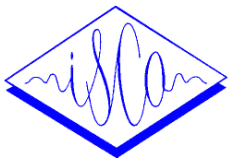


3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ASSIGNING INTONATION ELEMENTS AND PROSODIC PHRASING FOR ENGLISH SPEECH SYNTHESIS FROM HIGH LEVEL LINGUISTIC INPUT

Alan W Black[1] *Paul Taylor*[2]

[1] ATR Interpreting Telecommunications Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN. awb@itl.atr.co.jp
[2] Human Communication Research Centre, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK. pault@cogsci.ed.ac.uk

ABSTRACT

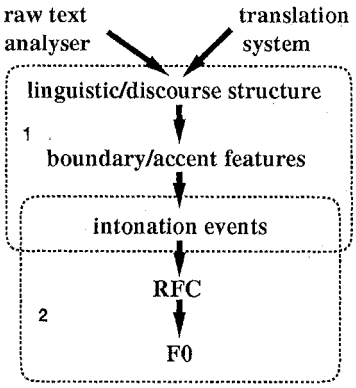
This paper describes a method for generating intonation events and prosodic phrasing from a high level linguistic description. Specifically, the input consists of information normally available from linguistic processing: part of speech, constituent structure and, importantly, speech act. The generated output contains explicit intonation events from which an F_0 contour may be generated. Prosody can be controlled via features in the input describing the *function* of words and phrases without direct reference to intonation. The results are evaluated against natural spoken sentences.

1. INTRODUCTION

This paper describes a method for generating intonation events and prosodic phrasing, for English speech synthesis, from high level linguistic factors. When a speech synthesizer is used as an integral part of a speech translation system, the input to the synthesis sub-system need not be simple unlabelled text. Structured information including speech act type, part of speech, and syntactic constituent structure already exists in earlier components of the translation system. This information can be used directly, enabling the synthesis system to produce better intermediate representations, and consequently better quality speech.

This method consists of four stages. The input takes the form of a syntactic tree where each node is labelled with a feature structure. In the first stage—accent assignment (based on [3])— each word is assigned a feature whose value is one of: accented, deaccented, emphatic or cliticized. The second stage assigns prosodic phrase breaks based on syntactic constituency, grammatical function, and constituent length [1]. Only one level of prosodic phrase break is predicted, and another two, "sentence" and "discourse" are identified explicitly in the input. The third stage takes the labelled (syntactic constituent structure) feature tree and rebuilds it into a prosodic phrase tree based on the predicted phrase breaks. In general, this builds a much flatter tree than the syntactic structure. However, importantly, the prosodic phrase tree need not have the same constituent boundaries as the syntactic tree. In the fourth and final stage, pitch accent and phrase features are realised as intonation events. The events are assigned based on the speech act type of the utterance (i.e. statement, question etc.). The final output consists of a prosodic tree whose leaves are words explicitly labelled with intonation events, which are then used to generate an F_0 contour.

This process needs to be positioned in the wider context of prosody generation in speech synthesis. The following diagram identifies the processes involved



This paper is primarily concerned with box 1. (Aspects of box 2, in our system, are described in [6].) As we progress down this model the description of prosodic events becomes more and more explicit.

Although we are working with a particular intonation model the techniques described here are intended to be general enough to fit with alternative models. That is, this paper describes a method for going from labelled text to a structure explicitly labelled with intonational events and prosodic phrase boundaries. In our case the labelled text is produced by a translation system but it could come from a raw text parsing system.

The rest of this paper details the method. Some discussion of its flexibility in controlling the generation of "non-default" intonation is also given. Next, results of how well it performs with respect to some natural spoken data are presented. Finally the "mismatches" it makes are described and some general discussion of this method is given.

2. METHOD

The input to this technique is some form of "syntactic" tree. As yet it is unclear what information is necessary but in the current implementation, input trees have the following information explicit:

- Part of speech: from crude set of around 8 tags.
- Major constituent structure: Noun phrases, verb phrases, prepositional phrases etc.
- Speech Act: question, yes/no question, statement etc.

