Proceedings of the Institute of Acoustics

AUTOMATIC ACCENT PLACEMENT IN ANOMALOUS TEXT SEQUENCES

A. I. C. Monaghan

Centre for Speech Technology Research, Department of Linguistics, University of Edinburgh.

1. INTRODUCTION

It is widely accepted that the main factors which determine intonation are semantics, pragmatics and speaker intent, in ascending order of importance. What this means is that, within the meaningful intonation patterns of their language, speakers can put whatever intonation they want on an utterance regardless of other factors: however, for communicative reasons, the patterns they usually choose are closely related to the semantic and pragmatic structure of the utterance. Rather less widely accepted is the notion that any correspondence between syntax and intonation is largely the byproduct of the relations between both of these and higher-level (semantic, pragmatic, etc.) structures: this is nevertheless the view taken here, and will not be justified in the present paper. For further discussion, see [1-4].

Unfortunately (for those of us working on text-to-speech systems!), there is currently no way of automatically extracting semantic and pragmatic information from text. All that is available to even the most sophisticated text-to-speech systems is a limited amount of syntactic structure and lexical information. It is indeed possible to assign an acceptable intonation contour using only this information: the latest evaluation of the text-to-speech system under development at Edinburgh University's Centre for Speech Technology Research (CSTR) revealed a success rate of around 70% in this task [11]. However, it seems reasonable to assume that the availability of higher-level information would allow significant improvements in performance.

Fortunately, there is at least one group of constructions whose semantic structure can be deduced with some certainty and which are clearly marked in text. Real text contains a high proportion of character strings which are not normal words, and which have therefore been regarded as a problem for text-to-speech systems and are usually either ignored or converted into words by such systems. These are all constructions containing characters other than lower-case letters, referred to as ANOMALIES because of their failure to correspond to the lower-case alphabetic 'norm' for text. Such forms include dates (1/2/34, 1986, '87), number strings (123, 12.34, 12.345, 123456), times (12:34, 12.34pm), and various types of abbreviation (ABC, CoHSE, RSSPCC, Ph.D): in current text-to-speech systems they are generally identified by a preprocessor module which attempts to determine the nature of any anomaly. If this particular subset of anomalies were marked for special treatment by the intonation rules, the amount of information deducible about their structure and function might well be sufficient to allow the consistent assignment of highly natural-sounding intonation to them. It is to this end that the proposals presented in this paper are directed.

The problems of implementing such a scheme are of course considerable, and the textual data is not as regular or as easily interpreted as the preceding paragraphs appear to suggest. It is hoped, however, that such a simplistic approach will prove to be helpful both in improving the intonational treatment of anomalous constructions and in indicating areas where more complex strategies are required. An initial description of the intonational behaviour of these constructions in the context of the CSTR text-to-speech system is presented in the following section, and this is intended to be used as the basis for an automatic treatment of such constructions. The final section of this paper discusses the strengths and weaknesses of this approach in the light of various points raised in this paper.

2. ANOMALIES

There are several distinct classes of anomaly which differ from each other in their intonational behaviour. The present description only addresses five common types: years, times, dates, number strings and abbreviations. (These include all the types which occurred in the CSTR evaluation mentioned above.) In the context of the CSTR intonation rules, the behaviour of each type can be described by answering two questions: what is the relation of this construction to a prosodic domain, and which items within it should receive accents if an optimal accentuation is to be produced? The meaning of these questions requires some explanation.

The accents which should be assigned to a construction are determined by the semantic and pragmatic functions of its constituents, as well as the predicted effect of rhythmic factors on those accents [5]. In our model, accents are assigned to almost all content words and approximately half of these accents are then deleted by the rhythm rule [12]. If a particular domain does not behave in accordance with our rhythm rule, the assignment of accents may need to be modified accordingly; there may be reasonable pragmatic or other grounds for this, as will be seen below.

