THE AUTOMATIC MARKING OF PROMINENCE IN SPONTANEOUS SPEECH USING DURATION AND PART OF SPEECH INFORMATION

Matthew Aylett *Matthew Bull*

Human Communication Research Centre,
University of Edinburgh

ABSTRACT

The work reported in this paper was the result of the need to label a large corpus of spontaneous, task-oriented dialogue with prosodic prominences. A computational model using only word duration, part of speech and a dictionary lookup of each word's canonical phonemic contents was trained against the results of a human coder marking prominence. Because word durations were normalised, it was possible to set a common threshold for all members of a form class above which the lexically stressed syllables were classed as prominent. The method used is presented and the relative importance of duration information, phonemic contents, syllabic context and part of speech information is explored. The automatic coder was validated against unseen material and achieved a 58% agreement with a human coder. Further investigation showed that three humans coders agreed no better with each other than each agreed with the computational model. Thus, although the automatic system did not conform very well to the performance of any one human coder, it conformed as well as another human coder might.

1. INTRODUCTION

This paper presents a practical approach to the assignment of prosodic prominences to a large task-oriented corpus of dialogues (the HCRC Map Task Corpus [1]). The size of the corpus - 128 dialogues, each several minutes long - made it desirable to assign prominences automatically. Our aim was to produce a prominence assignment model which would mimic the perceptions of a group of subjects as closely as possible. For the purposes of this study, a word was classed as prominent if it contained a primary stressed syllable, and non-prominent if it did not. It would, however, be quite possible to alter the constraints of the model to account for the more conventional notions of pitch-accent or nuclear-stress.

The HCRC Map Task Corpus has been word segmented by hand giving all word durations. Other acoustic factors such as amplitude, pitch and vowel quality (All of which affect the perception of vowel quality) are not as readily available. For this reason duration was used as our primary acoustic measurement for automatically determining prominence. The model predicts a duration for the stressed form of each word and assesses the probability the observed word is stressed by comparing its duration with the prediction.

2. BASIC MODEL

The basic model used a combined log distribution model of each phonemic segment (as in [5]), and assumed that a change in the duration of a word is divided equally among the segments of that word in terms of z-scores for duration. Therefore, the change between a word's predicted duration and actual duration could be measured in terms of a single z-score calculated for all of a word's segments. This value, called here the 'k-score', was used as a measure of how much a word had been 'stretched' or 'compressed' from a citation form.

The predicted duration, d, of any word may be expressed as:

$$d = \sum_{i=1}^{n} \exp^{(\mu(i)+k\sigma(i))} M \qquad (1)$$

where:

n = the number of phonemes in a word,
k = a constant function of average segment length,
μ = the mean log duration of a segment,
σ = the standard deviation of the log distribution of a segment's duration
M = an optional multiplier which defaults to 1. (see Table 1)

K-scores were calculated by assuming an initial k-score of 0 for each segment in a word. If the resulting value for the predicted word duration (according to the equation above) was higher than the observed word duration, a lower k-score (-0.001) was used. If the predicted word duration was lower than the observed duration, a higher k-score (+0.001) was used. This process was continued until the predicted and actual word durations were the same. The value of the k-score at this point was taken as a measure of the difference between predicted and observed word durations. A threshold k-score was set separately for each form class to maximise agreement with a single human coder. Words falling below the threshold for their form were labelled as unstressed.

There were two major constraints on the resources available to us. Firstly the amount of phonetically labelled spontaneous speech was limited to only two dialogues. Secondly the online dictionary we had available (CELEX [2]) is based on standard English pronunciation, whereas most of the Speakers in the map task have Glaswegian or other Scottish accents. In order to explore the effects of different factors on the success of the model and to work within these constraints six different models were tested. Two were controls and another four made different use of syllabic and phonemic information.

2.1. Control model

This model acted as a control. If a word was open class (in this case either an adjective, noun or verb - adverbs were regarded as closed class) then it was automatically stressed. If the word was closed class (anything else) it was regarded as unstressed. The success of this model gives an indication of possible success in assigning stress without any duration information.

2.2. Simple μ, σ model

One log distribution was used ($\mu = -2.7478(64ms)\sigma = 0.5702(-1sd = 36ms, +1sd = 113ms)$) for **all** phonemes, so that there was effectively no differentiation between phonemes. Expected word durations therefore depended on how many segments there were in any given word. Again this model acted as a control showing how good a model with no knowledge of either the phonemic contents or the syllabic structure would be at predicting prominence.

