LEXICAL STRESS AND PHONETIC INFORMATION: WHICH SEGMENTS ARE MOST INFORMATIVE?

David Carter*, Bran Boguraev* and Ted Briscoe†

ABSTRACT

Altmann (1986) investigates the relative usefulness of different parts of the speech signal for word recognition, and claims that a front end which is able to provide "fine class" (phonemic) information about segments in stressed syllables and "mid class" information about those in unstressed would on average be less helpful for word identification than one which was able to attach a fine class label to the same proportion of segments, but scattered at random throughout words. Altmann concludes that his results "demonstrate the desirability of a front-end which outputs occasional fine class information not systematically, but at random".

We show that Altmann's results are based on a faulty experimental procedure, which leads him to incorrect conclusions; fine class information about segments in stressed syllables is actually slightly more useful than that about randomly chosen segments.

1 INTRODUCTION

This paper is a response to claims by Altmann (1986) about the relative usefulness of different parts of the speech signal for large-vocabulary word recognition. Altmann accepts Huttenlocher's (1984) finding that a broad (manner of articulation) transcription only of segments in stressed syllables is more helpful for word identification than such a transcription only of those in unstressed. However, he introduces a "mid class" transcription that essentially distinguishes both manner of articulation and voicing, and uses it to obtain results suggesting that a front end which is able to provide fine class (phonemic) information about segments in stressed syllables and mid class information about those in unstressed syllables would on average be less helpful than one which was able to attach a fine class label to the same proportion of segments, but scattered at random throughout words. Altmann concludes that his results "demonstrate the desirability of a front-end which outputs occasional fine class information not systematically, but at random." If this is true, then it has important implications for the allocation of processing effort in front ends. However, we have been unable to reproduce Altmann's results, and indeed the data we have gathered provide evidence for a major error in the design of his experiments, and ultimately for the opposite conclusion.

2 ALTMANN'S PAPER

Altmann describes four experiments on a 18500-word lexicon (a version of the Oxford Advanced Learner's Dictionary) without word frequency information. In experiment (1), all the segments in the lexicon are given a mid class transcription. In the other three experiments, one third of the segments in the lexicon are given a fine class (phonemic) transcription, while the other two thirds are given a mid class transcription. In experiment (2), segments in stressed syllables are given a fine class transcription (almost exactly a third of all the segments in the lexicon are in stressed syllables). In experiment (3), the segments to be transcribed fine class are chosen at random, while in experiment (4), every third segment is transcribed fine class. Altmann reports the figures in table 1.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Mean class size (words)</th>
<th>% lexicon in unique (singleton) classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) all segments mid-class:</td>
<td>1.38</td>
<td>62%</td>
</tr>
<tr>
<td>(2) stressed segments fine:</td>
<td>1.22</td>
<td>75%</td>
</tr>
<tr>
<td>(3) random segments fine:</td>
<td>1.06</td>
<td>91.5%</td>
</tr>
<tr>
<td>(4) every third segment fine:</td>
<td>(results between those of (2) and (3))</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Altmann's (1986) results

Altmann notes that the payoff from extracting a fine class transcription of stressed syllables (2) is fairly small, giving an increase of only 13% in the uniqueness percentage. However, if random segments (3) are

*University of Cambridge, Computer Laboratory, Corn Exchange Street, Cambridge CB2 3QG, England
†Department of Linguistics, University of Lancaster, Bailrigg, Lancaster LA1 4YT, England
transcribed fine, the payoff is far higher. He explains this difference by noting that in (2), but not in (3), fine class segments will always be consecutive. Neighbouring segments will tend to be phonotactically more strongly correlated, and therefore less informative in combination, than non-neighbouring ones. This effect appears to outweigh the greater intrinsic informativeness of stressed segments.

However, if this explanation is right, one would expect transcription (4) to be still more discriminating, because in it, fine class segments are never neighbours. Altmann offers no explanation for the fact that (4) gives results intermediate between (2) and (3).

3 EXPERIMENTAL DESIGN

To interpret this anomaly, we need to look more closely at the whole process, used by Altmann and others before him, of evaluating a transcription by using it to partition the lexicon into equivalence classes and deriving statistics from the numbers (or frequencies) of words in those classes. The easiest way to derive equivalence classes is what we will call the assign-and-count method:

1. transcribe each word in the lexicon according to the current transcription (e.g. fine class for stressed segments, mid class for unstressed ones).
2. add it to the equivalence class defined by the resulting sequence of symbols;
3. finally, examine the set of equivalence classes for uniqueness, average size, or whatever quantities are deemed appropriate.

