AUTOMATIC ANNOTATION OF SPEECH CORPORA

Nick Campbell and Yoshinori Sagisaka

ATR Interpreting Telephony Research Laboratories

ABSTRACT - This paper describes a method for automatic annotation of prosodic events in speech corpora and extends previous work that detected prominences from segmental duration and energy measures. It details a way of differentiating prominence-related lengthening from boundary lengthening using durational clues alone, and discusses an anomaly in the phrasing characteristics of four speakers' readings of 200 phonetically-balanced sentences.

INTRODUCTION

There are many ways to produce synthetic speech; the way we have chosen is to select optimal units of variable size from natural speech sources for concatenation. This has been done with some success for Japanese [4], and we are now attempting to apply the same method for the synthesis of English speech. Selection procedures have been developed that allow us to substitute an appropriate source unit from a different articulatory context when the ideal unit cannot be found in a source database of limited size [5].

Prosody is considered as integral in the selection procedure, and we aim to select a unit from an appropriate prosodic as well as an appropriate segmental context. In order to do this, we need a source database of natural speech that is labelled for prosodic as well as segmental features. However, because of the large size of our corpora, and because of the problems involved in hand-labelling, we are now developing methods for the automatic prosodic annotation of speech.

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, phrase boundaries tend to be placed in accordance with syntactic rules when the actual perceived phrasing 'just doesn't make sense'. Competence knowledge can bias a prosodic transcription just as a transcriber's own dialect can bias a phonetic transcription.

Our corpora are used for training of the synthesis system, as well as for source units. Neural networks are trained to predict timing and pitch contours from repeated exposure to pairs of labels and data, but if the data is not accurately labelled, then the output of the networks will degrade considerably. 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.

There is a high degree of interaction between the duration, pitch and energy variations that signal prosodic events, but in this paper we will concentrate on the extent to which measures of duration can be used to determine the prominences and phrasing of an utterance. We will present our approach to one aspect of prosodic segmentation and discuss the problem of verifying results.

RECOGNISING PROMINENCES

Using a corpus of conference-registration dialogues that had been recorded in different versions to show contrastive prominence (focus) we were able to determine which part of each utterance had been assigned focus by using a combination of normalised and smoothed measures of the lengthening and energy of each phone [2].

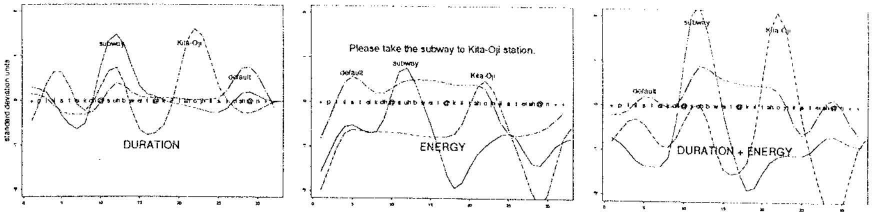


Figure 1: Plots of normalised and smoothed duration (left) show prominence in the utterance but don't distinguish between segments lengthened for stress and those lengthened phrase-finally. These three versions of the same sentences have been produced with focus on different words to convey different interpretations of the same word sequence. We can see that energy (middle) is high throughout the default reading, but drops quite sharply once the focal point has been reached. It starts lower in the marked readings. Energy shows focus well, but not which other words or parts of the utterance have been given prominence. When we combine the two measures (right), (here they have been added), we can see both the stress pattern and the focal prominence.

Because, in English, phones undergo lengthening from at least two different causes (from stressing, or from proximity to a boundary) length alone is an ambiguous clue, but the addition of an energy measure, which is typically weak in the case of pre-pausal lengthening, and strong in the case of stress lengthening, allows disambiguation of these two contexts. Figure 1 illustrates the combination of these two measures.

A test with data of three speaking styles, showed that we were able to correctly identify more than 80% of the stressed syllables, with a false-insertion rate of less than 5%, and to rank them in order of prominence to identify the focussed phrase in more than 75% of the utterances. Analysis of the errors in this test showed that there were cases of sentence-medial phrase-final lengthening that were being incorrectly tagged as stressed because the energy level remained high. Another measure is needed to distinguish stress from final lengthening in such cases.

DIFFERENTIAL LENGTHENING WITHIN THE SYLLABLE

Previous work [1] has shown that differential lengthening takes place on segments within the syllable under the two types of lengthening (see Figure 2). Boundary lengthening affects coda segments more strongly, and stress lengthening is strongest on onset segments. The 'slope' of lengthening through a syllable can therefore be used to distinguish between the two cases. In the rest of this paper we will show how this differential can be applied, to test the feasibility of prosodic segmentation using measures of duration alone.