2.3. Syllabic M model

M in equation 1 was varied to account for syllabic information whereas μ and σ were as above (the same for **all** phonemes). Table 1 shows the values for M used which depended on syllabic context. These values, based on durations established by [3] from measurements taken from a phonetically balanced read corpus [4], are proportions with regard to the mean segmental duration of a segment in a three segment stressed monosyllabic word. For example if a segment is predicted to be 100ms in a 3 segment stressed monosyllabic word then, if it is in an unstressed 4 segment monosyllabic word, the duration is reduced to 36.6ms.

Syllabic Multipliers					
		Syllabic Context			
		mono	initial	middle	final
Stressed	1 seg	1.632	1.088	1.008	1.600
	2 seg	1.163	0.775	0.718	1.140
	3 seg	1.000	0.680	0.630	1.000
	4 seg	0.949	0.632	0.586	0.930
	5 plus seg	0.887	0.592	0.548	0.870
Unstressed	1 seg	0.549	0.522	0.585	0.900
	2 seg	0.390	0.371	0.416	0.640
	3 seg	0.366	0.348	0.390	0.600
	4 seg	0.366	0.348	0.390	0.600
	5 plus seg	0.366	0.348	0.390	0.600

Table 1. Multipliers for different syllabic context. For example if a segment is in a three segment stressed mono-syllabic word the multiplier is 1.000, if it is in a four segment unstressed final syllable in a polysyllabic word the multiplier is 0.6 (see equation 1).

2.4. Syllabic μ, σ model

Instead of M, μ and σ were varied to account for syllabic context. Data was collected from two phonetically hand segmented spontaneous dialogues. For each syllabic context (for example stressed segment in 3 segment monosyllabic word) a log distribution of segmental durations was calculated giving a different μ and σ for each context. When estimating the k-score of a word these varying μ and σ were used in Equation 1. A problem with the syllabic multipliers described above (The Syllabic M model) was that they were calculated on the basis of a phonetically balanced read corpus. This model explored the advantage of using distributions calculated from spontaneous speech.

2.5. Phonemic μ, σ model

μ and σ now depended on a log distribution of segmental durations for each phoneme as observed in the balanced corpus in [4]. The CELEX online dictionary [2] was then used to establish the likely phonemic contents of each word in the corpus. The problem that resulted from this was whether it was valid to model different Scottish accents with Standard English data. However, what was important here was not the precise phonetic quality of any given segment, but rather its general class. As long as any differences in pronunciation are small enough that their corresponding durations are also similar, the predicted word durations should be relatively reliable.

2.6. Combined model

This model combined the Syllabic M model with the Phonemic μ, σ model so that both segmental content and syllabic context were represented.

3. METHOD FOR MANUALLY-LABELLING THE TEST DIALOGUES

Two test dialogues were selected from the Map Task Corpus. Three subjects **a,b,c** who were experienced phoneticians were presented with the dialogues. They could see a speech waveform, and hear selected segments of speech as much as they felt necessary. However, subjects were encouraged to make decisions as quickly as possible.

The subjects were asked to decide for each word in the dialogues whether that word sounded prominent in any way. They were not asked to make specific judgements about stress. If a word was perceived as prominent, the subjects marked the most prominent syllable in that word. All word and syllable boundaries had previously been marked for the subjects.

The word segmentation used by the subjects and the automatic model were not identical (See Figure 1). The subjects used word and syllable segmentation from the phonetically labelled dialogues whereas the automatic coder used word segmentation available for the whole corpus (its intended domain) and predicted syllable boundaries on the basis of which segmental model was