A minimal prosodic domain is defined [6] as the domain of operation of accent rules. Within such a domain, the rhythm rule is sensitive to the differences in intonational behaviour between predicates (verbs, adjectives) and arguments (nouns) and allows effects such as stress-shift to be modelled [12]. Stress shift or other rhythmic effects across domain boundaries are not permitted, since accents in one domain cannot influence accents in another. If the intonation of a particular constituent depends on a neighbouring constituent, the two constituents should therefore share a domain: if their behaviour is independent of each other, they should be in separate domains. To take a familiar example, the difference between the realisations of "fifteen" in *There are FIFteen MEN in a RUGby team* and *The NUMber fifTEEN is an INteger* (capitalisation indicates accented syllables) results from the prosodic domains involved. In the former, "fifteen" and "men" are in the same domain and so the accent on "men" shifts that on "fifteen"; in the latter, "fifteen" is immediately followed by a domain boundary and so there is no influence from subsequent accents.

2.1. Years

Years (written as four digits, or as two digits preceded by an apostrophe) are always accented on the first and last stressed syllables when pronounced in isolation: only contrastive usage licenses accents on other syllables. Thus, (1) and (2) are the only acceptable non-contrastive accent assignments for these utterances.

(1) NINEteen eighty-NINE

(2) TEN sixty-SEven

However, following material can affect this accentuation by causing the deletion of the second accent, thus:

(3) the NINEteen eighty-nine FESTival

(4) the NINEteen eighty-NINE edinburgh FESTival

The contrast between (3) and (4) is a result of the principle of rhythmic alternation which prohibits accents on adjacent items in the same domain, and therefore years do not necessarily constitute domains in themselves. However, it does not appear to be possible to delete or even to shift the first accent on a year constituent in non-contrastive usage:

(5) *the FAmous nineteen eighty-NINE edinburgh FESTival

(6) *the FAmous nineTEEN eighty-NINE edinburgh FESTival

This suggests that such a constituent must start a domain, but that it may combine with following material (up to the next prosodic boundary). Since rhythmic deletion can apply to these domains, they must be processed by the rhythm rule: marking the unaccented items in the year constituent as pragmatically deaccented (which, arguably, is what they are) will allow the rhythm rule to apply correctly to mimic the observed behaviour of these domains. The details of the implementation of these ideas will not be discussed in the present paper: see Section 2.5.

Time constructions (e.g. 22:10 and 5:55, pronounced as twenty-two ten and five fifty-five respectively) appear to behave in exactly the same manner as years:

(7) it's TEN forty-THREE

(8) the TWELVE thirty-two exPRESS

(9) *the FAmous four FIFty from PADDington

They can therefore be handled by the same rules, although different pronunciations (oh five fifty-five, five to six) or the addition of "a.m." and "p.m." may require special treatment.

2.2. Dates

Constructions giving days, months and years can be pronounced in one of two ways: "1/2/34", for instance, may be expanded to (10) or (11).

(10) the FIRST of the SECOND nineteen thirty-FOUR

(11) the FIRST of FEBruary nineteen thirty-FOUR

These forms appear to share one accent pattern, which simplifies things greatly and avoids the need to choose one or the other. However, the accent pattern for the year as part of a date is not the same as that for a year alone: although "nineteen" also seems quite acceptable with an accent in many cases, this can lead to 'over-accented' intonation as in (12).

(12) *the THIRD of MAY NINEteen TWELVE

Accents are therefore assigned to the first and last accentable items and to the month in these constructions.

Dates appear to constitute domains in themselves, in that their accentuation is not affected by preceding or following constituents. Examples such as (13) are simply ungrammatical, as dates cannot normally function as premodifiers in English, and (14) shows that preceding accents need not affect these constructions.

(13) *The fourteenth of July seventeen eighty-nine events.

(14) on the MORning of the FOURTH of juLY seventeen seventy-SIX

Dates of the form "12/9", meaning the twelfth of September, behave in the same way as those specifying a year except that the accent assigned to the year is not assigned:

(15) the TWELFTH of the NINTH is a TUESday

(16) he's igNORing the FOURTH of juLY

(17) *The twenty-fifth of December celebrations.

