Proceedings of the Institute of Acoustics

A STATISTICAL ANALYSIS OF SEGMENTAL DURATIONS USING AUTOMATICALLY-SEGMENTED DATA

B. Williams & D. McKelvie

Centre for Speech Technology Research, 80 South Bridge, Edinburgh EH1 1HN, Scotland

1. INTRODUCTION

Many studies of segmental duration using hand-segmented data have shown differences in segmental duration according to linguistic context (see [1] for a review of such studies). Effects such as sentence-final and phrase-final lengthening of segments have been discovered, with possible foot-level shortening of vowels. However, as these studies were carried out on data segmented by hand, it has not been proven that the results are applicable to speech recognition, where the segmentation must be carried out automatically. Therefore a study of segment durations has been carried out on data segmented by the (HMM-based) front end of a continuous speech recogniser. If it can be shown that there is any durational variation in this data according to linguistic context, then it can be argued that speech recognisers need to take linguistic context into account in the training of segment models. This would reduce the durational variance of the models, and probably increase their accuracy.

2. DATA PREPARATION

2.1 Data used

The data used was a list of 200 phrases and sentences taken from the domain of cytopathology laboratory reports. This kind of data is typical of the kind of task that a speech recogniser might be required to perform. Examples of the data are as follows:

(1) These very cellular smears contain large cohesive sheets of epithelial cells.
(2) Seven millilitres of orange fluid.

The sentences were recorded by an adult male speaker of RP English in an acoustically-treated room. The digitised speech then served as input to the front end of a continuous speech recogniser. The units recognised by the front end were largely phonemes, together with separate models for some clusters of obstruent plus approximant (e.g. /tr/) and for clear and dark allophones of /l/ (see [2] for details of the HMM-based front end). The output of the front end was a phoneme lattice. This served as input to the lexical access module of the speech recognition system.

2.2 Automatic syllabification of the lexicon
The lexicon used by the lexical access routine contained hand-produced phonemic transcriptions, with primary and secondary word-level stress. An algorithm was written to insert syllable boundaries in polysyllabic words (word boundaries were taken to force a syllable boundary). This algorithm encoded the phonotactic constraints of English. Where more than one syllabification was permitted by these constraints, then as many consonants as possible were taken to begin the next syllable. An exception was made in the case of /s/, which was syllabified with a preceding stressed vowel. This algorithm is essentially that outlined in [3] (chapter 2.2).

2.3 Phrase boundary marking in the corpus
Phrase boundaries were marked in the corpus by hand, but algorithmically. Due to the nature of the particular data used, it was found that each syntactic phrase ended in a noun, and there was no noun which did not end a syntactic phrase. Since an automatic phrase-level parsing algorithm could be expected to attain at least this level of accuracy, given such restricted data, it is reasonable to assume that the syntactic boundaries recognised (utterance, phrase, and word boundaries) are equivalent to automatically-assigned boundaries. This means that any results found would be valid for a complete speech recognition system.

2.4 From phoneme lattices to word strings
The phoneme lattice for each of the 200 utterances formed the input to the lexical access module, together with the syllabified lexicon and the phrase-marked text of each utterance (see [2] and [4] for details of lexical access). The lexical access module determined the best parsing into words against the phoneme lattice, and was functioning in training mode (where the text of the utterance is known). The output was a word lattice marked with syllable and phrase boundaries. Where a lexical phoneme did not appear in the phoneme lattice, the phoneme was ignored (since there was in such a case no duration to measure). Where the phoneme lattice contained surplus segments when compared against the lexicon, each surplus acoustic segment was merged with whichever of its flanking segments showed the greatest similarity to it in terms of lexical access costs (equivalent to phonetic similarity). This process generally resulted in surplus consonantal segments being assigned to consonant phonemes, and vocalic segments to vowels, but some exceptions occurred (eg. surface (acoustic) /oo/ (half-open rounded back tense vowel) was assigned to lexical /l/ in some cases).

