Speech Communication 13 (1993) SPC00646 S
North-Holland

Automatic detection of prosodic boundaries in speech

Nick Campbell

Advanced Telecommunications Research Institute, Interpreting Telephony Research Laboratories, Kyoto 619-02, Japan

Received 7 January 1993
Revised 16 June 1993

Abstract. This paper describes a method for automatic annotation of prosodic events in speech, using segmental duration information. It details a way of differentiating prominence-related lengthening from boundary-related lengthening, using durational clues alone, and discusses an anomaly in the phrasing characteristics of four speakers' readings of 200 phonetically-balanced sentences. An algorithm is described that uses syllable-level differences in normalised segmental duration measures to detect prosodic boundaries in a speech signal. Tests with read-speech data from four British-English RP speakers show high agreement between speakers with respect to the number of boundaries detected and the length of the phrases delimited by each pair of boundaries, but the correlation between speakers on actual boundary locations is low. There is particular disagreement between speakers in the case of a single function word linking two groups of content words. This discrepancy can be resolved if the boundary is taken to be at the function word location itself, rather than at one or other side of the word. These results are taken to indicate some freedom in the placement of prosodic boundaries in such cases, sometimes being cued by a syntactic boundary, and sometimes by a rhythmic one.

Zusammenfassung. Dieser Artikel beschreibt eine Methode, die eine automatische Notierung der Betonung in der Sprache unter Verwendung von Informationen über die Teilzeiten ermöglicht. Er beschreibt detailliert einen Weg zur Unterscheidung der prominenzbezogenen Längung von der grenzbezogenen Längung, unter Verwendung der Angaben über die Dauer und berichtet über eine Anomalie in den Satzdaten von 4 Sprechern, die 200 phonetisch ausgeglichene Sätze lesen. Weiterhin wird ein Algorithmus beschrieben, der die Unterschiede in der Silbenbetonung in standardisierten Zeitabschnittsmessungen verwendet, um die Grenzen der Betonung in einem Sprachsignal zu erkennen. Tests mit von vier englischen Sprechern abgelesenen Reden zeigen eine hohe Übereinstimmung zwischen den Rednern bezüglich der Anzahl der erfaßten Grenzen und der Lange der Satze, die von jedem Grenzpaar abgegrenzt werden, aber die Korrelation über die aktuellen Lagen der Grenzen ist ziemlich gering. Zwischen den Rednern besteht ein besonderer Unterschied bei einzelnen Worten, die zwei zusammenhangende Wortgruppen verbinden. Dieses Problem kann gelöst werden, indem man die Grenze direkt auf das einzelne Funktionswort legt anstatt auf die eine oder andere Seite dieses Wortes. Diese Ergebnisse zeigen eine gewiße Freiheit in der Placierung der Betonungsgrenzen in solchen Fällen, die manchmal durch eine Silbengrenze und manchmal durch eine rythmische Grenze definiert werden.

Résumé. Cet article décrit une méthode permettant un étiquettage automatique des évènements prosodiques de la parole, à partir de l'information foumie par les durées segmentales. Il précise une façon de différencier, à partir des seuls indices de durée, les allongements dus à la prominence de ceux dus à la présence d'une frontière, et expose une anomalie trouvée dans le découpage syntagmatique effectué par 4 locuteurs lisant 200 phrases phonétiquement équilibrées. On décrit un algorithme qui ublise les différences de durée normalisée au niveau syllabique pour détecter les frontières prosodiques dans led signal de parole. Des tests effectués sur des données de parole lue émanant de 4 locuteurs, anglais-britanniques montrent une forte concordance inter-locuteur en ce qui concerne le nombre de frontières détectées et la longueur des syntagmes délimités par chaque paire de frontière, mais la corrélation inter-locuteur sur la localisation effective des frontières est faible. On observe en particulier une difference nette, entre locuteurs, dans le cas de mots fonctionnels uniques liant deux groupes de mots lexicaux. Ce Problème peut être résolu si l'on considère que la frontière est sur la position du mot fonctionnel lui même plutôt que à gauche ou à droite du mot lexical. Ces résultats semblent montrer qu'il

0167-6393/93/$06.00 © 1993 – Elsevier Science Publishers B.V. All rights reserved

existe, dans ce cas, une certaine liberté dans la localisation des frontières prosodiques, qui peuvent êre déterminées soit par la frontière syntaxique, soit par des critères rythmiques.

Keywords. Speech-segmentation; duration; syllables; phrasing.

1. Introduction

This paper describes a method whereby segmental duration information can be used to indicate prosodic boundaries in the speech signal to enable chunking together of related words into phrases delimited by these boundaries. This chunking is an essential preliminary to semantic processing for language understanding systems and for the description of segmental and phrasal contexts for speech-source labelling for concatenative speech synthesis. The relation between segmental lengthening and proximity to a prosodic boundary has long been known (Gaitenby, 1965; Klatt, 1975; Scott, 1982) but detection of such lengthening has been complicated by the different "inherent" durations of segments and the different, often interacting, causes of segmental lengthening.

