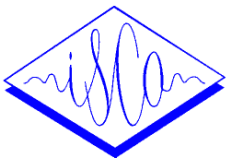

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

PROSODIC ENCODING OF ENGLISH SPEECH

W. N. Campbell

ATR Interpreting Telephony Research Laboratories
Hikaridai, Seika-cho, Kyoto 619-02, Japan

ABSTRACT

This paper describes an algorithm for determining stress groups in a spoken utterance from acoustic parameters of the speech waveform. It uses normalised and smoothed measures of duration and energy to produce an index of stress that is highest on focussed parts of the utterance and lowest at the boundaries between stress groups. 81% of stressed words were correctly identified, with a false detection rate of less than 5%. The location of contrastively focussed words in the utterance was correctly recognised in 76% of cases.

1 Introduction

In order to provide reliable databases for training the next generation of speech synthesisers, large volumes of natural speech must be both prosodically and segmentally labelled and annotated.

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. Since we require several hours of annotated speech for training purposes, some automation of the process is considered desirable.

While it must be acknowledged that current techniques for acoustic analysis of a speech waveform are far from adequate for the level of annotation that we would desire, it is believed that the benefits of automatic annotation outweigh any lack of sensitivity (or abstraction-generation) it may suffer from. For this reason, we are attempting to quantify the extent to which measurable acoustic-prosodic parameters such as fundamental frequency, amplitude and duration can be correlated with boundary, emphasis and focus events in the spoken utterance.

This paper details a method of locating boundaries and stress groups in the spoken uttterance to provide an anchor for automatic determination of pitch movements from the fundamental frequency of the waveform. In this preliminary stage of our project, we are using a database of 100 utterances that were recorded under controlled conditions. Each set of sentences consists of syntactically and semantically identical word-strings that differ in the amount of emphasis given to different sections of the utterance resulting from the use of contrastive stress to clarify (deliberate) misinterpretations.

The first task of prosodic analysis, which we are close to achieving, is to locate and differentiate the main emphases and boundaries in the speech. The second task is to annotate the speech signal with prosodic tags in an attempt to simulate the transcription that a human listener would produce. A major goal of this project is to determine a set of transcription tags that can be generated automatically and yet remain consistent with current methods of analysis, such as those that have been agreed upon at the recent MIT/NYNEX prosody workshops.

The type of speech being used at this stage of the work is far from spontaneous; nor is it completely natural, in the sense that it is free from external/non-speech noises and is controlled in form and content, but it was elicited in a dialogue situation and we believe that it can be considered representative of the quality of speech in many existing corpora. Since a large body of linguistic analysis is already available for these corpora, we believe that a workable mapping may be determined by stochastic methods between the linguistic information and the prosodic labels that we generate.

Recent developments in the automatic determination of break indices [7,9,10] have provided a method of quantifying the degree of affiliation between each pair of words in the utterance, thus reliably indicating prosodic boundaries. Stress indices are proposed here as a complement to these, to measure the prominence of each speech segment between the boundaries. Being data-driven and free of deterministic rules or values, the stress index is speaker-independent, and provides a clustering of phones into stress groups, with a ranking of each according to its relative prominence.

2 Data for a test

There are two criteria by which the stress index can be judged; whether it can locate the stresses accurately, and whether it can differentiate between the prominences. Results are presented below for each. A test database was constructed, and hand-labelled stress marking was compared with the output of the algorithm. Differential placement of contrastive stress allowed a measure of the ranking by providing precise knowledge of the speaker's intentions.

2.1 Materials

A series of recordings was made by a true bilingual speaker of English and Japanese. The speaker is female, college-educated, in her mid-twenties, born in Japan of a Japanese mother and an American father. At home she spoke English, and at school Japanese. She was judged by native-speakers to have American English as her first language. Her Japanese is said to be slightly accented but totally fluent. The speaker knew that the recordings were to provide source material for a bilingual speech synthesiser but was not aware that they would be used for an analysis of prosodic features. She has had no formal training in linguistics or speech-science and can be considered naive to the purpose of the experiment.

As part of a wider project to study the mappings from read speech to spontaneous speech, recordings were made of her using both languages in a variety of styles and registers. The subset with which we are concerned here consists of a series of twelve telephone dialogues between a receptionist and a caller, identical to those used in the ATR-CMU Conference-Registration Database Project.

