Proceedings of the Institute of Acoustics

THE DESIGN OF A SPOKEN CORPUS FOR DERIVING PROSODIC RULES

A. I. C. Monaghan

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

1. INTRODUCTION

This paper presents the design of a corpus to be used to derive rules for the automatic assignment of prosody from text in the English language part of the ESPRIT POLYGLOT project. The corpus is specifically designed to take advantage of the considerable amount of data already available on English prosody whilst adding to that existing data in a principled way. The material will be read by professional speakers - male and female - and has been chosen from broad-appeal periodicals, reflecting the commercial aims of the project: it is believed that this will give a representative corpus for the intended applications. The limitations on time and resources available to the project impose a low size limit - the corpus will total about 10,000 words of speech - but the material has been adapted to include numerous examples of rarer prosodic phenomena such as questions, exclamations, lists, parentheticals and extraposed constituents. The corpus consists of two magazine articles, each of c. 1,000 words, as running text, as isolated sentences, and as isolated words. Both the words and sentences are randomised, and additional sentences have been added which contain the rarer phenomena mentioned above. It is expected that this corpus will allow study of sentence-level and discourse-level prosodic effects.

The goal of the POLYGLOT project is to perform real-time text-to-speech (TTS) conversion from any one of seven European languages to any other of those languages. Accordingly, spoken data from all seven languages will be collected for various ends including the modelling of prosodic phenomena. There is a great deal of disparity amongst these languages in respect to the amount of previous research which has been carried out and thus in the level of new research required for POLYGLOT's purposes: in most areas, and certainly in the area of prosody, more literature is available on English than on any of the other six. It was therefore decided that the prosody corpus for English should attempt to fill the gaps in previous research rather than examine aspects of prosody which had already been studied for English but were still unexamined for a language such as Italian or Greek. The corpus is therefore intended to provide information on less common (and arguably more interesting) prosodic phenomena than have generally been investigated, and consequently contains a very high concentration of unusual linguistic constructs.

2. EXISTING ANSWERS AND REMAINING QUESTIONS

While the prosodic phenomena of many European languages are still largely uninvestigated, there is a long and well-documented tradition of such investigation for English. The major factors determining prosodic effects in English have been known for some time: Kingdon [1], Crystal [2] and Halliday [3] in the British tradition, and Pike [4], Bolinger [5] and Trager & Smith [6] in the American school, identified the importance of such factors as focus, context and intention as well as the more or less regular correspondences between grammar and prosody. The contribution of continental research into English has also been significant, with most recently work by Gussenhoven [7], Terken [8], Nooteboom & Kruyt [9] and Baart [10] making major contributions to our understanding of the role of prosody in speech. Many of the questions, therefore, which remain unanswered for e.g. Italian prosody have long since been more thoroughly

investigated for English than the POLYGLOT project's scope would allow (researchers at IPO in Eindhoven alone have devoted enormous resources to the study of English prosody over the last 20-odd years): it is thus clearly unnecessary and inefficient to undertake a re-examination of such questions within the POLYGLOT project.

2.1. Previous Investigations

The issues which can reasonably be said to have already been investigated to a degree which POLYGLOT could not hope to match include the following:

- Syntax-prosody correspondence [11,12]: there is a great deal of information available about the relations between syntactic structure and phonetic phenomena such as fundamental frequency and segmental durations.

- Phonetic correlates of prosodic categories [13,14]: much larger and more representative corpora than POLYGLOT can analyse (IPO have around 1000 "spontaneous and semi-spontaneous" ([13], p. 1250) British English utterances, IBM have over 50,000 words of radio broadcasts) have already provided published results.

- Perceptual factors in prosody [8,15]: the information available on what prosodic effects are perceived and what the categories of effect might be has long been sufficient for various stylised synthesis schemes to be proposed for English.

- Grammar of intonation [10,16]: linguistic constraints on the combination of phonetic or phonological phenomena have also been the subject of extensive research, particularly in the Netherlands.