Recent advances in speech technology, requiring the collection of large speech corpora for analysis and training material, have placed an increased emphasis on the annotation of speech. There is at the same time growing international agreement on a set of standards for the transcription of prosody (Silverman et al., 1992, TOBI) for which automatic or semi-automatic procedures are now being researched (Wightman et al., 1992; Wightman and Ostendorf, 1993). In order to provide useful databases for speech analysis, large volumes of natural speech must be both prosodically and segmentally labelled and annotated, and although this work can perhaps be adequately performed by human labellers trained to detect relevent events in the speech signal, it is expensive, time-consuming and unreliable. Furthermore, if the transcription conventions are revised at any time, the labour is wasted and the annotation process must be repeated. Some automation of the process is therefore considered necessary.

The segmental durations used in the following experiments were obtained semi-automatically by use of hidden Markov models to label and segment the speech corpus. The models were trained on a small number of hand-segmented phones and constrained by phoneme labels generated from an orthographic transcription of the speech (Edwards et al., 1992). Inter-labeller tests of manual segmentation consistency have shown end-points for 50% of the labels to be within 5 msec, and 90% within 25 msec. HMM segmentation yielded results of 50% within 12 msec, and 90% within 30 msec (Schmidt and Watson, 1991). Automatic segmentation is not as accurate as hand segmentation, but it is perhaps more consistent in the location of its inaccuracies, and can be used equivalently.

The following sections show that when these raw measures of segmental duration are normalised to factor out the effects of phonemic differences, the prosodic lengthening patterns become clear. If this lengthening is viewed in the context of a syllable framework, the effects of different causes of lengthening can be distinguished. A test using syllable-level analysis-by-synthesis of segmental durations shows that the effects of stress and pre-boundary lengthening in particular can be distinguished. The final section of this paper presents a prosodic boundary detection algorithm based on this differential, and describes the results of a test applying it to the speech data of four British-English RP speakers reading a set of 200 phonemically-balanced sentences.

2. Segmental lengthening

There is, of course, a high degree of interaction between the duration, fundamental frequency and energy variations that signal prosodic events in speech, but in this paper we will concentrate on the extent to which normalised measures of segmental duration alone can be used to determine the phrasing of an utterance. This will enable us to determine the usefulness of the segmental information, which is inherent in the labelling and can therefore be considered a cheap resource not requiring further access to the speech waveform or special signal processing.

Differences in global speaking rate will clearly have an influence on the length of the component segments, as will any local compression or expansion of segments that results from accommodation into a rhythmic framework. We will assume here, though, that the effects of this type of lengthening are uniform across all segments in a syllable and look more closely at two other causes of lengthening instead. Durational information is significant in the encoding of two aspects of prosodic structure; marking prominence and marking boundaries, but simple measures of segmental lengthening fail to distinguish between the two. Articulatory data from Edwards and Beckman (1988), derived from jaw movements, has revealed amplitude differences that suggest that the lengthening "profile" throughout a syllable should be different for the two cases, and an analysis of the segmental durations in the two cases supports this view (Campbell, 1989). The following experiment shows that when viewed within the framework of the syllable, the lengthening of segments tends to be more pronounced on initial (onset) segments in prominent syllables, and on later (coda) segments in preboundary syllables.

Because different articulatory gestures produce sounds with different durational characteristics, some normalisation is required before these effects become clear. One very simple way to normalise durations is to assume a Gaussian distribution and express the durations in terms of deviation from the mean determined for each segment type in units of the standard deviation of the distribution of all tokens of that type. This produces a unitless number typically in the range of ±3 that expresses the *lengthening* of the segment and filters out any phone-specific durational characteristics. Another transform that has been found effective is the two-parameter Gamma, which can be optimised by maximum likelihood estimation (Crystal and House, 1986; Levinson, 1986). This allows a slightly better fit to the individual distributional characteristics of different phones by modelling the skew separately instead of assuming a symmetrical (Gaussian) distribution, but for our present needs the choice of distribution type seems to make little difference. A transform into the log domain renders the majority of segmental duration distributions sufficiently close to Gaussian.

By converting each segment's duration into a measure of its relative lengthening, we can see the effects of prosodic influence on the signal without interference from the segmental aspects and can more easily discern the structuring. It is of particular interest to view this structuring from the level of the syllable.

3. Differential lengthening within the syllable

It has been shown that much of the variance in segmental duration can be predicted from the syllable duration under an assumption of elasticity (Campbell and Isard, 1991). The strong form of the elasticity hypothesis accounts for the phonetic aspects of segmental duration by assuming that each segment in a syllable is lengthened equivalently, in terms of its distribution, to accommodate to the duration determined for the syllable as a whole by its prosodic environment. That is, for any given syllable, there should be a number k of standard deviations such that the length of any segment in the syllable is equal to $\mu_{\text{seg}} + k\sigma_{\text{seg}}$, where μ_{seg} and σ_{seg} are the mean and standard deviation, respectively, of durations of the particular segment type. The phonetic or articulatory constraints on segmental length are thus modelled by the individual distributions, leaving the prosodic causes of lengthening to be accounted for by higher-level factors. Weaker forms of the hypothesis take into account the fact that lengthening profiles can be different for different classes of phone in different prosodic contexts. If all types of lengthening were equivalent in their effects, then the strong form of the elasticity hypothesis would apply. Because they differ, the weaker forms are required but use can be made of the differences in identifying the lengthening type and, by implication, the phrase-final syllables and the prosodic phrase boundaries.