Sentence boundaries are already explicitly indicated and each sentence has a speech act. The input is a feature structure tree. A simple set of defaults is applied before further processing. A typical input structure for the utterance "*Hello. Is this the conference office?*" is of the form

```
(((CAT D))
 (((CAT S) (IFT Greeting))
  (((LEX hello))))
 (((CAT S) (IFT YNQuestion))
   (((CAT Copula) (LEX is)))
   (((CAT NP))
    (((CAT Noun) (LEX this) )))
   (((CAT NP))
    (((CAT Det) (LEX the)))
    (((CAT Noun) (LEX conference)))
    (((CAT Noun) (LEX office))))))
```

The first stage in this method is to mark prosodic phrase boundaries. This algorithm is based on Bachenko et al. [1]. Sentence internal boundaries are identified based on constituent type, grammatical function, and constituent length. The result is the addition of `PhraseLevel` features to nodes in the input feature structure. In the above example, as it is short, no boundaries are predicted.

The second stage assigns accent information to each word. The assignment algorithm is based on the work of Hirschberg [3]. Each word is assigned a feature with one of four values: accented, deaccented, emphatic or cliticized. It should be made explicit that at this point these features are *not* intonation events themselves, these are at a higher level of abstraction. The added features are indications of the prosodic function of a word, later processing will assign actual intonation events based on these features (and other information).

The accent assignment algorithm is rule driven and assigns its values based on word class (function words vs. content word—which is defined in terms of part of speech) and context. Although we have implemented almost directly the Hirschberg algorithm described in [3, p 373] there a few differences. First we do not include, as yet, any "local/global focus". Second as the complex noun analyser (described by Sproat [4]) used by Hirschberg is not as explicitly described we have designed our own. Although much simpler and less complete, it has proved adequate for the present, though will need to be extended later. Sproat's system depends on a significant dictionary of compound nouns which we do not have, however our input will normally have much of the internal structure of compounds already defined.

After accent assignment our example becomes

```
(((PhraseLevel :D) (CAT D))
 (((PitchRange two) (PhraseLevel :S)
   (CAT S) (IFT Greeting))
  (((HAccent +) (LEX hello))))
 (((PitchRange two) (PhraseLevel :S)
   (CAT S) (IFT YNQuestion))
  (((HAccent -) (CAT Copula) (LEX is)))
  (((CAT NP))
   (((HAccent +) (CAT Noun) (LEX this))))
  (((CAT NP))
   (((HAccent -) (CAT Det) (LEX the)))
   (((HAccent +) (CN Stress)
     (CAT Noun) (LEX conference)))
   (((HAccent -) (CN Unstress)
    (CAT Noun) (LEX office))))))
```

That is all phrases are marked with the features `PhraseLevel` and `PitchRange` (`two` is the default). All words are marked with the features `HAccent`, `+` means accented, `-` means deaccented. The words *hello*, *this* and *conference* are marked accented while all others are deaccented. The final compound noun has further markings identifying its internal structure, as generated by the complex nominal analyser.

This finishes the first part of the method. So far we have merely labelled the input with more explicit features regarding the function of its contents. The second part consists of two further stages which use that information to assign actual intonation events and prosodic boundaries.

The third stage takes the syntactic tree which is labelled with both phrase break and Hirschberg accent features and restructures it into a prosodic phrase tree. Each node labelled with a phrase break feature becomes a non-terminal node in the prosodic phrase tree. It dominates all words in the syntactic tree up until the next node labelled with a phrase break (traversing the syntactic tree in pre-order). The result is typically a much flatter tree with one level of phrase below a sentence. The result of this procedure on our example is as follows, (the tree structure is also changed slightly).

```
(:D ((PhraseLevel :D) (CAT D))
 (:S ((PitchRange two) (PhraseLevel :S)
      (CAT S) (IFT Greeting))
  ((hello (HAccent +) (LEX hello))))
 (:S ((PitchRange two) (PhraseLevel :S)
      (CAT S) (IFT YNQuestion))
  ((is (HAccent -) (CAT Copula) (LEX is)))
  ((this (HAccent +) (CAT Noun) (LEX this)))
  ((the (HAccent -) (CAT Det) (LEX the)))
  ((conference (HAccent +) (CN Stress)
               (CAT Noun) (LEX conference)))
  ((office (HAccent -) (CN Unstress)
           (CAT Noun) (LEX office)))))
```

