In order to gauge performance for new speakers and new sentences, biased and unbiased testing sets were formed. The biased set consisted of new syllables from speakers and sentences used for training. The unbiased set contained syllables from sentences and speakers not seen before.

2. SYLLABLE STRESS CLASSIFIERS

Three different approaches were developed in order to address the automatic syllable stress classification problem: a connectionist approach, a stochastic approach, and a rule-based approach. For the first two methods the values of the extracted syllable features were normalized in the [-1, +1] range.

2.1. Neural Network Classification

Syllable stress classification depends on context. In neural networks under supervised learning, recurrent architectures include context temporally, while feedforward architecture can include it spatially . In a previous study both types of networks were examined but best results were obtained with the feedforward network and was therefore chosen for a complete investigation of this classification problem [8]. The standard sigmoidal activation function was employed. Training was implemented using a fast version of backpropagation with two additional enhancements to further speed up training. They were "adapt33", which initially updated the network weight values every 3 examples and then increased this update size by 3 every twenty epochs, and "adapt55" which worked the same way but on 5 examples.

The feedforward network consisted of input, hidden and output layers, and best results were obtained with two hidden layers, whose sizes were empirically determined. The output layer had three neurons; one for each stress type. The input layer size was a multiple of the number of signal features used (i.e., six).

Syllable context was involved by including up to three neighboring syllables, as represented by their corresponding features in network development and testing. This way, a number of input syllable configurations were tested ranging from the syllable in question and three syllables before it (context 3-0), making a total of four input syllables, to one syllable before it (context 1-0), for a total of two input syllables. For reasons of economy in computation, smaller input configurations were favored over larger ones which were discarded if no improvement was provided.

For all neural architectures tested, classification of secondary stress in the three stress classes problem failed. However, when the overall performance rate was computed it exceeded 70%. Performance of the biased set was better than the unbiased set. The two stress type (Stress/No-stress) situation was also tested by considering PS and SS as a single stressed class. The final best results with this type of classifier were obtained for contexts 1-0 and 2-0, and they are shown in Table 2, for the different hidden layer sizes and training enhancements used.

Context, Method, Hidden layer size	Data Set	Overall Rate	Primary Stress	Secondary Stress	Zero Stress
1-0, adap33, 10 and 7	biased	82.06	84.83	68.69	85.22
	unbiased	84.24	88.96	70.48	86.34
1-0, adap55, 10 and 7	biased	82.06	84.83	68.69	85.22
	unbiased	84.58	88.34	71.43	86.96
2-0, adap33, 13 and 10	biased	81.68	86.90	69.70	83.16
	unbiased	84.24	87.73	72.38	86.34
2-0, adap55, 13 and 10	biased	81.87	87.59	70.71	82.82
	unbiased	84.41	87.73	73.33	86.34

Table 2: Best neural network test set recognition rates (%) for the all-speaker group and two stress classes.

2.2. Probabilistic Classification

Speech production has been successfully modeled as a stochastic process, and the hidden Markov model is a powerful technique used for speech recognition. There are many similarities with the task of syllable stress classification where probabilities of left-to-right state transitions must be determined and no recursions are permitted. The difference in this case is that syllabification is already provided and thus the need for hidden states does not arise. As a result, in this work Markov chains were used to model stress in continuous speech. The system training amounted to specifying the number of states and determining the state transition probability matrix [9].

The feature space was discretized in 32 or 64 vectors. Codebook design was based on the Euclidean distance metric and the basic ISODATA algorithm [10]. Both supervised and unsupervised design approaches were examined. In the supervised case, the portions of the codebook were reserved for each stress class, with a 30-20-50 split for the PS, SS, NS classes respectively. However, no performance benefits were gained from this method over an unsupervised codebook design, which was finally adopted.