To illustrate this difference, a test was performed using the strong form of the elasticity hypothesis to predict segmental durations from known syllable durations in a corpus of speech. A comparison was then performed between the predicted segmental durations and those observed in the original data. Systematic differences between the observed and predicted durations show where other factors than simple accommodation to the syllable length are having an effect.

3.1. Procedure

Syllable durations in a 200-sentence phonetically-balanced corpus (see Appendix A for examples) were calculated by summing component segmental durations, and then an appropriate value of k was determined for each so that the durations of the component phones, predicted as below by solving equation (1), would sum to the known syllable durations.

Log-transformation of the segmental durations was applied to minimise the positive skew seen in the data. Applied to the log-transformed data, a value for the factor k was determined by solving the equation

$$D = \sum_{i=1}^{n} \exp(\mu_i + k\sigma_i),$$

where D is the total duration of all segments in a given syllable, n is the number of segments in the syllable, and μ_i is the mean and σ_i the standard deviation of the log transform of measurements for all the tokens in the database corresponding to segment i.

To measure the difference between preboundary lengthening and prominence-related lengthening, the sentences were labelled for degrees of each. Four levels of prominence, or degree of stress, were determined subjectively by listening to the readings of the four speakers, and the data were labelled accordingly:

1: none (unstressed syllables);
2: secondary (similar to secondary lexical stress);
3: primary (similar to primary lexical stress);
4: sentence or phrasal stress.

Four levels of boundary were similarly determined, using a simplified form of break indices as in the TOBI system of analysis, to indicate the degree of prosodic discontinuity between each pair of syllables in the readings:

1: word-medial syllables and clitics;
2: phrase-medial word boundaries;
3: intonation phrase boundary or strong disjuncture marked by a pause or virtual pause;

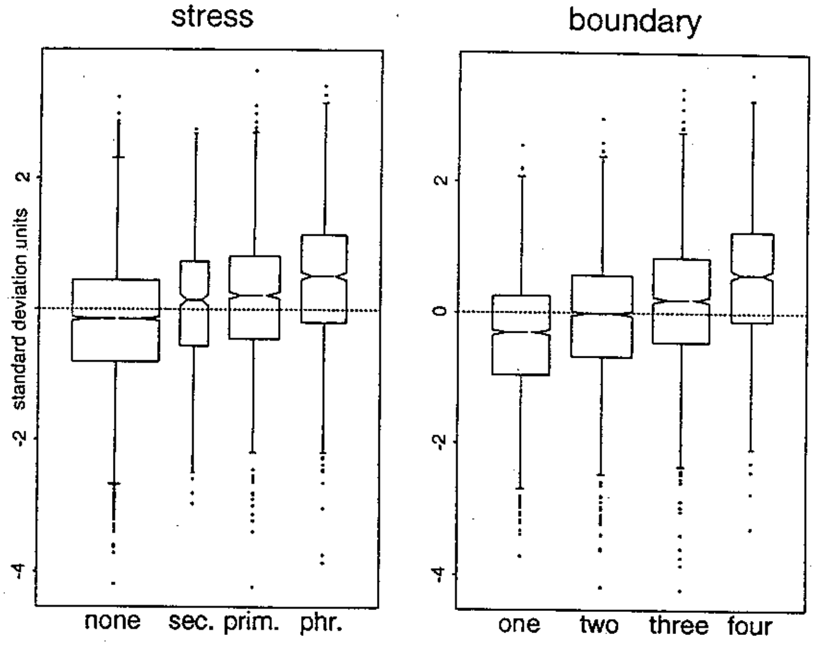

Fig. 1. Normalised duration scores showing lengthening in four classes of prominence and preboundary position.

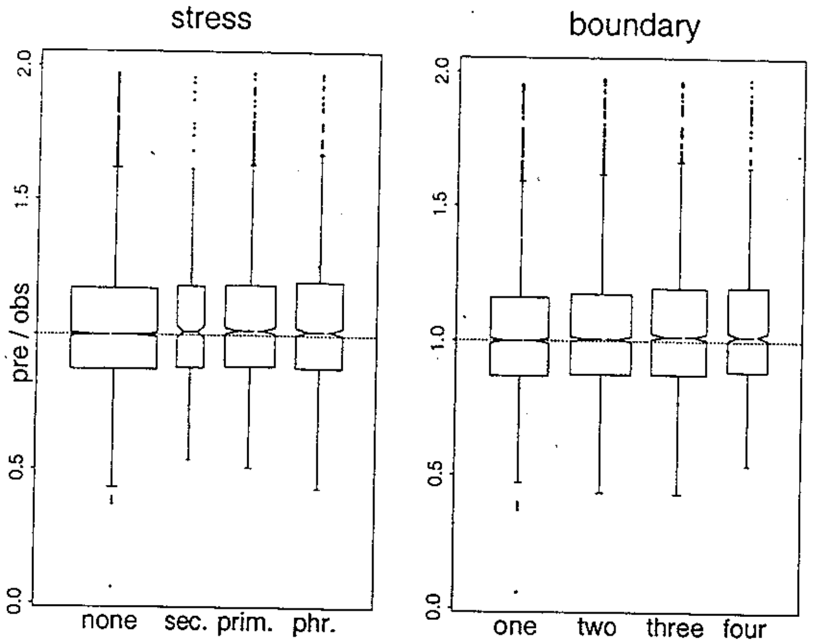

Fig. 2. Measures of fit (predicted/observed) after a prediction of segmental durations from the syllable duration assuming strong elasticity. Classes and data are as for Figure 1.