The final stage involves the realisation of intonation events on words. These are both pitch accents and boundaries. Up to this point our representation is independent of any particular intonation model. The events realised are a function of a sentence's speech act type and the feature markings for `HAccent`. The speech act type is identified by the feature `IFT` (Illocutionary Force Type). A typical rule is

```
YNQuestion: START
            HAccent +     E_fall
            HAccent ++    E
            TAIL          E_late,rise
```

That is on a sentence labelled with IFT `YNQuestion`: add no features at the start, add E_{fall} (fall event) to all words labelled with `HAccent +`, add E to words labelled `HAccent ++` and on the final word add a $E_{late,rise}$ (boundary rise) to give the question the characteristic final rise.

Apart from START and TAIL, which refer to the first and final words in the sentence, all other parts refer to features in the input tree. Basically any feature (and value) may be associated with any intonation event.

The result after the application of these rules to our example is

```
(:D ((PhraseLevel :D) (CAT D))
 (:S ((PitchRange two) (PhraseLevel :S)
      (CAT S) (IFT Greeting))
  ((hello (HAccent +) (LEX hello))
      (E fall)))
 (:S ((PitchRange two) (PhraseLevel :S)
      (CAT S) (IFT YNQuestion))
  ((is (HAccent -) (CAT Copula) (LEX is)))
  ((this (HAccent +) (CAT Noun) (LEX this))
      (E fall))
  ((the (HAccent -) (CAT Det) (LEX the)))
  ((conference (HAccent +) (CN Stress)
               (CAT Noun) (LEX conference))
      (E fall))
  ((office (HAccent -) (CN Unstress)
           (CAT Noun) (LEX office))
      (E late rise))))
```

Now we have a structure with explicitly marked prosodic phrase boundaries, pitch accents and boundary tune. These are described using intonation events suitable for input to

lower level parts of the system. Specifically, we have mapped from a "linguistic structure" with no explicit prosodic information to one where all necessary information is explicit.

Within our system we actually generate intonation events of the form described in [6], as it has been shown that this encoding proves a rich representation of intonation of natural spoken speech, if this method predicts the right events we will be able to construct natural sounding intonation. Briefly, in that lower level process, events are mapped using declared event attributes, which are parameterized per speaker, into RFC elements. RFC elements offer a symbolic representation of the F_0 which can straightforwardly be mapped to the actual contour [5].

3. CONTROLLING PROSODY

It should be made clear that in the simplest case the above method assigns so-called "discourse neutral" intonation patterns to its input. That is in the absence of information to the contrary a "reasonable" intonation is predicted. But a major advantage of this method is that it offers systematic methods for controlling a much wider range of prosody.

In this method, there are effectively two dimensions in the control of generating prosody. Given the same words, different intonation patterns may be assigned if different initial features are specified. One dimension is controlled by the IFT feature (specifying speech act information). For example the same set of words may be a statement in one instance and a yes/no question in another. By simply changing the value of the IFT feature we can get a different intonation tune for these utterances. Also the type of pitch accents on words may be a function of the speech act. Although the accent assignment algorithm may predict the same values for each word, how they are realised is dependent on the IFT based rule. Question-type sentences may use different types of pitch accent from declarative statements.

The second dimension in controlling the prediction of prosodic events involves controlling words within a sentence. For example although prepositions are normally not accented they may be accented in contrastive situations. For example if a speaker wishes to emphasize the fact that a book was *on* a box (rather than *in* the box), we can represent this with the input,

```
(((CAT S) (IFT Statement))
 (((CAT NP))
  (((CAT Det) (LEX the)))
  (((CAT Noun) (LEX book))))
 (((CAT VP))
  (((CAT Copula) (LEX is)))
  (((CAT PP))
   (((CAT Prep) (LEX on) (CONTRASTIVE +)))
    (((CAT NP))
     (((CAT Det) (LEX the)))
     (((CAT Noun) (LEX box)))))))
```

Another feature used in our current description is that of "focus". By "focus" we mean a prominent word in a sentence. Words marked in the input with the feature `Focus` are given larger accents. Moving the `Focus` feature to different words allows control of where the prominences in the utterance are perceived.