The text of the dialogues was presented to the speaker in advance to allow her to familiarise herself with the content and look up any unfamiliar words (e.g. 'Telephony'). In the recording studio the following day she was asked to read each dialogue from start to finish. No emphasis was placed on the fact that the sentences formed a conversation or could be attributed to different

10.21437/ICSLP.1992-218

speakers. After a break, a second set of readings was produced. This time she was instructed to differentiate between the speakers, and first read only the sentences of speaker A, then, in a different session, only the sentences of speaker B, thus taking a role in the conversation, but with no feedback, and alone in the studio. A third set of readings was produced the following day with another person in the studio, without eye-contact, using two microphones. This time the dialogues were performed as 'realistic' conversations, such as might be used for a radio play. She was by now familiar with the supposed context of the utterances and able to produce them quite spontaneously.

2.1.1 Initial test database

To provide a baseline set of training material for the test of the stress index algorithm, three further sets of utterances were recorded using a subset of the sentences capable of being read with different interpretations. For example, sentence 4 of dialogue 9, '*Please take the subway to Kita-Oji station.*' was originally produced in response to a request for information about how to reach the conference site. Emphasis was placed on 'subway' and 'station'. It could, however, have been produced with quite different emphasis in response to a question asking whether it is not better to take a taxi to the station or, alternatively, how far the visitor should stay on the subway train. In an extreme case, the first word could be emphasised to persuade someone reluctant to use the subway at all.

There are thus several possible interpretations of the same phone sequence, depending on the background of the conversation and the intentions of the speaker. Since the sentences are syntactically identical, the differences can only be encoded in the prosody.

In all, 27 sentences were selected, with an average of three contrastive interpretations for each, forming a set of 100 utterances altogether, of which 73 were marked for contrastive stress. The sentences were ordered in two lists, with the words to be emphasised printed in block capitals. In the first list, Set A, the default unmarked reading was followed by each of the marked versions, with emphasis shifting from early to late targets throughout the sentence in increasing order. In the second list, Set B, the ordering of the sentences was randomised both for ordering within the sets, and for ordering between them, under the constraint that there was no relation between any two adjacent sentences in the list.

The differences were explained to the speaker, making sure that she understood the type of cue sentences that could elicit each of the readings, and she was asked to read the two lists 'naturally, but in a way that makes clear which interpretation is intended'. After a break, I entered the studio and in a series of mini-dialogues, elicited each of the response sentences (Set C) by apparently misunderstanding her intended meanings.

2.2 Labelling the data

Transcription files were generated at Edinburgh University (CSTR) from the orthography of the texts and used to guide HMM segmentation for location of the phone boundaries.

Precise determination of phone boundaries is probably not possible in speech, but a test of the agreement between hand and HMM-segmented data [3] indicates that more than 80% of the boundaries were within 25 msec, and about 60% were within 15 msec. A comparison of the boundaries determined by two human labellers [6] showed a similar degree of consistency. HMM segmentation was therefore used to provide approximate phone durations for the above readings. Hand-labelling of the stress and intonation was performed Two sentences in Set B and three in Set C were not labelled.

2.3 Normalising and smoothing

Since we are laying the foundation for a pitch labeller, use was made only of duration and energy information. Raw phone durations have been used in stress detection [4,5], but the information they provide about the underlying prosody is masked by phone-specific durational characteristics. When they are normalised to remove these phone-specific effects, and smoothed to reduce boundary placement differences, they can be more informative.

To normalise the durations, means and standard deviations (SD) were calculated, for each phone type, over all tokens in the data. The observed duration of each token, minus the mean for its type, was then expresssed in SD units. In this way, the lengthening or shortening of each token relative to the mean for its type, was converted to a standard unitless value, and any 'inherent' features of duration were filtered out. The resulting values were normally distributed about a zero mean, i.e., were almost all within the range of plus-minus three.

Smoothing of these values was done using Tukey's 4(3RSR)2H twiced [8]. Twicing is the process of smoothing, computing the residuals from the smooth, smoothing these and adding the two smoothed series together. Smoothing is done with a running median, window width 3, end-values copied on. A Hanning window (i.e., a moving average with weights 1/4, 1/2, 1/4) is applied after smoothing to reduce large gaps or jumps in the smooth.