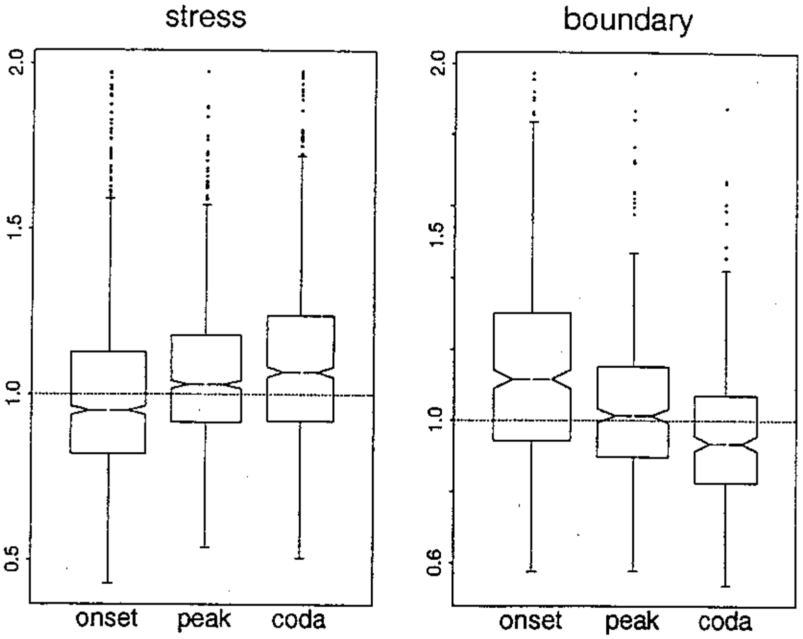

Fig. 3. The same data as in the previous figure, but factored this time by position of the segment in the syllable. We can see that onset segments are overpredicted in pre-boundary syllables, and that coda segments are overpredicted in prominent (stressed) syllables.

4: end-of-utterance, full intonation phrase boundary.

Figure 1 shows boxplots of the normalised durations (z-scores) of the segments in these contexts. The horizontal edges of the boxes mark the 25th and 75th percentiles of the distributions. Notches on the median line indicate significance at the 5% level in the difference of the distributions when there is no overlap. The figure shows no difference in lengthening between primary and secondary levels of stress, but confirms clear and increasing differences in the lengthening of syllables bearing lexical and phrasal stress. All four levels of boundary strength show clear durational correlates.

3.2. Results

After prediction, a mean fit of 1.0 was obtained, measured by calculating predicted/observed durations. This is to be expected since the method ensures that the total durations will be the same, but of more interest is how any error in prediction is distributed amongst the component subcategories of phone. Figure 2 shows the second and third quartiles of the distributions to be approximately 0.80 and 1.15, indicating that 50% of the results fell within this range. A chi-square value of 0.378 (df = 19) shows the distribution of error not to be significantly different from normal. Comparison with Figure 1 shows that the gross lengthening differences between classes of prominence and boundary lengthening have been accounted for, but although the boxes now appear to be level, there remains a lot of variation about the zero mean still to be explained.

Comparison of the fit factored into consonant and vowel components separately showed no significant difference in prediction accuracy, but when we further subcategorise the consonants into those that are in onset position and those that are in coda position in the syllable a distinction becomes clear. Figure 3 presents the same data as Figure 2, but factored this time into onset, peak and coda segments; it shows that much of the prediction error (variability) in the earlier figure can be accounted for by the differential effects of stress and finality on lengthening within the syllable. Predicted durations for onset consonants were typically less than observed in the case of stressed syllables, and those for coda segments greater than observed. This confirms that onset consonants are typically lengthened more than coda ones within a syllable when it is lengthened by prominence, and shows the reverse to be the case for pre-boundary lengthening.

3.3. Discussion

The strong form of the elasticity hypothesis overpredicts the lengthening of coda segments in prominent syllables and onset segments in pre-boundary syllables. Weaker forms are shown to be required to account for these different types of lengthening separately.

These results are in accordance with those of Edwards and Beckman from articulatory data of jaw movements, and lead to the conclusion that an automatic algorithm for spotting prosodic events should be able to distinguish phones that are lengthened in phrase-final position from those that are lengthened by stress, by consideration of the phone's position in the syllable in conjunction with the lengthening profile for that syllable. In this way, it should be possible to locate and differentiate both phrase boundaries and stressed syllables from duration measurements alone.

4. Locating prosodic boundaries

Having shown that differential lengthening takes place on segments within the syllable under the two prosodic conditions, we can now determine the extent to which "slope" of lengthening through a syllable can be used to detect prosodic boundaries. The issue of prominence will not be addressed further in this paper.

A program was written that calculates the profile of lengthening within a syllable, comparing the length of each phone with that of its immediately preceding neighbour to determine if the lengthening is increasing, which would indicate proximity to a boundary, or falling, which would indicate prominence. This "slope" was used to differentiate between the two types of lengthened syllables to indicate potential prosodic boundaries in the 200 sentences.