But this method is only valid if, whenever the pronunciation of a word matches the symbol sequence for a class, that word is put in that class. Otherwise, the classes are not equivalence classes at all, and no useful statistics can be extracted from them. To see this, suppose we run Altmann's experiment (3) (random segments transcribed fine) using assign-and-count. Then bit may be transcribed \( /b\) frontvowel \(\text{unvoicedstop}/\), and bet may be transcribed \( /\text{voicedstop frontvowel t}/\). If this happens, then the two words are put in different classes, but in neither case does the second segment receive the fine class analysis necessary for discrimination, and furthermore, the pronunciation of each word matches the symbol sequence for the other class as well as its own. This is clearly inappropriate. A front end that performs this transcription should enable bit to be distinguished from bet only one third of the time, i.e. when the vowel is transcribed fine; but "assign-and-count" will put bit and bet in different classes on about 86% of occasions. That is, unique identifiability will be predicted far too often.

To correct this, we must replace assign-and-count by an assign-and-lookup procedure in which a word is transcribed and lexicon entries are then matched against the resulting symbol string; the size and/or frequency count of the word set (not equivalence class) returned are then noted. This approach deals correctly with our bit/bet example; the word set for bit will exclude bet when, and only when, the vowel is transcribed fine. This strategy was adopted in repeating Altmann's experiments.

When the notion of an equivalence class is replaced by that of a matching word set, certain commonly-used statistics become meaningless. For example, because, in general, every word in the lexicon may generate a different word set, the idea of a mean class size (that is, the number of words in the lexicon divided by the number of classes) is no longer useful; neither is the concept of the percentage of singleton (unique) classes. However, the parallel notions of expected class (or set) size (the average, over words in the lexicon, of the sizes of the word sets matching their transcriptions), and the percentage of words that are given a unique transcription (that is, one that no other word in the lexicon matches, as evidenced by a singleton word set), are still appropriate, as is, we will see later, the concept of "percentage of information extracted" (Carter, 1987).

Altmann, as we saw above, presents figures for "mean size of equivalence classes". This appears, from a comparison of his figures with the histogram in his figure (1), indeed to be a mean, rather than an expected, size, which is undefined for an assign-and-lookup experiment.

4 REPEATING THE EXPERIMENTS

The obvious question, then, is whether the procedure Altmann used was equivalent to assign-and-count or to assign-and-lookup, and if, as his use of mean classes sizes suggests, it was the former, whether this makes any substantial difference to the results.

To test this, we re-ran his experiments (1) to (4) under both methods, measuring the percentage of the lexicon uniquely identified (this being a statistic that Altmann uses and that is definable under both methods). The lexicon used was derived from the Longman Dictionary of Contemporary English.
with the word frequencies of Kucera and Francis (1967), as described in Carter (1987). The results of our experiments are shown in table 2. In our experiment (4), the choice of whether to transcribe fine segments 1,4,7..., 2,5,8... or 3,6,9... in each word was made at random. Space reasons preclude a full discussion here of the reliability of our results or their comparability with Altmann’s, but there are good reasons to be confident of both: we averaged experiments (3) and (4), with their random element, over 100 runs, and the differences between lexicons are unlikely to have much effect on the differences between figures for different experiments (as opposed to the figures themselves).

<table>
<thead>
<tr>
<th>Lexicon size</th>
<th>Altman’s figures</th>
<th>Assign and count</th>
<th>Assign and lookup</th>
</tr>
</thead>
<tbody>
<tr>
<td>18500</td>
<td>62%</td>
<td>69.2%</td>
<td>69.2%</td>
</tr>
<tr>
<td>12850</td>
<td>75%</td>
<td>79.6%</td>
<td>79.2%</td>
</tr>
<tr>
<td>12850</td>
<td>91.5%</td>
<td>93.0%</td>
<td>78.7%</td>
</tr>
<tr>
<td>12850</td>
<td>[(2)&lt;(4)&lt;(3)]</td>
<td>88.4%</td>
<td>78.7%</td>
</tr>
</tbody>
</table>

Table 2: Altmann’s experiments repeated

Experiments (1) and (2) confirm Altmann’s finding that a fine transcription of stressed segments gives better, but not enormously better, discrimination than an all-mid-class transcription. But it is clear that the superiority in Altmann’s results of (3) to (4), and of both to (2), is reflected only in the (inappropriate) assign-and-count figures; in the (appropriate) assign-and-lookup ones, no significant difference appears. In other words, it appears that Altmann’s central claim, that transcribing a random one-third of segments fine is more discriminating than transcribing stressed segments fine, is based on an inappropriate experimental procedure.

The assign-and-lookup figures, unlike the assign-and-count ones or Altmann’s results, are also consistent with the expected effects of phonotactic correlations. The similarity between the values for (2), (3) and (4) suggests that the greater intrinsic informativeness of stressed segments (transcribed fine consistently only in (2)) is more or less offset by the effect of correlations between fine-transcribed neighbours, which are more numerous in (2) than in (3) or (4).