Every syllable n is part of a spoken sentence consisting of syllables *1* (first syllable) to N (last syllable), and it is characterized by a feature vector $f_n = \{f_{n1}, f_{n2}, f_{n3}, f_{n4}, f_{n5}, f_{n6}\}$. In the codebook design process every syllable feature vector was assigned to a vector k_n, $\{k=1,\ldots, K\}$, with K = 32 or 64. Syllable context 1-0 was modeled in a first order Markov chain with state S_1 being the past syllable and state S_2 representing the present syllable. This way, the transition probability from the past to the present state can be written as:

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

where $j = \{0, 1, 2\}$ and it is the stress type (NS, SS, PS). The state transition matrix had the form:

$$S_1 = (k_{n-1}) \quad \begin{matrix} \\ 1 \\ \vdots \\ K+1 \end{matrix} \begin{matrix} S_2 = (k_n, j_n) \\ 1,0 \quad 1,1 \quad 1,2 \quad \cdots \quad K,0 \quad K,1 \quad K,2 \end{matrix} \begin{bmatrix} . & . & . & ... & . & . & . \\ & & & & & & \\ . & . & . & ... & . & . & . \end{bmatrix}.$$

Vector K+1 represented the default for syllable at n = 0, the undefined sentence-initial context. Syllable context 2-0 was modeled in a second order Markov chain with state S_1 and S_2 representing the past syllables and state S_3 being the present syllable. In this case, the transition probability can be written as:

$$\Pr\left(S_3 = (k_n, j)\right) = \Pr\left(S_3 = (k_n, j) \middle| S_2 = (k_{n-1}), S_1 = (k_{n-2})\right).$$

By combining states this model was reduced to a first order Markov chain and the corresponding state transition matrix was:

$$S_1 = (k_{n-2}, k_{n-1}) \quad \begin{matrix} \\ 1,1 \\ \vdots \\ 1,K+1 \\ \vdots \\ K+1,1 \\ \vdots \\ K+1,K+1 \end{matrix} \begin{matrix} S_2 = (k_n, j_n) \\ 1,0 \quad 1,1 \quad 1,2 \quad \cdots \quad K,0 \quad K,1 \quad K,2 \end{matrix} \begin{bmatrix} . & . & . & ... & . & . & . \\ & & & & & & \\ . & . & . & ... & . & . & . \\ & & & & & & \\ . & . & . & ... & . & . & . \end{bmatrix}$$

where S_1 and S_2 merged into a new S_1 and new S_2 is same as the original state S_3.

During testing, two problems had to be overcome. One problem was the resolution of ties in the state transition matrix for which a default and an enhanced method were used. The first used the {NS, PS, SS} ordered list of preferences, with no stress being the most frequently occurring and secondary stress being the least likely to occur. The second method selected the stress type according to the type that occurred most often during training, for that codebook vector. The other problem was caused from the zero entries in the transition matrices which resulted from certain syllable sequences not being seen during training . In this case, a scheme for alternative assignment of state transitions was developed, which shifted states in the transition matrix, in turns, according to the distance that had to be covered in the feature space by each attempted shift. Stress type was then decided based on the new transition that was assigned, which was closest to the original and had occurred at least once during training.

The recognition of secondary stress for this classifier was also low but quite better than for the neural networks. Performance of the biased set was better than the unbiased set for separate speaker genders. Best recognition results for Stress/No-stress classification were obtained for context 1-0 and codebook of size 32 as shown in Table 2. Codebook index denotes the method of selecting every 25th or 33rd training syllable for the initial assignment of centroid vectors at the start of the iterative codebook design process. Classification success during training is also included for reference.

Codebook Index Tie Method	Data Set	Overall Rate	Primary Stress	Secondary Stress	Zero Stress
25 default	train	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	train	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	train	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	train	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

Table 2: Recognition rates (%) for all speakers, two stress classes, context 1-0 using codebooks of size 32.

2.3. Rule-based Classification

This study was completed with the development of a rule-based classifier for the classification for contexts 1-0 and 2-0. The first step in the method used for generating rules was to measure stress classification performance in the absence of context. This was done by testing if any of the features used were sentence maxima or minima as in [11], and then automatically assigning them to be primary stressed (stressed for two stress classes) or zero stressed respectively. For the complete training set, it was found that syllables had at least two maxima and minima in their feature vector. No syllable feature vectors contained both three maxima and three minima, so this was the smallest number allowed in the assignment of stress. This approach was moderately effective in stress classification, with best results obtained for zero stress and for the unbiased data set.

The remaining step for rule generation was to meaningfully include the context. This was achieved by resetting all context features, f_i, to be the difference between their corresponding syllable feature, $syll(f_i)$, and the original context feature, $cont_{old}(f_i)$, as follows:

$$cont_{new}(\mathrm{fi}) = syll(\mathrm{fi}) - cont_{old}(\mathrm{fi}).$$

A statistical analysis of these context 1-0 and 2-0 differenced features was performed for both the female and male speaker training sets on the basis of the stress class type of their corresponding syllable *syll*. Examination of the means revealed good separability between the three stress classes for both of the context differenced means, particularly 1-0, for most features.