It should be noted that although we were able to distinguish four levels of boundary in the data, no attempt will be made to distinguish between them here. The algorithm produces a binary decision, triggered by a simple reset in the lengthening profile. It will be of interest to see just what constitutes a boundary in these terms, and a large part of the later discussion will focus on how we judge the correctness of such decisions.

4.1. *Differential lengthening of syllables*

In order to determine the lengthening differential within a syllable, segmental durations were first normalised per phone type by subtracting the type mean and expressing the residual in terms of the type variance (*z*-score normalisation). This step in the process removes the effects of phone-related durational differences from consideration, leaving only the higher-level prosodic timing effects. The first difference of these scores

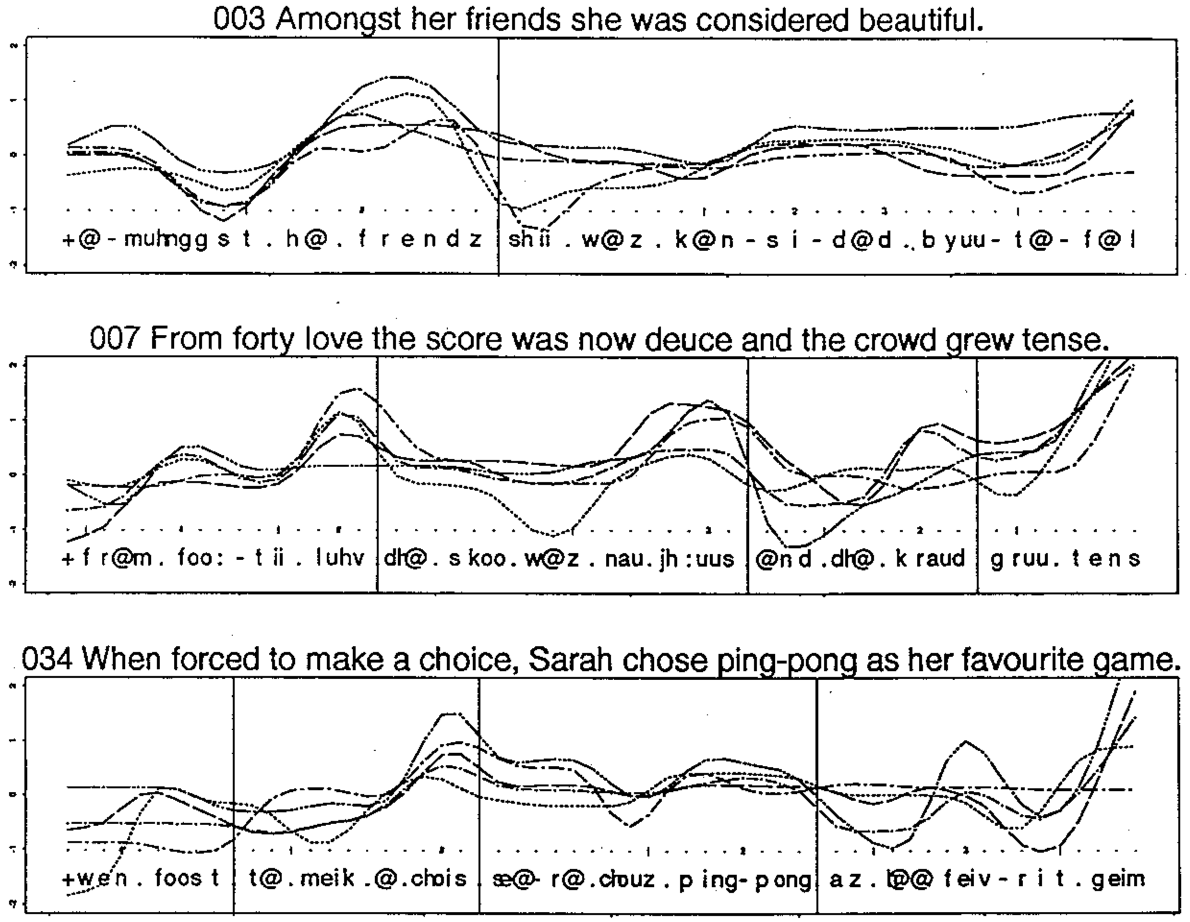

Fig. 4. Plots of normalised segmental durations for four speakers and three sentences, showing a high degree of similarity in the lengthening contours. The *y*-axis shows normalised segmental lengthening in SD units. (Raw scores have been smoothed for these plots to allow clearer comparison). Vertical lines indicate prosodic phrase boundaries as determined by the algorithm. The fifth line (dot-dash) indicates the mean profile of lengthening.

can be expected in the general case to increase throughout the syllable with phrase-final lengthening, and to decrease with stressed lengthening. Location of a downward reset in an increasing syllable-level lengthening profile can therefore be taken to indicate a prosodic boundary typical of a phrase-final or sentence-final position.

The algorithm to perform the segmentation thus has two components; the first calculates the slope of lengthening within each syllable by differentiation of the lengthening values, and the second compares the degree of slope between each pair of syllables and indicates a break when an increase in this slope is reset.