2.3. Number Strings

For present purposes, a number string is any string of digits (interrupted only by commas and decimal points in appropriate places) which occurs in text and does not function as a date or similar construction. Number strings are expanded as sequences of cardinal numbers and the decimal point where appropriate.

In general, expanded number strings consist of items which can receive accents and items which cannot. The former include the 'units' zero to nine and the 'tens' ten to ninety; the latter include the words "hundred", "thousand", "million", and so on. Accent patterns are currently generated by assigning accents to all the accentable items and then applying the rhythm rule [5] from right to left to delete every second accent. This results in accentuations such as the following:

(18) SIXteen thousand FOUR hundred and twenty-THREE

(19) TWO million seven hundred and SIXTY-four thousand and TWO

(20) SEVEN hundred and FORTY-two million sixTEEN hundred and eighty-SEVEN

It should be obvious than (20), and even (19), would not be judged entirely natural: however, if instead of ignoring the unaccentable items we allow them to delete the accent of the item which precedes them, we produce the following which are rather more acceptable:

(21) SIXteen thousand four hundred and twenty-THREE

(22) TWO million seven hundred and SIXTY-four thousand and TWO

(23) SEVEN hundred and FORTY-two million SIXteen hundred and eighty-SEVEN

(24) FOUR-hundred and TWENTY-seven thousand eight hundred and forty-THREE

The solution would appear to be a compromise, allowing the accent to be deleted only if there is more than one accentable item before the next unaccentable one: only implementation will show how effective this compromise is.

There are two further problems in deciding which items in a number string should receive accents. The first seems to depend on whether the construction is functioning as a modifier, and the second involves the decimal point.

There appears to be a regular exception to the unaccentability of words such as "hundred", "thousand", and so on: in cases where these words come at the end of a domain (generally, the end of a noun phrase), they can and must be accented. The reason for this behaviour is not clear - it may be that numbers in this position are functioning differently (e.g. not modifying a following noun), or it may be simply the result of the phonological pressure to place accents at the right edges of domains - but the behaviour itself is clear enough:

(25) PROject TWO THOUsand

(26) The POpulation of SCOTland is about FIVE MILLion

(27) PICK a NUMber between TEN and three HUNdred

(28) she was aWARDed SIX hundred THOUsand

This behaviour can be modelled quite easily in the CSTR system by assigning such words to a special syntactic class and checking whether the final item in a domain belongs to this class. This is clearly not a particularly theoretically principled solution, but until the reasons for the observed behaviour have been determined there is some justification for the view that any solution that works is as good as any other.

In number strings incorporating a decimal point, the word "point" never receives an accent in non-contrastive usage: however, it does have the effect of splitting the construction into two sections which behave very differently. The section before the point behaves as though the point were not there, and the section following the point behaves unlike a number string. Thus, we find accents as we would expect before the point but something rather different after it:

(29) TWO hundred and seventy-SIX

(30) TWO hundred and seventy-SIX point FIVE three eight one NINE

(31) TWO hundred and seventy-SIX point FOUR TWO

A strategy of assigning accents to the first and last decimal places appears to produce acceptable accentuations for up to five decimal places. Moreover, the simplicity of this strategy has much to recommend it in an automatic system. Although larger numbers of decimal places than this may sound somewhat unnatural if no accents are interposed between the first and the last place, the rarity of such strings in text allows this problem to be disregarded at least for the present: the accentuation in (30) and (31) is not difficult to achieve within the current CSTR rules.

Something of the relation between number strings and domains should be clear from the above examples, and from well-known examples such as (32). The influence of subsequent material on the accent patterns of number strings in such examples indicates that such constructions do not necessarily constitute a domain in themselves:

(32) FIFteen MEN

(33) there were TWENty-five PEOple on the BUS

(34) the BISHop orDAINED THIRty-seven PRIESTS today

(35) JOHN'S friend PAUL had a BUDget of NINEty-five thousand POUNDS

Examples (34) and (35) indicate that number strings must start a domain, although in informal experiments some listeners judged these (and (25) and (26) above) to be unnaturally over-accented. This is consistent with other anomalies, but again only evaluation of some implementation will show whether a domain boundary is appropriate in most cases.