2.5 Formation of the data matrix

The output word lattice, after the adjustments described above, was then transformed into a data matrix of one row per (lexical) segment. Each segment was associated with the following information:

3) Lexical segment (one of 44 phonemes).
4) Duration in milliseconds.
5) Stress level of syllable containing segment (primary, secondary, or unstressed).
6) Position in syllable (first, second, etc. segment).
7) Number of segments in relevant syllable.
8) Position of relevant syllable in word (initial, medial or final in polysyllabic word; or monosyllabic word).
9) Position of relevant syllable in phrase (initial, medial or final).
10) Position of relevant syllable in utterance (initial, medial or final).

This data matrix formed the input to the statistical analysis described below.

3. STATISTICAL ANALYSIS

3.1 Form of analysis

The data yielded a total of 4357 segments. These were broken down into subsets according to the parameters described above (eg. one subset contained all cases of lexical /s/ that were syllable-initial and word-initial but not utterance-initial or phrase-initial). The mean and standard deviation of the segment duration in each subset was determined. An approximate check on the shape of the distribution in each relevant subset revealed that the distribution of most of them was reasonably close to Gaussian. T-tests were then carried out on various pairs of subsets. Many subsets contained too few cases to yield statistically significant results. The results given here are all significant to at least the 5% level (many are significant to the 1% level). In each case, a two-tailed test was carried out: since it was the directionality (as well as size) of any duration difference which was under examination, a one-tailed test would have assumed the very facts that it was hoped to prove.

3.2 Results

The results are presented separately below for vowels and for consonants.

3.2.1 Vowels. In the case of (primary and secondary) stressed tense vowels (ie. diphthongs and long monophthongs), utterance-final, phrase-final and word-final lengthening effects were found for polysyllabic words. Fig. 1 shows boxplots of the durations of such vowels in the following contexts:

11) In utterance-final syllables in polysyllables.
12) In phrase-final syllables in polysyllables which are not also utterance-final.
13) In word-final syllables in polysyllables which are not also utterance-final.
14) In word-initial syllables in polysyllables.

Each box shows the median and inter-quartile range of the duration, while the whiskers indicate the extent of the data up to 1.5 times the inter-quartile range. The relevant statistics are given below.

15) Utterance-final: mean=240.00 ms, s.d.=169.99 ms, n=18.
16) Phrase-final: mean=167.50 ms, s.d.=29.52 ms, n=8.
17) Word-final: mean=130.98 ms, s.d.=37.38 ms, n=46.
18) Word-initial: mean=102.62 ms, s.d.=41.10 ms, n=145.

The durations in 15 are significantly different from those in 16 to the 5% level. In this case, due to the paucity of data, a two-tailed Mann-Whitney U-test was used: this is a non-parametric test based on rank ordering. The durations in 17 are significantly different from those in 16 to the 5% level (t-test), while those in 18 are significantly different from those in 17 to the 1% level (t-test).

In the case of stressed short vowels in monosyllabic words, the effects found were those of utterance-final lengthening ($p<0.05$) and phrase-final lengthening ($p<0.01$). The statistics were as follows.

19) Utterance-final: mean=148.65 ms, s.d.=46.82 ms, n=52.
20) Phrase-final, non-utterance-final: mean=126.56 ms, s.d.=40.43 ms, n=32.
21) Non-phrase-final monosyllables: mean=95.47, s.d.=31.32, n=32.

3.2.2 Consonants. The consonant giving the largest number of significant results was /s/, especially in syllable-initial position: because of the paucity of data at this level of analysis, far fewer significant results could be obtained for other consonants. Syllable-initial lexical /s/ showed both utterance-initial lengthening and word-initial lengthening. Fig. 2 shows boxplots for the duration of syllable-initial /s/ in these contexts. The relevant statistics are as follows.