The differentiation reduces the effects of stress lengthening, yielding negative values from decreasing slope of lengthening throughout the syllable, and accumulates positive values for increasing slope typically found in phrase-final position. Local maxima in the differential indicate a reset in the slope and trigger a boundary decision. A limit requiring more than one syllable per phrase prevents the algorithm from over-generating boundary hypotheses; in the exceptional case where a phrase does actually consist of one single syllable, it is hoped that other stronger cues to boundary status will be present.

```
1. syllable_count = 0
2. for (each segment in the syllable)
      sum + = this_z_score-last_z_score
   slope = sum/number_of_segs_in_syll
3. syllable_count + = 1
   if ((last_slope > this_slope) and
      (syllable_count > 1))
   then insert prosodic boundary marker
      and reset sylable_count to 0
```

4.2. A test of the segmentation method

A test of the algorithm was performed with the durations of four speakers' readings of 200 sentences, presented as a continuous stream, with no indication of any boundaries between the sentences. Differences were calculated between each pair of durations, disregarding syllable boundaries, and then means were taken within each syllable to calculate the slope. Resets in a rising slope triggered a boundary insertion at the reset location, with the condition that there be more than one syllable in each phrase group.

The test was performed using data from readings of the 200 SCRIBE phonetically-balanced sentences (2398 words) by four speakers of British English. Input for the test was in the form of a string of labels with associated normalised durations; output was in the form of a label string with prosodic boundary markers inserted. Syllable boundaries were marked in the input string. To analyse the results of the segmentation, phone labels were aligned to orthographic words and the words grouped into phrases according to the boundary marking produced by the algorithm.

Figure 4 illustrates the degree of agreement between the lengthening profiles for the four speakers. It also shows (with vertical lines) where the algorithm inserted a boundary marker, and we can see clear instances of prominence-related lengthening that did not trigger a marker. See for example "forty" in the sentence #007, and "favourite" in sentence #034. The following examples, chosen at random, illustrate the distribution of the boundary decisions. In each sentence the number following a word indicates the number of speakers for which a boundary was inserted at that point (maximum = 4):

```
007 From forty love 4 the score was now 1
deuce 3 and the crowd 3 grew tense 4
020 She flicks 3 through a 1 magazine 3 when
she gets 3 a chance 4
026 It's strange 4 that I slept 3 for 1 so
long 3 since 1 I wasn't 2 feeling tired 4
043 Jane adored 3 Maths 1 and French 3 but
hated 1 the rest 3 of school 4
062 Water was 2 cascading 1 down 1 the moun-
tain 2 at 1 a 1 rate 1 of 1 knots 4
065 Our butcher 4 makes his own 3 pork and 1
beef sausages 4
171 Alf's brother 4 was totally absorbed 1
in 3 the virtuoso performance 4 of Bach's
Toccata 2 and 2 Fugue 4
195 I yearn 3 for the day 4 when smoking is
banned on 4 public transport 4
```

Results for all four readers showed a high degree of uniformity in the number of prosodic units determined in this way; in the 200 sentences, the number of phrases per speaker were 751, 757, 753 and 754. A high degree of consistency was also noticed between speakers with

Table 1
Phrase lengths

Number of words	1	2	3	4	5	6	7	8
Number of tokens	96	198	1453	890	280	73	9	3

respect to the number of words between each boundary. The average was 3.4 words per phrase, with a standard deviation of 0.97.

Since all four readers were reading the same set of sentences, and were familiar with the content of the sentences, it is reasonable to suppose that their readings would be similar and would display similar groupings of the words, with boundaries being inserted at identical positions. This, however, was not the case. The correlations between results for each pair of speakers averaged only $r = 0.48$, indicating that although they showed approximately the same number of boundaries in total, in many cases they were not placing them in the same locations in the text. Further analysis was therefore performed on the number of speakers for whom a boundary was inserted at each position; i.e., on the amount of agreement between speakers with respect to boundary location.

A clear positive correlation was found between the number of boundary insertions at each location and both (a) the number of closing brackets produced by a syntactic parse of the sentences ($r = 0.56$), and (b) the break-indices ($r = 0.62$). These results indicate a close agreement between the detected boundaries and the linguistic structure of the text, but as Table 2 shows, there were a large number of boundary locations where only one or two speakers' data indicated final lengthening. If this is a reliable method of boundary detection, then it shows a high degree of individuality in the phrasing.

Table 2
Boundary locations showing the number of speakers at each insertion location, and the percentage of speakers with respect to total number of insertions

all four speakers	285	38%
three speakers	217	22%
two speakers	343	22%
one speaker	542	18%
no speakers	1011	

4.3. *Individual differences in phrasing*

To clarify what is happening at the individual boundaries, we must examine the points of difference further. Many of the locations where only one or two speakers inserted a boundary occurred around a single function word that linked two groups of content words. Examples of such words are "*to*", "*that*", and "*with*" in the following excerpts: "it's difficult to choose between ...", "Tom says that ancient Saabs are ...", and "into battle with all the forces ...". Some speakers inserted a boundary before the word, linking it with the following phrase; some inserted the boundary after the word, linking it with the preceding phrase. Because there are many instances of split boundary locations around single function words, we can assume that the position of the boundary may be *at* the word, rather than before or after it, and regroup such pairs accordingly. Re-analysis after grouping together boundaries split either side of a single function word shows that the majority of disagreements are resolved in this way, yielding a much closer agreement of 73%, as shown in Table 3.