RECOGNISING BOUNDARIES

By examining the differential lengthening in onset and coda segments in data from readings of 200 phonetically-balanced sentences by four speakers of British English we were able to distinguish sentence-medial final-lengthening without reference to energy values.

Procedure

To apply this differential, a program was written that calculates the slope of lengthening within a syllable, comparing the z-score (type-mean minus token duration expressed in standard deviation units for the type) of each phone with that of its neighbours to determine if the lengthening is increasing throughout the syllable (final case) or decreasing (stressed case). This 'slope' was used to differentiate lengthened syllables to indicate potential prosodic boundaries in the 200 sentences.

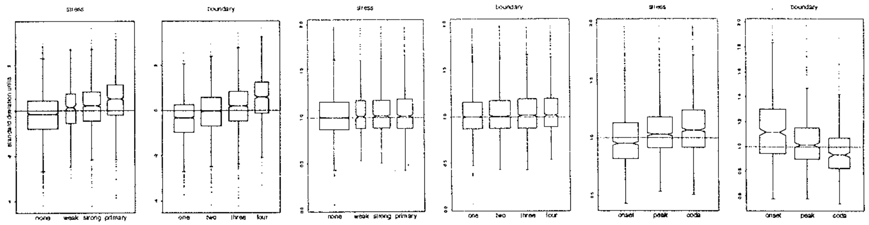
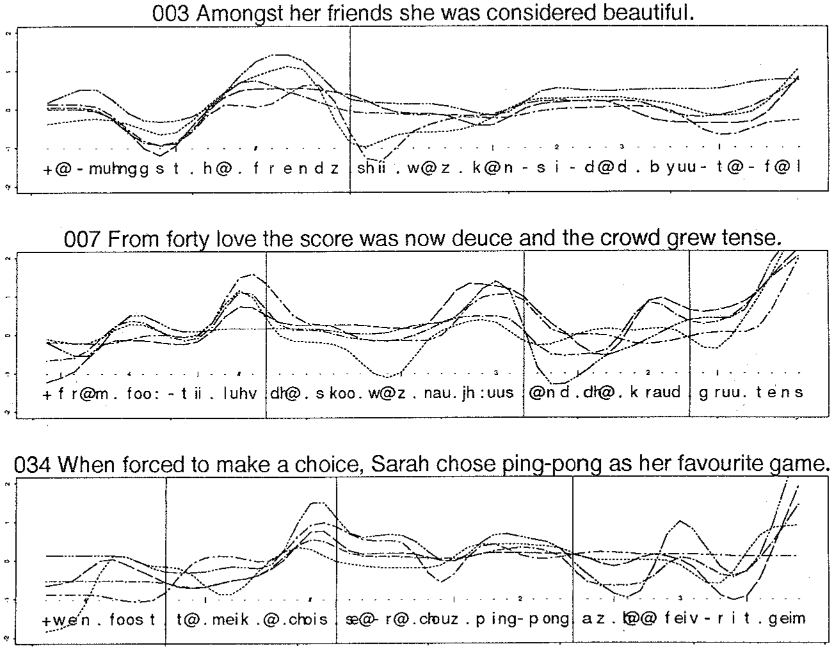


Figure 2: Four levels of stress and final lengthening were marked in the database (left two figures). Prediction of segmental durations on the basis of syllable durations factors out this difference (middle), but error within the syllable (right two figures) shows onset segments to be overpredicted in final syllables, and coda segments to be overpredicted in stressed syllables.


003 Amongst her friends she was considered beautiful.

007 From forty love the score was now deuce and the crowd grew tense.

034 When forced to make a choice, Sarah chose ping-pong as her favourite game.

Figure 3: Plots of normalised segmental durations for three sentences, showing a high degree of similarity in the contours. Vertical lines indicate prosodic phrase boundaries determined by the algorithm. (Raw scores have been smoothed for these plots to allow clearer comparison).

Algorithm

The program essentially has two components; the first calculates the slope within each syllable, the second compares the slope between syllables and indicates a break when the slope is reset:

```
for (each segment in the syllable)
        sum += this.z.score - last.z.score
slope = sum / number.of.segs.in.syll

if ((last.slope > this.slope) and (count > 1))
        then break.phrase
```

The tendency for lengthening to increase through a final syllable will ensure a larger slope value, whereas the strong initial lengthening of segments in a stressed syllable will result in a lower value. Because normalised durations are used, any segment-specific durational effects have been factored out of this measure of lengthening, which is thus more representative of the prosodic effects.

Results

Each syllable in the 200 sentences was assigned a slope value, and when a downward reset was noted in the slope, a prosodic phrase boundary was inserted in the text. 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. However, in only 284 contexts was there unanimous agreement (27% of possible contexts), and there were 542 cases (12.8% of the total possible contexts) where a boundary was marked for only one speaker alone.

