Prosodic patterns in spoken English by B. Altenberg, Lund Studies in English 76, Lund: Lund University Press, 1987.

The objective of this study (part of the research project, "Text Segmentation for Speech", in progress at the Survey of Spoken English, Lund University) is "to explore the grammatical predictability of certain prosodic features in English for application in automatic text-to-speech conversion" (p. 11). Altenberg's contribution is primarily one of textual analysis, leading to proposals for predictive rules for the assignment of prosodic features to a written text. He does not report on any attempt to implement his algorithms within an actual text-to-speech system.

The study is divided into eight chapters, of which 1 and 8 are a brief introduction and conclusion. Chapter 2 compares the relationship between speech rate (measured in words/sec), tone-unit (TU) rate (in secs/TU), and TU length (in words/TU) in 10 recordings of different speech varieties in the London-Lund Corpus; the evidence suggests a relative consistency in average TU length over a variety of speech styles, despite large differences in speech rate. Chapters 3–7 all concentrate on a single recording, or text, of around 5000 words, representing one variety of speech: the non-spontaneous, partially scripted monologue. Chapter 3 analyses the distribution over this text of four prosodic features: stress, onset, booster and nucleus, and also studies the relative frequency of the different nuclear tones. Chapter 4 is the core of the book, occupying over 40% of the total text; it looks for correlations between the division into TUs and the grammatical structure of speech. Chapters 5, 6 and 7 explore the grammatical predictability of stress, TU onset and nucleus, respectively.

Any thorough and careful analysis of a text can make a useful contribution to our stock of knowledge, and Altenberg's study is meticulous. He acknowledges from the start that prosodic choices may be primarily motivated by speaker choice and situational factors, and only secondarily by grammatical factors, but justifies his exercise on grounds of expediency familiar to all workers in text-to-speech. In the absence of an adequate pragmatic model, we must do our best with any linguistic knowledge that can be accessed by the system. Increasingly sophisticated automatic parsers will, in principle, be able to supply a grammatical analysis.

One of Altenberg's major achievements must be his extensive refinement and revision of Crystal's (1975) model for assigning TU boundaries to sentences. The revised model is based on an exhaustive top-down grammatical analysis of the monologue text. By distinguishing between a variety of clause types, together with their degree of cohesion and position in the sentence, the author achieves an impressive 93% success rate (95% coverage) in predicting TU boundaries at sentence level for this text. At clause and phrase level, the proportion of boundaries which can be predicted grammatically is much lower but an algorithm applying only in expanded clauses and phrases still has a respectable success rate (70–80%).

In investigating the "prosodic potential" (for stress, onset and nucleus) of the different word classes, Altenberg demonstrates that a traditional classification into (say) "open-class" and "closed-class" items, or even into 12 "primary word classes", produces categories which are highly heterogeneous. A more delicate analysis reveals "secondary" classes which may show a much better correlation with prominence or the lack of it. For example, a subset of "closed" items, including wh-adverbs, ordinals, quantifying and compound pronouns, and "very", have a stress potential exceeding 95%. However, the prognostic value of such an analysis turns out to be limited; the probabilistic rules can

only make predictions for categories where prominence is extremely likely or unlikely, giving a coverage of 62% for stress and only 44% for onset assignment. Nucleus placement is further complicated by "marked tonicity", deaccenting, and "compound tonicity". In simple TUs of more than one word, a rule assigning the nucleus to the very last word apparently works better than the conventional wisdom of selecting the last open-class word. An initial division of all lexical items into "salient" and "non-salient" classes improves the tonicity rule by another 2%, giving a 90% success rate for the *simple* units. Selection of compound tonicity (some 11% of the TUs in the main text), and marked tonicity generally, are speaker-governed and remain grammatically unpredictable.

Altenberg's rules present a considerable challenge to any automatic parser. To operate successfully, they need access to such subtle distinctions as "nonrestrictive" *vs.* "restrictive" relative clauses, and to a highly detailed word-class classification. It is not yet clear just how cost-effective an appropriately advanced parser would be in terms of improved prosodic output. The limited coverage of the rules is a real problem, which will not be solved by extensive analysis of yet more texts. Decisions must still be made about material not covered by the rules. In "scoring" the output of his rules as a percentage of direct matches with the original text, the author does not address the real issue, which is whether the output is *plausible*. The revised algorithms must be applied to *other* texts and the output compared with that generated by cruder algorithms, such as those based on simpler parsers, or on the punctuation of the input text and a basic content/function word distinction. "Plausibility" is highly subjective, but perhaps a more objective measure could be obtained by inviting a number of competent speakers to record an indentical text in near-identical circumstances and using their range of prosodic choices as a yardstick.

A major misgiving about the value of Altenberg's study relates to his choice of text for analysis. He has chosen a prepared, but largely unscripted, talk on the life and history of the village of Stoke Poges—described as "in many ways a suitable authentic model for text-to-speech conversion" (p. 31). But text-to-speech systems, by definition, require *text* as input, fully scripted, normally complete with punctuation. The task is then to interpret the text and deliver it in an intelligible and informative manner. The appropriate analogy must be with a speaker who is reading aloud from a prepared script and who is faced with the same interpretative choices as the machine, not with a speaker who is putting his grammatical and prosodic structures together simultaneously. And the availability of punctuation to guide intonational phrasing is not trivial; real speakers certainly use it, existing text-to-speech systems rely heavily on it and many of the major boundaries predicted by Altenberg's algorithm would normally be marked by punctuation in any case. Why did the author not choose recordings of (say) news bulletins, where a written script was known to be involved? If the study were directed towards a system in which the language itself were automatically generated (some kind of "concept-to-speech" system), then Altenberg's choice might make more sense. Furthermore, we are given no indication as to the variety of English used in any of the texts; but we would surely be wrong to assume that "spoken English" is prosodically homogeneous (e.g. Brown *et al.*, 1980; Knowles, 1984). Listening to the original recording has revealed to me that the main speaker has a local accent, probably not prosodically "marked" by comparison with RP (though he shows what might be considered an idiosyncratic fondness for high rising tones). Readers without access to the recording need this information.