In many of the cases where ambiguous phrasing was noted, the medial word grouped more closely with the following words in terms of syntax, but with the previous words in terms of rhythm, closing the previous foot and anticipating a stress on the following word or phrase. Table 4 shows examples of such ambiphrasal words where speakers were divided in their boundary placement. The individuality in phrasing can therefore

Table 3
Revised counts of boundary insertions after resolving those split around a single linking function word

all four speakers	553	73%
three speakers	172	17%
two speakers	52	3%
one speaker	187	6%
no speakers	1434	

Table 4
Examples of ambiphrasal words; showing the numbers of speakers who lengthened the preceeding word

Two and two	The table 2 is 2 made so sloppily	He emphasized 2 his 2 strengths
	I always 2 seem 2 to follow	into battle 2 with 2 all the forces
	It's difficult 2 to 2 choose between	Mashed potatoes 2 are 2 more fattening
	The world 2 is 2 becoming increasingly	Tom says 2 that 2 ancient Saabs
	Vernon 2 helped 2 himself to dessert	Gordon's words 2 were 2 lost amidst
	We really 2 will 2 need to defrost	We need 2 to buy 2 some more
Three and one	The smell 3 of 1 the freshly	The topic 3 of 1 Jeff's thesis
	I slept 3 for 1 so long	It's a 3 shame 1 that architects
	Clara went 3 through 1 a phase	The opposition 1 claim 3 that present
	The food 3 varies 1 from place	He glimpsed 3 the 1 traffic warden
	The questionnaire 3 was 1 short	It's obvious 3 that 1 the student
	The walkers 3 took 1 a detour	It was 1 a 3 sheer fluke
	He caught 3 a 1 glimpse of	I get 1 a 3 craving for

be accounted for as a trade-off between syntax and rhythm, sometimes resulting in an unstressed function word being (syntactic) phrase-initial, as at the onset of a prepositional phrase, and sometimes being (prosodic) phrase-final, as the last syllable of a rhythmic foot.

It was not the case that one speaker was consistently delaying a boundary, as might be supposed from a simple examination of the total counts, but rather that different speakers chose different boundary points for different sentences, maintaining approximately the same spacing between boundaries and the same grouping of lexical items in all cases. Speakers tend to keep a regularity in their boundaries, although not necessarily inserting them at the same place.

5. Discussion

At this point, we should produce further figures to show what percentage of boundaries were correctly recognised, but at issue here is also the question of how to judge such "correctness". Garding and Gerstman (1960) showed with data using different pitch-accent locations that listeners tended to 'correct' their perception in line with their expectations about what can be expected to carry stress. Perceptually, the lengthening distinction around these words is not immediately obvious on listening alone, and it can be difficult to notice these differences. In a hand-labelled transcription it is likely that many of the boundaries would be "correctly" assigned to coincide with the syntax, resulting in a "wrong" score for an automatic algorithm.

Manual labelling of prosodic events is subject to perceptual filtering, and there can be a tendency for domain knowledge to override acoustic facts. When the placement of a stress is somewhat ambiguous for example, lexical knowledge can override, causing it to be marked on a full ("stressable") syllable rather than on a neighbouring schwa, in spite of the speaker's actual performance. Similarly, in a hand transcription, phrase boundaries tend to be placed in accordance with syntactic rules when the actual perceived phrasing "just doesn't make sense". Without specialist training, a transcriber's own dialect can bias a phonetic transcription; even with training, competence knowledge can bias a manual prosodic transcription when the differences are small.

Our corpora are used for training a speech synthesis system, as well as providing source units for concatenative synthesis. Stochastic models are trained to predict timing and pitch contours from repeated exposure to pairs of labels and data, but if the data are not accurately labelled, then the prediction of the models and the synthesis quality will degrade considerably. For our purposes then, the labelling must closely reflect the speech as it was actually produced, and should be based on acoustic rather than on perceptual features if we are to properly model the speaker characteristics of the source data. For this, automatic labelling may be essential, and we may have to trust a bootstrapping approach that employs acoustically-based segmentation, rather than hand-labelling the speech, and then having to design an algorithm that is robust to the differences in the acoustics.

In the majority of cases there is little difficulty in judging acceptibility of the output of the automatic segmentation, and as we saw above, detected boundaries frequently coincide with a syntactic phrase boundary, or at locations where a comma could be inserted in the orthography. Such cases accounted for 432 out of the 501 locations where three or more speakers' data were in agreement. Of more interest are the locations where there is less agreement; although the segmentation based on normalised durations shows a poor correlation between the individual readings of the sentences by different speakers, closer examination suggests that it may be revealing differences that would go unnoticed in a manual transcription. The issue appears to be not so much one of "correctness" as of personal choice of phrasing. Of the 343 locations where boundaries were determined for only two speakers, 144 of these were paired around a grammatical (function) word sandwiched between two lexical (content) words.

With respect to fundamental frequency patterning, it has already been noted (Vaissière, 1992) for these data that function words falling between two pitch groups can cluster with one or the other as a matter of speaker-dependent personal choice; the above results show that there is similar variability in durational phrase-marking as well. We can conclude that although the automatic prosodic segmentation produced different results for different speakers reading the same texts, this is not a weakness of the algorithm, but a feature of the speech that is better revealed by a non-perceptual analysis.