2.4. Abbreviations

The term "abbreviations" covers a multitude of sins: almost all textual anomalies could reasonably be described as abbreviations. Its usage here is much more restricted, in that it encompasses only those alphabetic anomalies which are not acronyms. By this definition, *NEC, FRCP, B.Sc, Ph.D, RSSPCC* all qualify as abbreviations but *DEC, FRIBA, CoHSE, RS232, 3M* do not. Those which do qualify do not appear to differ significantly in their intonational behaviour, despite variations in their orthographic forms:

(36) EN ee SEE (NEC)

(37) PEE aitch DEE (Ph.D)

(38) BEE ess SEE (B.Sc)

(39) TEE gee double-you YOU (TGWU)

(40) ARE ess ess pee see SEE (RSSPCC)

As with the strings of individually-pronounced digits after a decimal point, these strings of individually-pronounced letters appear to require accents on the first and last items only. The same argument also applies regarding the possible unnaturalness of this treatment for very long abbreviations, although a corpus of 2500 random abbreviations contained only one seven-letter and a handful of six-letter exemplars.

The accent on the final element of an abbreviation can be deleted as a result of subsequent accents as in (41), and this even occurs in very long abbreviations as in (42):

(41) this is the BEE bee see NEWS at nine o'CLOCK

(42) i went to the ARE ess ess pee see see OFFices today

However, the accent on the first element does not seem to be affected by preceding accents:

(43) the SECond EE ee see SUMMit

(44) doctor OWEN'S ESS dee PEE

It therefore seems likely that a treatment whereby an abbreviation starts a new domain but need not finish it will yield appropriate accent patterns for these cases.

2.5. Implementation
The implementation of these ideas, either as extensions of the current CSTR intonation rules or as what would amount to an ACCENT GRAMMAR for such constructions, has not yet been accomplished. Because of the nature of text-to-speech systems, it is essential that the preprocessor should extract as much information as possible from text and pass it on in a form which can be interpreted by subsequent rules: it is inevitable, therefore, that the implementation of rules for the intonational treatment of anomalies will require modifications to the preprocessor and to certain intervening modules, as well as to the intonation rules. Despite this, we foresee no great difficulties in principle in implementing all the above proposals within the framework of the CSTR intonation module.

We are currently engaged in implementing the above rules and evaluating their output in the CSTR text-to speech system. Given our approach to synthesising intonation, which is one of implementing crude but intuitive rules and then using their errors to drive a process of continual improvement and refinement [7], we see no objection to implementing speculative rules such as those above and then testing them on text corpora and using the results to produce a new improved rule set. This strategy allows fast algorithm development, as new or alternative rules can be tried without resorting to extensive data collection and analysis for every modification, and is widely used in cognitive research.

3. DISCUSSION

There are several assumptions and assertions made in the preceding section, some of which deserve more discussion than the present paper allows. This section attempts to point out areas for further investigation and to elaborate briefly on some of the less complex issues raised.

The most obvious question arising from the preceding description is whether such a treatment is appropriate: are the questions of accents and domains really the ones which need to be answered, or should we be looking at the precise function (grammatical, semantic or pragmatic) of anomalies in text and assigning intonation from that? There are two levels of answer to this question. From the standpoint of someone trying to build a machine to do the impossible, i.e. assign natural intonation from unrestricted text, there is a compelling case for holding that any approach which works is a good approach to take. The counter-argument, of course, is that there is little point in trying to make an impossible task easier and that what we ought to be doing is investigating the higher-level factors governing this task and trying thereby to bring it into the realm of the possible: from the standpoint of a theoretical linguist, this is probably the only justifiable course of action. To the extent that it is possible to satisfy both the theorist and the technologist, a compromise would seem to be the best solution for both short- and long-term goals but the issue remains unresolved.

From a practical point of view, the next question is whether the description given is accurate and complete: if it is not, it may be more trouble than it is worth to incorporate it into an automatic system. We take the view that the main purpose of development systems such as those at CSTR is to determine whether a description satisfies these criteria, and to this extent the proof of the theory is in the implementation. This may not meet with the approval of all theoretical linguists, but that is not a problem unique to this theory.