This processing gives us a value for each phone that is derived from an estimate of its duration[1]. Higher values indicate a section of the speech that has been said more slowly than is normal for this speaker with these utterances, and lower values predominate where she has been speaking more quickly. The resulting values, the stress contour, vary in ways that are quite consistent with our intuitions about the structuring and prominences of the sentences; boundaries between phrases appear as troughs of low values, and information units appear as 'humps' or peaks of high values. Content words and and other items that have been given prominence in the speech tend to be marked as humps; the more the prominence, the higher the peak. However, final lengthening [1,10] can cloud judgements of stress if they are made from length values alone.

2.4 Combining duration and energy

Energy information can also assist in the location of emphasised areas of the speech [5]. Since there is a strong correlation between lengthening for emphasis and increased energy in the emphasised region, and since there is also a strong correlation between boundary-related lengthening and a *decrease* of energy as the utterance trails off into the boundary, then the energy contour can be used to distinguish the two forms of lengthening. Because declination was observed in the energy contour throughout the shorter utterances, the residuals of a linear regression of the smoothed data were used.

The values of a similarly normalised and smoothed energy contour are also unitless numbers representing degree of prominence. Since they are similarly scaled, the simplest combination of these two measures is by direct addition. Adding energy values to those of the durations serves to increase the strength of the values that are related to emphasis, while decreasing those associated with proximity to a boundary.

Local maxima and minima were then determined within the summed smoothed values and these were taken as the centres of the stress peaks and boundaries respectively. The values of each maximum were taken as the stress index for the group it covered. They were ranked to provide a number (1 highest) to indicate the relative emphasis of each. The largest number (lowest rank) is

[1]Silences also have duration, and these values are therefore also normalised, but since many non-linguistic factors may also be involved in the duration of a silence, and since silence-detection is becoming quite robust, length values for silence are reset to zero before smoothing.

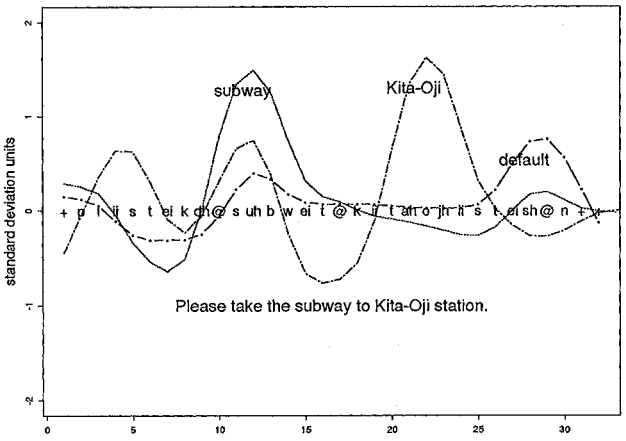

Figure 1: Values of lengthening for three readings.

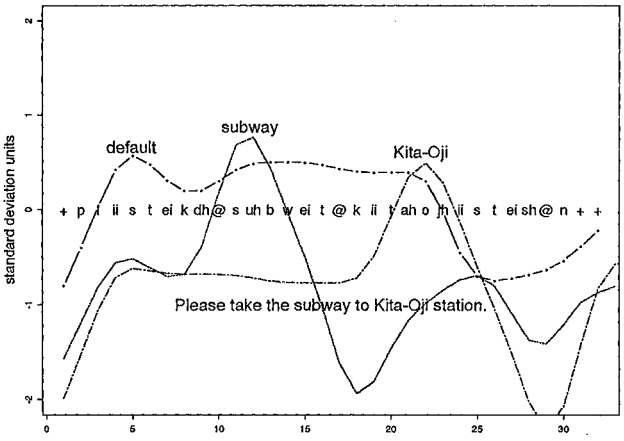

Figure 2: Values of energy for the three readings.

the same as the number of stress groups in the utterance (unless, of course, the lowest two happen to share exactly the same value).

3 An example sentence

In this section we will look in detail at the contours for one sentence from Set A. Figure 1 shows three readings of the example sentence mentioned above. The MRPA (machine-readable phonemic alphabet) labels for the phones are displayed along the middle of the plot at the zero line. The abscissa shows a count of the phones in the utterance; silences at the start and end of the utterance are also counted and included. The ordinate shows SD units for the individual phone duration distributions. The values for each reading are labelled 'default', 'subway', and 'Kita-oji' respectively.

Four 'humps' or prominences are clearly discernible in this plot; on the words 'please', 'subway', 'Kita-Oji', and 'station'. These are the key content words in the sentence. The first (default) reading has only three, on 'please', on 'subway', and on 'station'. The second reading shows a large prominence on 'subway' and then falls below the level of the default for the rest of the sentence, showing only a slight rise over 'station'. The third reading shows three clear prominences, rising on each key word to reach a maximum on the 'Kita-Oji', which is focussed in this utterance. The line drops sharply from that point, falling to the lowest value on 'station'.