Although features can be specified in the input which are directly realised as particular intonation events it is important to reiterate that the intention of this descriptive method is that the input be marked with abstract *functional* features rather than specific intonational features. Thus systems generating the input to this method need not have knowledge about prosody itself.

Although this method is flexible and powerful enough for our current test set, the rules are still rather *ad hoc*. The IFT to intonation event rules are generated by hand. And although, as we will see, the results are reasonable, there may be more principled ways to derive these rules.

4. DOES IT WORK?

In order to evaluate the results of this method for predicting intonation events from labelled input we tested it against a small number of naturally spoken sentences from the CMU Conference Registration Database. Independently, around 60 sentences from one male speaker were labelled (semi-automatically) with intonation events (see [6]). The sentences were hand parsed. The parsed sentences were used as input to the above algorithm and the results were compared against the actual intonation events of the labelled natural forms. These tests were applied to the "default" predictions, (no special focus features were added to the input).

First let us consider phrasal boundaries. In the 57 test sentences there are only 6 labelled phrasal breaks (signified by silence). The described algorithm predicts 11 breaks, 5 of which match the actual breaks. If we state that phrasal breaks in the natural speech are silence and/or significant drops in F_0 then there are 10 actual breaks, 9 of which are predicted correctly plus two others. An example of one difference is in the following sentence, the original only has a phrase break after "*Yes*" while an extra one is predicted before "*is*".

> *Yes # two hundred dollars per person # is required as a registration fee*

Although different from the boundaries chosen by the original speaker this boundary is not unreasonable, as are the other differences. Of course only a small number of phrases actually exist in the data so it is difficult to gain any real notion of its effectiveness. However it would appear that the algorithm slightly over predicts (with respect to this particular speaker and these utterances) but in acceptable places.

A second point is about the type of boundary accent assigned. The boundary tunes are dependent on the IFT marking and there is a direct correspondence between the IFTs and the type of boundary tune (even though the IFTs are selected for pragmatic reasons not prosodic reasons). Thus all boundary tunes predicted match the boundary tunes in the original (given the broad classification of rising, falling or flat). A total of five IFTs are sufficient for these examples (`Interjection`, `Greeting`, `Statement`, `Question` and `Yes/No Question`). However some problems (discussed below) may be avoided if we extended the number of IFTs.

The more interesting result is with respect to pitch accent assignment. A simple measurement is used. If the natural labelled speech has an intonation event labelled, and an intonation accent is predicted on a word it is considered a match, or if no intonation event exists and no intonation accent is predicted it also considered a match, otherwise it is a mismatch. That is currently no check is done of the *type* of accent only if one is predicted or not.

The following table shows the results for four conversations (all from one male speaker).

Conversation (Sentences)	Words	Mismatch	% Difference
C01 (13)	89	13	14%
C02 (15)	96	26	27%
C03 (13)	101	23	23%
C04 (16)	133	34	24%
Total	419	97	23%

The better results for the first conversation are probably due to tuning the rules too much for that conversation. But it seems that with not too finely tuned rules the method is predicting around 75% of the intonation events in the natural speech.

5. MISMATCHES

It is important to note that mismatches are not necessarily wrong. Where a speaker chooses to place an accent sometimes has no clear bearing on the semantics or pragmatics (notwithstanding the question of whether an intonation event actually represents a significant pragmatic event). However on looking at the differences between the natural speech and the synthesized speech we can clearly identify four classes of mismatches.

The first class are over prediction of accents in non-important places. For example, while the speaker says (accented words are in capitals).

THIS is the office for the CONFERENCE.

the described method predicts

THIS is the OFFICE for the CONFERENCE.

Which also seems natural, even if not the same as the naturally spoken one.