By using the context differenced feature means, thresholds were allocated to each feature to determine whether it was likely to be associated with a syllable of a specific stress class type. First, for each feature and context type, the smallest primary stress mean and the largest secondary stress mean in the two training sets were chosen. Then, the thresholds for the primary stress class, for both contexts considered, were calculated by taking the average of

these two values. Similarly, for each feature and context type, the smallest secondary stress mean and the largest zero stress mean in the two training sets were chosen and then averaged together to get the thresholds for the secondary stress class, for both contexts considered.

For the two stress class situation two types of thresholds were calculated. The type 1 thresholds were simply a copy of the associated context and feature secondary stress thresholds, since the stressed class was comprised of both the primary and secondary stress classes. On the other hand, the type 2 thresholds were the computed average between the associated context and feature primary and secondary stress thresholds.

For the three stress class recognition task, the rule-based system's ability to distinguish secondary stress was low and comparable to the probabilistic approach. For the two stress classes task (stress/no-stress) classification performance was better for the biased data set for gender-specific systems, but on the combined set performance was about the same for both the biased and unbiased sets. Classification performance for context 1-0 was somewhat better than for context 2-0, and the results are summarized in Table 3 for the two stress categories.

Configuration. #, type thresholds	Data Set	Overall Rate	Primary Stress	Secondary Stress	Zero Stress
4, 1	train	72.67	62.21	56.14	82.82
	biased	70.65	57.24	56.57	82.13
	unbiased	70.68	60.12	49.52	82.92
3, 1	train	71.88	79.17	72.37	68.19
	biased	69.35	76.55	68.69	65.98
	unbiased	71.53	82.82	67.62	67.08
3, 2	train	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
2, 2	train	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

Table 3: Rule-based classification results (%) for all speakers, context 1-0, two stress classes

3. APPLICATION TO SPEECH RECOGNITION

The hypothesis that stress information can be useful in speech recognition was finally tested. This was done using a continuous speech recognition, also developed on the TIMIT database, which provided strings of recognized phonemes [11]. Phoneme recognition performance of the vocalic nuclei of stressed syllables exceed that of unstressed syllables by more than 70%. For the unbiased data that difference was greatest for the neural network classifier with 67% and 39% recognition rates, while for the probabilistic classifier the respective numbers were 64% and 42%. For the hand-labeled syllables the performance difference was even lower (61% and 44%), suggesting that the human classifier takes a lot more information into this decision-making process and does not give the kind of information that is appropriate for this task.

4. SUMMARY

This paper presented two new stress classification methods. The first was based on neural networks and it achieved performances of 81-84%. The second was a Markov chain which achieved performances between 78-80%. These were contrasted with a rule-based system which had performances of 67-75%. In a continuous speech recognition system vowels classified as stressed, using the developed techniques, could be correctly identified 70% more often that those marked as unstressed. This information could be used to enhance speech recognition.

Acknowledgment

Most of this work was done at the Department of Electrical and Electronic Engineering of the University of Melbourne.

5. REFERENCES

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

2. Clark, J. & Yallop, C., *An introduction to phonetics and phonology*, Basil Blackwell: Oxford, 1990.

3. Johns-Lewis, C. (ed.), *Intonation in discourse*, Croom Helm & College Hill Press: Great Britain, 1986.

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

5. Klatt, D., "Review of text-to-speech conversion for English", *J. Acoust. Soc. America*, vol.82, no.3, pp.737-793, 1987.

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

7. Waterson, N., *Prosodic phonology: The theory and its application to language acquisition and speech processing*, Grevatt and Grevatt: Great Britain, 1987.

8. Jenkin, K.L & Scordilis, M.S., "Automatic Methods of Syllable Stress Classification in Continuous Speech", *Proceedings of the 5th Australian International Conference on Speech Science and Technology*, vol.2, pp.731-736, 1994.

9. Taylor, H.M. & Karlin, S., *An introduction to stochastic modeling*, Academic Press: Orlando, Florida, 1984.

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

11. Hieronymus, J.L., "Automatic sentential vowel stress labelling", *Proceedings of Eurospeech*, pp.226-229, 1989.

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