Figure 2 shows the energy values for the same utterances. We can see that the energy in the default reading is high for all but the last word. In contrast, that of the 'subway' reading only goes above zero at one place, though it peaks three times. Values for the 'Kita-Oji' reading stay level until the marked word, and then fall steeply.

Figure 3 plots the values for duration added to those for energy. In this case, the differentiation is greatly enhanced and each contour displays four peaks, one for each of the content words. We can now see prominences on 'please' and 'subway' reappearing in the third reading, showing that they were indeed emphasised (or stressed) as content words in the sentence, but not nearly as much so as the focussed word.

By adding the two values we have combined different sources of information, incorporating any trade-off between the two, and as can be seen in the values for the default reading, have prioritised the key words in an intuitive order, so that 'subway' and 'Kita-Oji' now appear more prominent than the less informative 'station' and 'please'. The addition of the energy information has cancelled out much of the pre-pausal lengthening and increased

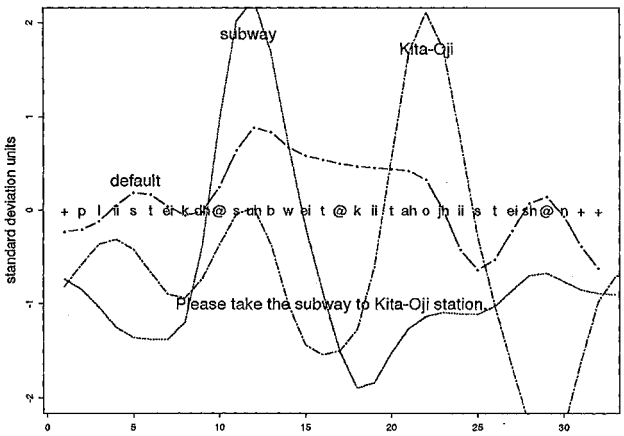

Figure 3: Summed values for energy and duration (see text).

that due to emphasis.

4 Locating emphasis in the speech

In order to test the generality of this stress index and show that it is not just confined to the untypical style of reading immediately contrastive pairs, similar plots were made for each sentence set, showing the variation in the combined values of smoothed and normalised duration and energy. Counts were made of the number of times a peak coincided with a target word in the utterance. Further counts were made of the number of times a target word was ranked highest in the stress hierarchy.

4.1 Recognising stress

Using contours generated for the unmarked (default) readings, a count was made of the number of times a stressed syllable fell within a window of three phones either side of the maximum in the peak. Table 1 shows that 81% of the syllables noted on careful listening as stressed (i.e. having a 'beat', rather than just having a full vowel [0]) were recognised correctly. The main cause of errors was when the algorithm grouped together syllables in compound words that had been transcribed with separate stresses. These typically included pitch accents that continued over more than just one word. 19% of these stresses were missed. False insertions

Table 1: Counts of stress recognitions.

	correct	missed	total	extra	correct/total
Set A:	109	9	118	5	92%
Set B:	108	31	139	4	78%
Set C:	78	30	108	8	72%
total:	295	70	365	17	81%
percent:	81%	19%	100%	<5%	

Table 2: Counts of focus recognitions.

	duration	energy	both	max	both/max
Set A:	34	43	55	71	77%
Set B:	42	42	55	70	79%
Set C:	40	33	52	73	71%
total:	116	118	162	214	76%
percent:	54%	55%	76%	100%	

accounted for less than 5% of the total number of stresses.

We can see that the more spontaneous, interactive speech of Set C has fewer stresses marked, and that the algorithm finds fewer of them. Listening to the recordings, we have an impression of greater use of pitch range in this style, and accordingly predict that the algorithm, as is, will be less successful transcribing spontaneous speech. It may be that prediction of a default contour is necessary, and that by differencing observed from predicted contours we will be more successful in locating stresses. Preliminary tests using different readings of the sentences show promising results, but there is not room here to present a detailed analysis.

4.2 Recognising focus

Using the marked (contrastive focus) readings, under the assumption that the target word would be the most prominent in each utterance, a comparison was made of the performance of the two parameters, separately and combined. Table 2 shows that both duration and energy alone can be used to locate the most prominent peak in just over half the utterances, and that when combined, they provided an accurate indication in 76% of the cases.