The accuracy of the prosodic transcription supplied by the Survey of Spoken English has apparently been accepted without question, though with some simplification, as has the theoretical framework within which it was made (based primarily on Crystal & Quirk, 1964). At the auditory level, since a complete agreement between transcribers in this field is a rarity, it would be interesting to know whether the author found any non-trivial discrepancies. At the theoretical level a fuller discussion of some of the more controversial issues would be useful. The nature of tone units themselves is belatedly discussed in Chapter 4, where they are presented as a fundamental structural unit in speech, definable in cognitive, textual, prosodic or grammatical terms. One might think, from the prosodic definition given—"a coherent intonation contour optionally bounded by a pause and containing (among other things) a salient pitch movement (the nucleus), normally at the end of the unit" (p. 47)—that segmentation based on auditory analysis was a simple matter, but there are of course many cases where alternative analyses are possible. Accepting TU subordination, whereby one TU is considered to be embedded within another, has obvious implications for the number of boundaries recognised as such. Similarly, a theory which allows compound tones frequently requires the transcriber to decide, in a given instance, between a compound analysis, a sequence of two TUs, or a mononuclear analysis (e.g. the extended fall–rise tone). The transcriber will make a judgment in good faith, recognising that there is often no clear boundary between alternative analyses. Altenberg observes that, functionally, there is often a "deaccented" variant which can commute with the compound analysis (examples, p. 187), but statistically he treats compounds as a separate class, even though they show obvious parallels with "marked" tonicity.

Issues such as these are the subject of continuing healthy debate amongst Survey workers and the theory behind the transcription conventions has evolved over many years; one cannot therefore assume complete theoretical consistency across all the transcribed texts. In statistical surveys, such as Altenberg's study, the discrepancies may be insignificant, but since there are so many competing conventions for transcribing prosodic phenomena, those working in the field of text-to-speech synthesis, or of intonation analysis generally, will always want to be able to reassess someone else's transcription in the light of their own theoretical standpoint. At the very least, a study like this which concentrates so heavily on a single recording, should include the transcribed text, in full, as an appendix, allowing us to reclassify phenomena if we so wish. An accompanying cassette would be even better.

Despite these reservations, the study represents a very real achievement. It demonstrates clearly that good grammatical parsing can *help* us to assign prosodic features to a text but will never give us a complete answer.

The reviewer would like to thank Professor Greenbaum and staff at the Survey of English Usage, University College London, for access to the original recording and transcript of the main text under discussion.

Jill House

References

Brown, G., K. L. Currie & J. Kenworthy (1980) *Questions of intonation*, London: Croom Helm.
Crystal, D. (1975) *The English tone of voice*, London: Edward Arnold.
Crystal, D. & R. Quirk (1964) *Systems of prosodic and paralinguistic features in English*, The Hague: Mouton.
Knowles, G. (1984) "Variable strategies in intonation". In D. Gibbon & H. Richter (Eds), *Intonation, accent and rhythm*, Berlin: de Gruyter.

Knowledge-Based Speech Pattern Recognition by Michael Allerhand. The Fifth Generation Computing Series. Kogan Page.

Michael Allerhand's monograph, published in Kogan Page's Fifth Generation Computing Series, proposes an integration of parametric and structural approaches to automatic speech recognition (ASR) using the formalism of attributed grammars.

The first chapter introduces the problem with its difficulties and discusses possible approaches to ASR. The author supports the use of a knowledge based (KB) approach when *unrestricted* speech recognition is the goal. According to the author, other approaches merely based on statistical methods have the advantage of using learning algorithms with acceptable complexity but have a relatively restricted performance.

Chapter two contains a review of syntactic pattern recognition and fuzzy set theory and introduces an isolated word recognizer in which lexical access is based on the order and duration of only voiced, voiceless and silence features.

Word patterns are described using structural information (segment order) and quantitative information (duration distributions).

A context-free description of syllable structure in terms of features is used. Phonotactic constraints upon sequences of intermediate manner of articulation categories are expressed. The syllabic grammar is used to parse strings of features into manner of articulation classes. Strings of these classes are matched with strings in the lexicon to access particular words.

Chapter three starts with an important statement that seems to be confirmed by recent results in ASR. The author says that the use of simple models creates an inherent performance limitation; a plateau is reached and further improvement cannot be made. Models of sufficient detail cannot be inferred from observations of speech data alone. A context-free grammar is then proposed based on liguistic knowledge obtained by expert linguists.

Methods based on dynamic time warping and hidden Markov models (HMM) are also reviewed and their limitations are discussed.

Chapter four discusses feature extraction through property detectors. Mathematical techniques for feature selection are reviewed. Methods based on hierarchical linear regression are discussed in details and applied to waveform segmentation.

Chapter five contains a useful review of decision algorithms and introduces a composite pattern recognition model for ASR. This model is based on rewriting rules of the form:

$$q_i \xrightarrow{P_{ij}} \{\delta_{ij}\} q_j$$