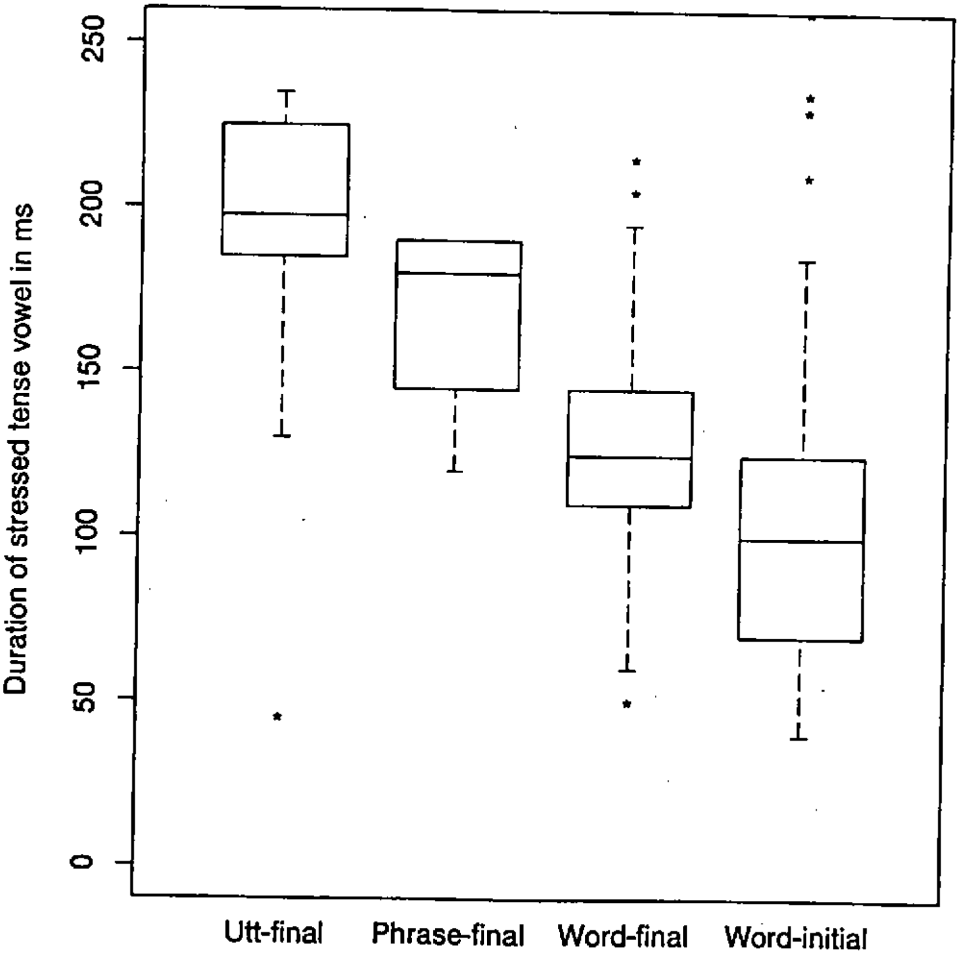

Fig. 1: Duration of stressed tense vowels in polysyllabic words, by syllable context.