An analysis of the 52 prominences not detected by the algorithm showed that 6 cases (3%) were due to a higher initial prominence (energy peak), 9 (4%) to stronger stress on another related word elsewhere in the utterance, 9 to confusion from pre-pausal lengthening utterance-internally, 14 (7%) to similar lengthening utterance-finally, and 14 to signalling of the focus with pitch-prominence rather than stress. This indicates that energy values alone may not be sufficient to differentiate final lengthening, and suggests that incorporating break-indices or syllable-based measures [1,9] may be desirable.

It was also noticed that in many cases, the emphasised syllable shared its prominence with the immediately preceeding syllable. These cases included not only cliticised determiners and particles, but also the seemingly unrelated endings of previous words, suggesting that the emphasis on a syllable may be generated in advance, disregarding syntactic or morphological boundaries. Human labellers may ignore such durational clues when determining where to place marks by hand because of linguistic knowledge; it is open to question whether automatic labellers should be constrained by similar knowledge.

5 Summary and Discussion

It is already well established that there is useful information about focus events in the pitch contour of a speech signal (see for example [2]). This paper shows that duration when combined with energy clues, after normalisation and smoothing, can be reliably used to extract similar information. The output of an automatic segmentation system provided data that was sufficiently robust to enable efficient stress grouping, with successful location of three out of four emphasised words in each set of sentences.

The normalisation and smoothing procedures reduce dependence on the measures for any one particular speech segment, and make the method more robust to small differences in transcription or phonemic alignment when two performances of the same utterance are being compared.

Because the resulting contour can be successfully used to locate emphasis on marked words, we intend to utilise the stress index for grouping and hierarchical ranking when transcribing the rest of our corpus. It remains to be shown whether there is justification for noting more than the currently transcribed three levels of stress, but now that we have the data, a more comprehensive analysis of the linguistic correlates will be performed.

Such a system can be used not only as part of a prosodic segmentation algorithm, as described in the introduction, but also has applications in both speech synthesis and recognition. As the technology progresses from recognising just the segments of an utterance towards 'understanding', or recognising the intention of an utterance, so the use it must make of these prosodic clues will increase. If a segment hypothesis is generated, then use can be made of the both the observed duration 'contour' and the difference between this and a predicted one for the same string.

At ATR we are also exploring the feasibility of learning synthesisers, that adapt to new model input. It is possible that a speech synthesiser could be trained to amend its default pronunciation of an utterance from spoken examples, rather than by heuristic programming. The above algorithm would enable a synthesiser to identify the direction of required change by comparing its generated default with the input from its user. This will allow even non-specialist users a higher degree of control over their devices, and facilitate personalisation of the working environment.

Acknowledgement

The author is grateful to CSTR for their assistance in the labelling of the data, to the management and friends at ATR for support and comments on an earlier draft of this paper, and to prosody@purccvm for information on who is doing what where.

References

[0] Beckman, M. (1986) Stress and non-stress accent. NPA 7, Foris.
[1] Campbell, W. N. (1991) Prosodic Segmentation of Recorded Speech, PERILUS XIV: Current Phonetic Research Paradigms, (University of Stockholm) pp 101-106.
[2] Chen, F. R. & Withgott, M. (1992) The use of emphasis to automatically summarise spoken discourse. Proc ICASSP-92, I 229-232.
[3] Edwards, K., Schmidt, M. S., and Jack, M. A. (1992) Evaluation of an HMM-based automatic speech segmentation system applied to ATR speech data. Report prepared for ATR.
[4] Hieronymus, J.L. (1989) Automatic sentential vowel stress labelling Eurospeech 89, 226-229
[5] Hieronymus, J.L. & Williams, B. W. (1991) An investigation of the relation between perceived pitch accent and automatically-located accent in British English. Eurospeech 91, 1157-1160
[6] Hieronymus, J.L. (1992) personal communication.
[7] Price, P., Ostendorf, M., Shattock-Hufnagel, S. & Fong, C. (1991) The use of prosody in syntactic disambiguation. JASA 80 2956 - 2870.
[8] Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts.
[9] Wightman, C. W. & Ostendorf, M. (1992) Automatic Recognition of Intonational Features. Proc ICASSP-92, I-221-224.
[10] Wightman, C. W., Shattock-Hufnagel, S., Ostendorf, M., & Price, P. (1992) Segmental durations in the vicinity of prosodic phrase boundaries. JASA 91 1707-1717.