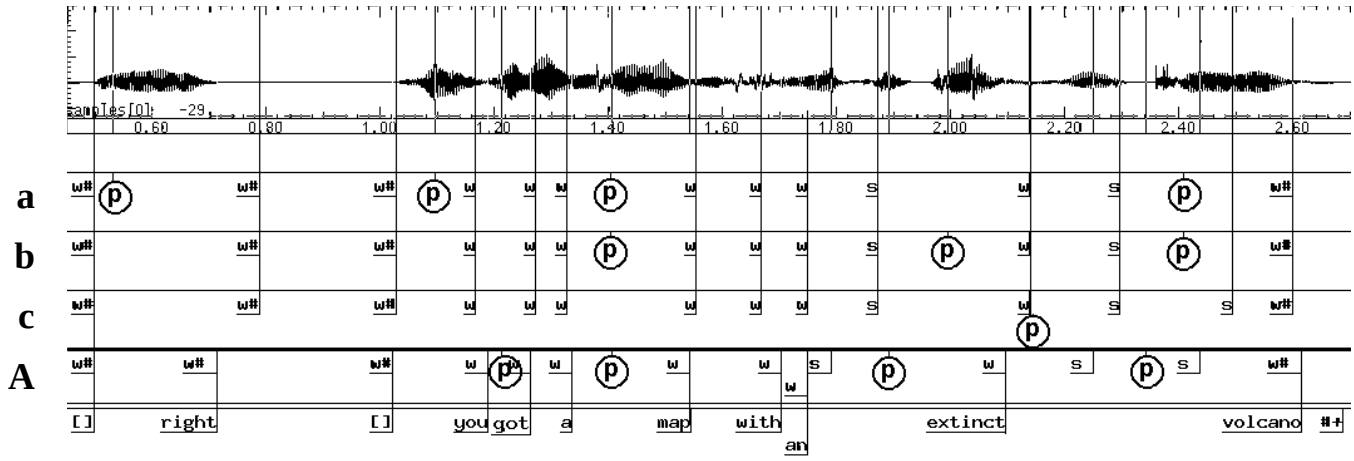

Figure 1: An example of some prominences coders by three human subjects (a,b,c) and the automatic coder (A). **w** marks word boundaries, **s** inter word syllable boundaries, and a circled **p** prominences on the speech "right, you got a map with an extinct volcano". The speech waveform is shown at the top. subjects were asked to mark prominences at the start of the syllable nucleus

used. The error in word boundary placement was low (mean of 0ms and standard deviation of 17ms) and although the syllable boundary error was higher (a mean varying from 11ms to 23ms with an standard deviation varying from 37ms to 42ms depending on the model) 85% of words in the Map Corpus are monosyllabic. Most errors caused by differences in syllabic and word boundaries were avoided by setting a threshold for matching stress markers (within 30ms of each other). A dynamic programming algorithm was used to count agreement between the automatic and human coders.

4. RESULTS OF A COMPARISON OF DIFFERENT MODELS

Each model was run on a test dialogue, and evaluated in terms of the numbers of stresses which agreed with or differed from the stresses marked in the manually-labelled test dialogue. The results were as shown in Table 2. The models were then applied to an unseen dialogue coded by coder **a**. The new dialogue contained speech from one speaker from the training dialogue and one new speaker (See Table 3).

The combined model was the most successful with the training data but not with unseen data. For the unseen data the syllabic M model is most successful agreeing with the human coder 58% of the time. It would seem that this model generalises more effectively across speakers and new data. For this reason the syllabic M model was selected as our final model. A comparison between the syllabic M model, **A**, and all three human coders who coded the training set is shown in Table 4.

There is good agreement between **a-A, b-A,** and **b-a**. In other words, the model predicted stress placement much like two of the subjects. The third subject, **c**, seemed to agree equally poorly with the other subjects as with the model.

5. CONCLUSION

As stated at the beginning of this paper, our primary objective was to solve a coding problem over a large corpus. The automatic coder selected, although its results have to be treated with caution, was a fairly good approximation to a human coder. Apart from solving a practical problem these results have some interesting implications.

Phonemic content was not as important as syllabic context when normalising duration. This was particularly true for long words where segments are significantly reduced. The syllabic context was a fundamental factor in this reduction. More surprising was that combining phonemic and syllabic information produced only a minor improvement in results when using the training data and appeared to be worse at generalising duration change across speakers and unseen data. Phonemic content is not independent of syllabic context. For example the phoneme ð occurs mostly as "th" in the word "the". Because of this the distribution calculated from a large numbers of observations of ð will underestimate the duration of this significantly in a stressed open class context e.g. the "th" in "mother". This lack of independence between phonemic contents and syllabic structure is widespread. Taking the stops **s,k** we find a marked difference in the frequency that syllables containing them are of a particular segmental length. 53% of syllables containg **s** are 2 or 3 segments in length whereas 73% of **k** syllables are this length and an enormous 94% of ð are 2 or 3 segments long. Because of syllabic structure, vowel and consonant distributions are also markedly different. For example 74% of syllables containing the dipthong **aI** (The 'i' in 'bite") are 3 segment syllables. This lack of independence between phonemic content and syllabic structure together with the fundamental important of syllabic structure in word duration means that generalising durational effects on the basis of syllabic context rather than phonemic content appears to be more effective.