A second area is where accents extend over more than one word. For example, in a phrase like *"I would like"* the speaker may often rise on the *"I"* and fall on the *"like"*. The method would predict two separate accents, one on *"would"* and another on *"like"*. Although such long accents could be treated within this method (as a rise and a fall) a better description of this phenomena would be more appropriate.

The third area where there are significant mismatches is in certain fixed phrases. In zip codes there is typically only a 50% match. The method produces acceptable accents but not the same as the original speaker. Another phrase, *"thank you very much"*, seems to be used with differing intonation which this method does not capture. It does however seem reasonable to think these types of phrases have particular intonational properties and can justify specific treatment.

The fourth area, which is the most serious, is where the mismatch of predicted accents make a significant semantic or pragmatic difference to the sense of the utterance. One such example is in the following sentence, the method predicts the following

PAYMENT SHOULD BE MADE by BANK transfer.

That is individual accents are added to SHOULD, BE and MADE while the original has only one accent over the three words. The extra accents give the listener the impression the speaker is a little upset and perhaps the listener has not yet payed the fee.

There is also a fifth class, where no clear reason could be identified.

There is the possibility that, the speaker said the sentences in a non-default way, and hence the default prediction (as no *extra* controlling features are added to the input) is not expected to match anyway. But this does not seem to be the case. Most of the mismatches are not considered to be alternative readings of the sentence which would require specific marking in order to obtain—the sentences, although part of a dialogue are read in a fairly standard way.

6. DISCUSSION

This first point to make is that the output is typically acceptable. It is rare that the wrong semantic message is conveyed by wrongly predicated intonation. However, non-misleading intonation is only a beginning, in order for synthesized speech to sound more natural a wider range of intonation (and finer control) is required.

Although our test set does consist of natural speech it is acted dialogue rather than spontaneous dialogue. The technique described here is designed to cater for more complex forms of intonation as found in spontanious speech, thus perhaps the current test set does not show off the capabilities of this technique to their full.

Although the evaluation of prediction of prosodic events is always difficult, this does not mean that it should not be attempted. However, what is important is to realise that the resulting figures are not exact. Fine tuning which only marginally improves the results is probably not significant.

The above method currently predicts: sentence-internal prosodic breaks, intonation events at sentence boundaries and accents on words. Further control is available through specific marking, such as contrastive stress and focus. It does not (at present) offer a facility for phrasal accents over a number of words or allow specification of boundary tunes at sentence internal boundaries. Also accents typically are resolved to one of only a few types which does not reflect rich range of accents actually used. Improvements in these areas are being investigated, but these extensions still fit neatly within the current framework.

This work is implemented within the CHATR generic speech synthesis system [2]. This environment allows the rules to be developed interactively. Synthesized utterances generated from both this method and the originals can easily be compared, listened to, and displayed.

7. SUMMARY

This paper presents a method for assigning prosodic boundaries and intonational events to labelled text in a flexible but systematic way. Importantly the input is labelled with "standard" linguistic information identifying functional aspects of the utterance. The output however has explicitly labelled intonation events. Different intonation patterns may be generated based on different high level input features.

REFERENCES

[1] J. Bachenko and E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170 1990.

[2] A. W. Black and P. Taylor. CHATR: a generic speech synthesis system. In *Proceedings of COLING-94*, Kyoto, Japan, 1994.

[3] J. Hirschberg. Using discourse content to guide pitch accent decisions in synthetic speech. In G. Bailly and C. Benoit, ed, *Talking Machines*, pp 367–376. North-Holland, 1992.

[4] R. Sproat. Stress assignment in complex nominals for English text-to-speech. In *Proc. of ESCA Workshop on Speech Synthesis*, pages 129–133, Lund, Sweden, 1990.

[5] P. Taylor. The Rise/Fall/Connection model of intonation. Speech Communications, forthcoming, 1994.

[6] P. Taylor and A. W. Black. Synthesizing conversational intonation from a linguistically rich input. In *Proc. ESCA Workshop on Speech Synthesis*, Mohonk, NY., 1994.