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 noted 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 mountain 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
```

Figure 3 plots example results. We can see that words that are lengthened by stress as in 'favourite' (bottom example) are not marked whereas those that mark the end of a prosodic unit are identified.

Discussion

At this point, we should produce figures to show what percentage of boundaries were correctly recognised, but at issue here is the question of how to judge such 'correctness'. In the majority of cases there is little difficulty and, as we can see from the examples above, boundaries frequently coincide with locations where a comma could be inserted in the orthography or at a syntactic phrase boundary. 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. A good example of these is in sentence 171 (examples above) at 'Bach's Toccata 2 and 2 Fugue', where speakers seem to be equally divided about whether to group the conjunctive with Toccata or Fugue, or at 'absorbed 1 in 3 the virtuoso' where three speakers grouped the preposition with the verb, and one with the noun.

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. Vaissiere has noted (personal communication) that function words between pitch groups cluster with one or the other as a matter of speaker-dependent personal choice; it seems that there is similar freedom in durational clustering as well.

Ambiguous phrasing

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 1 shows examples of such ambiphrasal words where speakers were divided in their boundary placement. In a hand-labelled transcription it is likely that many of these will be 'correctly' assigned to coincide with the syntax, resulting in a 'wrong' score for an automatic algorithm.

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

Table 1: Examples of ambiphrasal words; numbers count speakers who lengthened the preceeding word.

Correlations between the boundary locations across speakers were low in this test, averaging about r=0.48, but closer examination of the results indicates that speakers tend to keep a regularity in their boundaries, although not necessarily inserting them at the same place. 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.

Cases where a boundary was inserted for the majority of speakers at a syntactically inappropriate point are shown in Table 2. In almost every case, the 'boundary' preceeds a strong stress. Cutler et. al. [3] showed evidence for stress-based segmentation of English speech on the basis of rhythm; it is possible that here

The price 4 range	The chill 4 wind	round to 3 the side
I can't 4 pretend	These practical 3 jokes	malt 3 loaf
little 3 iced buns	most of the 3 scenes	strawberries have 4 oozed
company 3 directors	discussed in 3 depth	dear 4 old bishop
it's in 4 vogue	explore the 4 relationship	it was the 3 alcohol

Table 2: Boundaries at 'unexpected' positions, showing contexts where three or more speakers exhibited such boundary lengthening.

too we are seeing evidence for a rhythmic relocation of syllables, with final lengthening taking place in lieu of a full pause before the primary stressed syllable.

On the other hand, the phonemic segmentation of speech to produce measures of duration is not an exact process, and caution is due when drawing conclusions from an unproven technique, such as described here, which can be sensitive to small differences in values. We suggest that while these results are interesting, further study is needed before conclusions can be drawn. We have shown, however, that the acoustic prosodic segmentation of speech is practical and worthy of further study, and maintain that although there may be differences between these results and those of a manual segmentation, the differences may be well founded.

CONCLUSION

Large corpora of speech are now being collected in many countries for a variety of speech technology applications. Techniques exist for the automatic and semi-automatic segmentation of speech waveforms to produce a phonemic labelling of the utterance, but similar techniques do not yet exist for the automatic labelling of prosodic events in the speech signal.

This paper shows with multi-speaker data of British English that significant information regarding the prosodic segmentation of an utterance can be achieved from simple transforms of segmental duration using labels obtained by either manual or hmm segmentation. Normalisation (with or without smoothing to reduce phone-specific effects and segmentation uncertainties) yields duration contours from which a prominence index for boundary location, stress detection, and hierarchical ordering of focus can be obtained.

There is a high degree of inter-speaker agreement in the contours, and we believe 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.

We intend to use data segmented in this way to provide source units and training material for a multilingual speech synthesiser, and for the alignment of a pitch contour for further prosodic labelling of our speech corpora. Being objective and theory-independent, the prominence index forms a good base for comparison with perceptual analyses of the speech and, together with boundary marking, enables identification of the more subjective aspects thereof.

REFERENCES

[1] W. N. Campbell (1991) Prosodic segmentation of recorded speech. In PERILUS XIV. Stockholm University, Sweden.

[2] W. N. Campbell (1992) Prosodic encoding of English speech. In Proc ICSLP-92, Banff, Canada.

[3] A. Cutler, J. Mehler, D. Norris, and J. Segui (1992) The monolingual nature of speech segmentation by bilinguals. In Cognitive Psychology 24, 381-410.

[4] Y. Sagisaka (1992) ATR ν-talk speech synthesis system. In Proc ICSLP-92, Banff, Canada.

[5] W. J. Wang, W. N. Campbell, N. Iwahashi, and Y. Sagisaka (1992) Unit selection for English speech synthesis using regression trees. In Proc Acoustical Society of Japan, 1–5–10.