5 AN ALTERNATIVE MEASURE

We have now shown that Altmann’s figures cannot be relied on, because of his apparent use of assign-and-count. But we have not in fact shown his conclusions to be wrong, because even though our use of the assign-and-lookup method is valid, there are several reasons why the uniqueness figures in the rightmost column of table 2 are not the best guide to the effectiveness of the various transcriptions.

Firstly, as Altmann indeed recognises, word frequencies should be taken into account. Secondly, it is not sensible to assume, as Altmann does and as we have done so far, that monosyllabic words are always unstressed; it is more realistic to assume that, for careful continuous speech, content words will be stressed and function words unstressed. Under these two assumptions, the (frequency weighted) proportion of segments in stressed syllables becomes about 42.4%, or 3/7. Thus experiment (3) should transcribe a random 42.4% of segments fine, and in experiment (4), the effect of even spacing for this proportion of fine segments can be achieved by transcribing fine segments in a word, for integer n, where k is selected at random from {0, 1, 2, 3, 4, 5, 6} for each word.

Furthermore, it is argued in Carter (1987) that neither uniqueness values nor expected class size figures give the best indication of the informativeness of a transcription, even when weighted by word frequency; rather, what is needed is a measure of the percentage of information extracted (PIE) by each transcription, where “information” is used in the strict information-theoretic sense to mean the complement of uncertainty or entropy.

The PIE for a transcription can be defined, even in the absence of equivalence classes, in terms of the expected uncertainty once the set of words corresponding to the symbol sequence has been retrieved:

\[
P_{\text{IE}}(L) = \left(1 - \frac{\sum_{\omega_i \in L} \sum_{\omega_j \in S_i} q_{ij} \log q_{ij}}{\sum_{\omega_i \in L} p_i \log p_i} \right) \times 100\%
\]

where, for a lexicon \(L\), \(p_i\) is the probability of word \(\omega_i\) occurring (i.e. its proportional frequency); \(S_i\) is
the word set retrieved from the symbol sequence for \( \omega_i \) and \( q_{ij} \) is the probability of word \( \omega_j \) (in \( S_i \)) occurring once we know that a word in \( S_i \) has occurred:

\[
q_{ij} = \frac{p_j}{\sum_{\omega_k \in S_i} p_k}
\]

The PIE values for the new set of experiments were as shown in table 3.

<table>
<thead>
<tr>
<th>Segments transcribed fine (as opposed to mid)</th>
<th>PIE values</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) None</td>
<td>94.3%</td>
</tr>
<tr>
<td>(2) Stressed</td>
<td>97.3%</td>
</tr>
<tr>
<td>(3') Random 42.4%</td>
<td>96.7%</td>
</tr>
<tr>
<td>(4') Segments ( 7n+k+0,2,4 )</td>
<td>97.0%</td>
</tr>
</tbody>
</table>

Table 3: Frequency-weighted PIE values with monosyllabic content words treated as stressed

These results in fact have a similar pattern to those in table 2. They show that, if word stress and word frequencies are taken properly into account, there is no evidence that a fine transcription of random segments, or of evenly spaced segments, gives a more informative result than one of stressed-syllable segments. On the contrary, it is the stressed case, (2), that yields a slightly superior result, leaving only 2.7% of the information to be extracted, as against 3.0% in the evenly spaced case and 3.3% in the random case. This is in spite of the fact that correlations between adjacent segments will reduce informativeness more in the stressed case than in the other two.

6 CONCLUSION

Thus Altmann's conclusion, as well as his method, is incorrect. If anything, a speech recogniser should, as Huttenlocher's work first implied, concentrate on extracting fine class information from segments in stressed syllables. Their extra intrinsic informativeness outweighs the disadvantage of phonotactic correlations between contiguous segments.

However, the fairly small size of the difference between the results for (2), (3') and (4') means that if mid class information has been extracted, there is only a small advantage in subsequently concentrating on stressed segments rather than on random or evenly-spaced ones. It may be that this advantage is so small that the recogniser should concentrate its efforts wherever the signal quality is highest, whether the segments in question are stressed or not. But if the highest-quality segments are the stressed ones, as they often will be, the recogniser should not artificially have its attention shifted from them, as Altmann's work would lead us to believe.

7 REFERENCES


8 ACKNOWLEDGEMENTS

This work was supported by research grant no. GR/D/4217.7 from the UK Science and Engineering Research Council. We are grateful to the Longman Group Ltd for kindly allowing us access to the LDOCE typesetting tape for research purposes, and to Mary Smith for helpful comments on a fuller version of this paper, copies of which are available from the first author.