Two further points warrant some discussion: firstly, the issue of extendability to contrastive usage and other explicit exceptions. The rules in the CSTR system explicitly exclude contrastive and emphatic usages, but they are designed to be flexible enough so that when information on such usages is available it can be easily incorporated. In some ways, particularly in the case of the exceptional behaviour of domain-final "hundred", "thousand", etc., the treatment of anomalies seems to qualify as the use of such higher-level information and it might be expected that the mechanisms for recognising and treating anomalies would be capable of extracting other information as well. It must be emphasised, however, that we believe some element of understanding to be essential to the treatment of contrast and emphasis and that the mechanisms suggested for handling anomalies do not incorporate any such element. It is only by virtue of the exceptional textual characteristics of these anomalies that any higher-level information is deducible from them: even the identification of dates or years written out in full, as in *nineteen eighty-nine*, presents major problems on which the present observations have no bearing. Indeed, there is an empirical question to be answered regarding the factors which determine whether, for instance, a date is written in full or as digits: it may be that alternative textual forms correlate with different higher-level specifications and consequently have distinct intonational characteristics. No work on such factors has been carried out to our knowledge.

Secondly, in view of the preceding point, it should be asked whether there are any other classes of items which are readily identifiable from text and whose appropriate interpretation is relatively clear. Two classes of item come to mind in this context: proper names, generally identifiable by their initial capital letter, and punctuation. The latter has already been investigated [5] and partially incorporated into the CSTR intonation routines, and the former is high on the list of areas requiring investigation although the problems involved in interpreting proper names have been described at length by philosophers and linguistics alike, among them Strawson [8], Levi [9] and Sproat & Liberman [10]. There may also be other classes of items amenable to the above approach which will be revealed by corpus analysis in the future: if there are, we hope that our model will be readily adaptable in order to take advantage of them.

4. REFERENCES

[1] E O SELKIRK, *Phonology and Syntax: The Relation between Sound and Structure.* MIT Press, Cambridge Mass. (1984).

[2] D R LADD, *The Structure of Intonational Meaning: Evidence from English.* Indiana University Press, Bloomington Ind. (1980).

[3] D BOLINGER, *Intonation and its Parts.* Edward Arnold, London (1986).

[4] C GUSSENHOVEN, *On the Grammar and Semantics of Sentence Accents.* Foris, Dordrecht (1984).

[5] A MONAGHAN, 'Generating Intonation in the Absence of Essential Information', in W. A. Ainsworth (ed.), *Proceedings of the 7th FASE Symposium* vol. 4 pp. 1249-1256. IOA, Edinburgh (1988).

[6] A MONAGHAN, 'Phonological Domains for Intonation in Speech Synthesis', in J. P. Tubach & J. J. Mariani (eds), *Proceedings of the European Conference on Speech Technology and Communication,* vol. 1 pp. 502-505. CEP, Edinburgh (1989).

[7] A MONAGHAN, 'A System for Left-to-Right Intonation Specification from Text', in J. Laver & M. Jack (eds), *Proceedings of the European Conference on Speech Technology,* vol. 2 pp. 25-28. CEP, Edinburgh (1987).

[8] P F STRAWSON, *Individuals.* Methuen, London (1959).

[9] J N LEVI, *The Syntax & Semantics of Complex Nominals.* Academic Press, London (1978).

[10] R SPROAT & M LIBERMAN, 'Toward Treating English Nominals Correctly', in *Proceedings of the 25th Annual Meeting of the ACL* pp. 140-146 (1987).

[11] A MONAGHAN & D R LADD, 'Evaluating Intonation in the CSTR Text-to-Speech System', in *Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases,* Noordwijkerhout, September 1989, pp. 3.6.1-3.6.4.

[12] A MONAGHAN, 'Rhythm & Stress Shift in Speech Synthesis', *Computer Speech & Language* 4 (1), pp. 71-78 (1990).