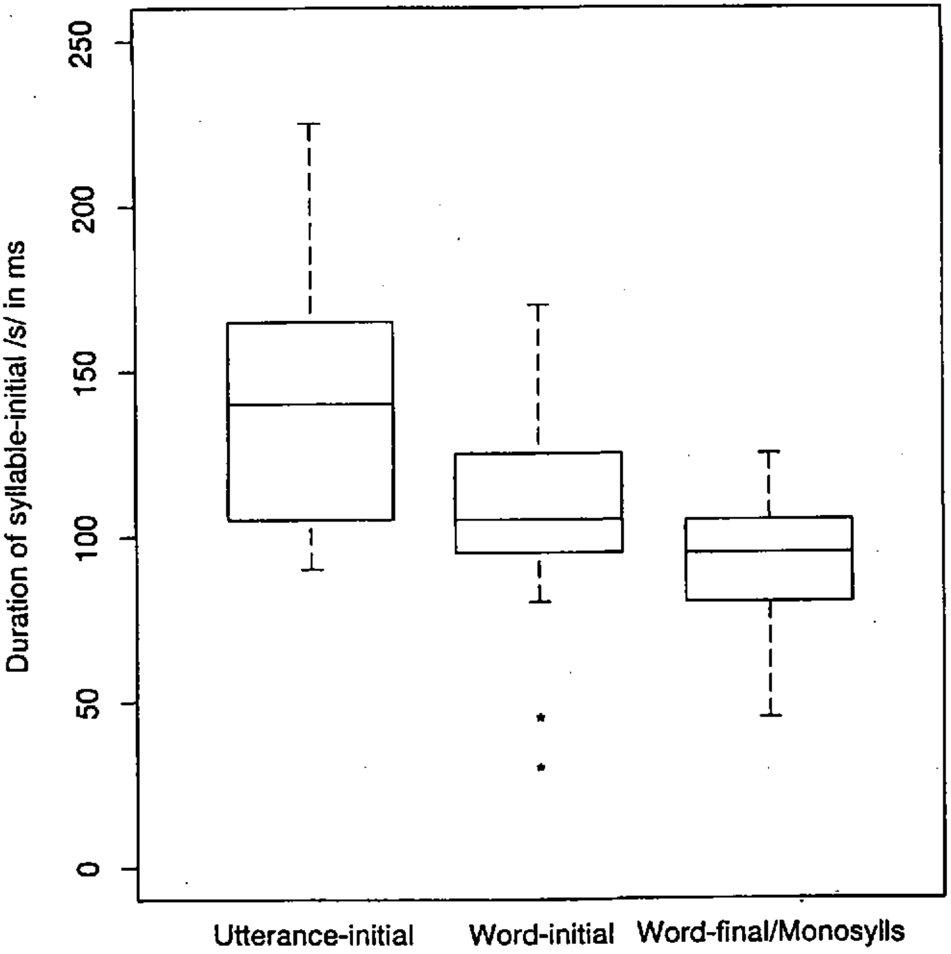

Fig. 2: Duration of syllable-initial /s/, by syllable context.

22) Utterance-initial syllables: mean=174.77 ms, s.d.=110.06 ms, n=22.
23) Word-initial syllables (in polysyllables) that were not initial in utterance: mean=106.92 ms, s.d=26.35 ms, n=39.
24) Word-final syllables of polysyllables; also monosyllables; mean=94.48 ms, s.d.=17.39, n=29.

The durations in 23 are significantly different from those in 22 to the 0.1% level (t-test), while those in 24 are significantly different from those in 23 to the 5% level (t-test). The durations of syllable-initial /s/ in monosyllables and in word-final syllables of polysyllables did not differ significantly, and so these contexts have been treated as one.

4. CONCLUSION

Some degree of utterance-final, phrase-final and word-final lengthening has been demonstrated for at least some vowels. This agrees with results from studies using hand-segmented data (see [1]). In addition, a degree of utterance-initial and word-initial lengthening has also been demonstrated for syllable-initial /s/: while this may be specific to /s/, it nonetheless demonstrates that linguistic boundaries are relevant to consonantal duration.

Because the results concern only those subsets that yielded significant results, it might seem that the phenomena found are arbitrary and limited in scope to a few very specific linguistic contexts. However, in many cases there were appropriate trends in similar subsets that could not be shown to be statistically significant. It is hoped that a repetition of the analysis with a larger amount of data will be able to show that the results found are by no means as isolated as might seem at first sight.

The results show that linguistic context, at several levels, has an effect on the duration of vowels and consonants even in automatically-segmented data. This indicates that an HMM-based speech recogniser might achieve even higher accuracy if linguistic context were taken into account in the training of the segmental models. Further research is necessary to establish the extent of these effects.

5. REFERENCES

[1] D.H. KLATT, 'Linguistic uses of segmental duration in English: Acoustic and perceptual evidence', *Journal of the Acoustical Society of America*, vol 59, pp. 1208-1221 (1976).

[2] F.R. MCINNES, D. MCKELVIE & S.M. HILLER, 'The structure and performance of a modular continuous speech recognition system', this volume.

[3] E. FUDGE, '*English Word-Stress*', London: Allen & Unwin, 1984.

[4] H.S. THOMPSON, D. MCKELVIE & F.R. MCINNES, 'Robust Lexical Access for Continuous Speech Using Dynamic Time Warping and Finite-State Transducers', *Proceedings of the European Conference on Speech Communication and Technology*, Paris, September 1989, vol. 2, pp. 59-62.

6. ACKNOWLEDGEMENTS

This work was supported by the Information Engineering Directorate/Science and Engineering Research Council as part of the IED/SERC Large Scale Integrated Speech Technology Demonstrator Project (SERC grants D/29604, D/29611, D/29628, F/10309, F/10316) in collaboration with Marconi Speech and Information Systems and Loughborough University of Technology.