6. Conclusion

This paper shows with multi-speaker data of British English that significant information regarding the prosodic structuring of an utterance can be found from simple transforms of segmental durations obtained by either manual or HMM labelling. Normalisation to reduce phone-specific timing effects yields duration profiles from which prosodic boundary locations can be obtained.

The algorithm presented above appears to be successful in the location of prosodic boundaries that mark the edges of intonational phrases. It shows that there is some individual freedom in the marking of a prosodic boundary, especially in the case of a single function word linking two groups of content words, which sometimes groups with the preceding phrase, in accordance with rhythmic principles, and sometimes with the following, in accordance with syntactic principles.

There is a high degree of inter-speaker agreement in the profiles, and evidence that the events located by these processes correspond to meaningful linguistic events in the speech. Speaker-specific variation shows individual interpretations of the linguistic structures and suggests that one general rule for all may not provide the best model of the speech processes.

Acknowledgment

The author is grateful to the anonymous reviewers of an earlier draft of this paper for helpful suggestions, and to the management and friends at ATR for support and comments.

Appendix A. Test sentences

The 200 SCRIBE sentences were constructed to provide examples of the permissible demisyllables in English, with almost all combinations of single consonants (in both initial and final position) and vowels, as well as examples of consonant clusters up to length four. The sentences were read by four adult speakers of British English. The first twenty-five sentences are reproduced below to ilustrate the type and length of utterance.

001. The price range is smaller than any of us expected.
002. They asked if I wanted to come along on the barge trip.
003. Amongst her friends she was considered beautiful.

004. The smell of the freshly ground coffee never fails to entice me into the shop.
005. I'm often perplexed by rapid advances in state of the art technology.
006. John could lend him the latest draft of his work.
007. From forty love the score was now deuce and the crowd grew tense.
008. The Presbyterian minister managed to curb the drinking habits of the loitering youths.
009. The bulb blew when he switched on the light.
010. It is futile to offer any further resistance.
011. They launched into battle with all the forces they could muster.
012. The chill wind caused them to shiver violently.
013. The government triumphed four years ago and we have every reason to believe that it will triumph again.
014. He jerked round in an instant to face his assailant.
015. He emphasized his strengths while concealing his weaknesses.
016. The table is made so sloppily that it tilts.
017. It was important to be perfect since there were no prompts.
018. I ran for cover whilst he hurled several stones.
019. We have proof that the regime wields sufficient power in the North to exploit the entire population.
020. She flicks through a magazine when she gets a chance.
021. Thank goodness it's Friday and time to go home.
022. Itches are always so tempting to scratch.
023. I'll hedge my bets and take no risks.
024. The length of her skirt caused the passers-by to stare.
025. I always seem to follow my instincts rather than reason.

References

W.N. Campbell (1989), "Syllable-level duration determination", *Proc. European Conf. on Speech Technology, Paris*, pp. 698–701.

W.N. Campbell and S.D. Isard (1991), "Segment durations in a syllable frame", *J. Phonetics, Special issue on Speech Synthesis*, Vol. 19, pp. 37–47.

T.H. Crystal and A.S. House (1986), "Characterisation and modelling of speech-segment durations', *IEEE Internat. Conf. Acoust. Speech Signal Process.*, Vol. 51.11.4, pp. 2791–2794.

J. Edwards and M. Beckman (1988), "Articulatory timing and the prosodic interpretation of syllable duration", *Phonetica*, Vol. 45, pp. 156–174.

J.R. Edwards, M.E. Beckman and J. Fletcher (1991), "The articulatory kinematics of final lengthening", *J. Acoust. Soc. Amer.*, Vol. 89.

K. Edwards, M.S. Schmidt and M.A. Jack (1992), Evaluation of an HMM-based automatic speech segmentation system applied to ATR speech data, *CSTR-ATR Technical Report*, Edinburgh.

J. Gaitenby (1965), The Elastic Word. Technical Report 1–12, Haskins Laboratories New Haven CT, Status Report on Speech Research SR-2.

E. Garding, and L.J. Gerstman (1960), "The effect of changes in the location of an intonation peak on sentence stress"; *Studia Linguistica*, Vol. 14, pp. 37–59.

D.H. Klatt (1975), "Vowel lengthening is syntactically determined in a connected discourse", *J. Phonetics*, Vol. 3, pp. 129–140.

S.E. Levinson (1986), "Continuously-variable duration hidden Markov models for automatic speech recognition", *Comput. Speech Language*, Vol. 1, pp. 29–45.

M.S. Schmidt and G.S. Watson (1991), "The evaluation and optimisation of automatic speech segmentation", *Proc. Eurospeech '91, Geneva*, Vol. 2, pp. 701–704.

D. Scott (1982), "Duration as a cue to the perception of a phrase boundary", *J. Acoust. Soc. Amer.*, Vol. 71, No. 4, pp. 996–1007.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C.W. Wightman, P.J. Price, J. Pierrehumbert and J. Hirschberg (1992), "TOBI: A standard for labelling English prosody", *Proc. ICSLP-92, Banff, Canada*, pp. 867–870.

Vaissière (1992), Personal communication.

C.W. Wightman and M. Ostendorf (1993), "Automatic labelling of prosodic features", submitted.

C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf and P.J. Price (1992), "Segmental durations in the vicinity of prosodic phrase boundaries", *J. Acoust. Soc. Amer.*, Vol. 91, No. 3, pp. 1707–1717.