Despite the considerable differences between the ATR database (used to calculate syllabic post modifiers and the phonemic dis-

tributions) and the spontaneous Glaswegian speech in the corpus, the models that used this data did better than the model which used the spontaneous speech to calculate distributions for segmental and syllabic context. Possibly the phonetically marked up data was too sparse to model such duration effects. However another possible explanation is that the read speech was less variable meaning that, in small cell sizes, the means and standard deviations calculated were more accurate. Thus, although the model overestimated the expected duration of the words in spontaneous speech, it did so consistently. When thresholds were generated by comparing to human coding decisions these more consistent results led to better performance.

Comparison between Models: Training Data				
	hits	misses	false alarms	%accuracy
Control	252	323	132	35.64
Simple μ, σ	367	208	158	50.07
Syllabic M	420	155	196	54.47
Syllabic μ, σ	381	194	169	51.21
Phonemic μ, σ	397	178	177	52.79
Combined	427	155	196	55.45

Table 2. Comparison of six models used to determine stress placement (against coder a). Accuracy = hit/(hits + misses + false alarms) as a percentage.

Comparison between Models: Unseen Data				
	hits	misses	false alarms	%accuracy
Control	365	336	127	44.08
Simple μ, σ	462	239	181	52.38
Syllabic M	538	163	220	58.41
Syllabic μ, σ	467	234	180	53.01
Phonemic μ, σ	493	208	214	53.88
Combined	531	170	249	55.89

Table 3. How well all six models performed when presented with unseen data and a new speaker (against coder a). Accuracy = hit/(hits + misses + false alarms) as a percentage.

Comparison between Model and Subjects				
	hits	misses	false alarms	%accuracy
a-A	420	155	196	54.47
b-A	400	104	216	55.56
c-A	253	63	363	37.26
b-a	390	114	185	56.60
c-a	260	56	315	41.20
c-b	262	54	242	46.95

Table 4. Cross-comparison of each subject's stress assignments and the assignments of the syllabic M model to the training data (Where A is the automatic model, and a, b, and c are the subjects). Accuracy = hit/(hits + misses + false alarms) as a percentage.

Given the disagreement between coders it might be worth asking whether making a simple binary decision on prominence in spontaneous speech is possible. Perhaps a graduated coding would have produced better results. Although this may present problem in terms of intonational phonology it seems difficult to justify a coding system which leads to such poor agreement. Work carried out by Grover et al [6] suggests that although boundary strength can be reliably coded on a four point scale, a magnitude estimation scale produces better results for the judgement of prominence.

However the results here are far from conclusive. We had only two phonetically segmented dialogues available to produce and test our model. Individual differences in these dialogues may well have confounded our results. Practical problems such as the unavailability of an on-line dictionary for Scottish (or even rhotic) pronunciation and differences in word segmentation may also have caused significant problems.

Overall the automatic coder was sufficient for our own purposes and the duration normalisation described here, despite clear drawbacks, offered a practical solution for comparing word durations. The results from the human coders raise questions concerning the overall practicality of marking binary prominence but given the limited scale of the study further work would be required to explore this issue. The contributions of syllabic context, phonemic contents and word class to a model of duration change were not entirely predictable. Our results suggest that syllabic context is the primary factor in a words duration especially when generalising across speakers.

REFERENCES

[1] Anne H. Anderson, Miles Bader, Ellen G. Bard, Elizabeth Boyle, Gwyneth M. Doherty-Sneddon, Simon Garrod, Stephen Isard, Jacqueline C. Kowtko, Jan M. McAllister, Jim E. Miller, Catherine F. Sotillo, Henry S. Thompson, and Regina Weinert. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.

[2] R. H. Baayen, R. Piepenbrock, and L. Gulikers. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995. Version 2.5.

[3] W. N. Campbell. *Multi-level timing in speech*. PhD thesis, Sussex University, 1992.

[4] W. N. Campbell. Multi-level timing in speech. *Advanced Telecommunications Research Institute Technical Report*, 1993.

[5] W. N. Campbell and S. D. Isard. Segment durations in a syllable frame. *Journal of Phonetics*, 19:37–47, 1991.

[6] C. Grover, B. Heuft, and B. Van Coile. The reliability of labelling word prominence and prosodic boundary strength. In A. Botinis, G. Kouroupetroglou, and G. Carayiannis, editors, *Proceedings of an ESCA Workshop: Intonation: Theory, Models and Applications.*, pages 165–168. ESCA and The University of Athens, October 1997.