The fact that so much is known about English prosody compared with the prosodic phenomena of other languages has enabled the implementation of numerous automatic prosody components in commercial and developmental English TTS systems. Some of these, including most of the commercial systems, have no provision for making use of the type of linguistic knowledge which is available to current TTS research, and are thus severely limited in their quality and applicability in that even if the system is told that the heuristics it uses are inappropriate in a particular case it has no way of incorporating such information. There is, however, a growing number of speech output systems making principled use of the higher-level linguistic information which is becoming available both within the system and from the large body of research literature: such systems include those developed at Edinburgh [17], Utrecht [18], IPO [13], IBM(UKSC) [19,20], AT&T [21,22] and British Telecom [23]. The knowledge used in the prosody components of these systems is largely common to all of them, and correspondingly the problems which these systems encounter in synthesising prosody are also shared. Whether one subscribes to the Chomskian view that syntax determines prosody, or the Bolingerian view that speakers determine prosody, the areas which pose problems for synthesis are the same:

> "At this stage, we are not yet in a position to explicitly account for the functional properties of intonation, such as its relation to syntax, semantics or pragmatics. However, we believe that our melodic description provides the necessary groundwork for later linguistic analysis." ([13], p. 1250)

> "Unfortunately, it is not currently possible for any automatic system to perform the syntactic, semantic and pragmatic analyses which are generally acknowledged ... to be essential to natural intonation. To compensate for this lack of vital information, a set of heuristics based on linguistic knowledge has been developed which allows us to mimic semantic and rhythmic structure" ([17], p. 1249)

DESIGN OF A PROSODY CORPUS

2.2. Unresolved Problems

The problem in defining rules for the automatic assignment of English prosody, then, is the effect of grammatical and pragmatic context on the prosodic specification of speech or, more precisely, the influence of higher-level linguistic and real-world knowledge on the desired pronunciation of running text. It is therefore intended that POLYGLOT concentrate on these problematic areas rather than attempting to cover ground which has been thoroughly documented in the literature. There are two main reasons for adopting this approach.

Firstly, and perhaps less importantly given the aims of the project, such an approach will result in a clear addition to our understanding of the role of prosody in speech, rather than merely confirming or at best slightly refining the knowledge already available.

Secondly, and crucially for a project as ambitious as POLYGLOT, this approach will concentrate on the areas where other systems currently fail to perform adequately, giving POLYGLOT the edge over other TTS systems in these areas and advancing text-to-speech synthesis to a point where real text rather than linguists' examples can be given a highly-acceptable prosodic realisation.

3. THE CORPUS DESIGN

Perhaps the most significant departure in the design of this corpus from the standard spoken corpus format is that we propose to base the entire corpus on two sizeable passages of running text, rather than on sets of pragmatically unrelated sentences. This will allow paragraph-level and discourse-level effects to be investigated. The goal of modelling grammatical, careful read speech for eventual synthesis applications (not to mention the massive increase in analysis effort which would be required) preclude the recording of spontaneous speech. In order to obtain fluent, broadly acceptable speech, professional speakers (initially one male and one female) will be recorded: in addition, since read dialogue has little to recommend it over read monologue from the point of view of naturalness - and indeed since it is unclear what degree of naturalness is required from synthesis systems [24] - in the first instance, we propose to record monologue text only.

Despite the fact that the corpus is based on running text rather than isolated sentences, it should be obvious that the recording of all sentences from the texts as isolated sentences as well, and of all words from the isolated sentences as isolated words, is an essential procedure if observations of sentence-level and discourse-level effects are to be possible. These three levels of data will therefore all be recorded, so that the isolated words can act as a "control" for the isolated sentences which will in turn act as a "control" for the same sentences in context. The comparison of prosodic phenomena across different styles of text (formal vs. casual, verbose vs. concise, etc.) is also important if the results of any analysis are not to be limited to a very specific speech style. To this end, we have chosen texts representing very different styles of prose.

The decision to record only one speaker of each sex is not only based on considerations of available resources. The use of a single speaker is desirable in an exercise where his or her prosodic behaviour is to be modelled: it is well known that English speakers vary greatly in the strategies they employ to convey prosodic information, and so it could well be counter-productive to analyse the behaviour of some small (<100) number of different speakers. The choice of a professional speaker is likely to ensure that his or her prosodic behaviour is highly acceptable to a wide spectrum of listeners, and can therefore be usefully modelled for synthesis in commercial systems. In order to ensure that the corpus reflects the properties of actually-occurring text, passages have been chosen from a financial business publication and a contrasting popular magazine. The combined length of the passages will total over 2,000 words of running text for each speaker. A further factor in the choice of texts was that they should not be overly topical, to avoid disinterested productions by the speakers.

DESIGN OF A PROSODY CORPUS

In addition to recording each sentence in isolation (in random order, and before recording the passage as a monologue, of course), further examples of unusual sentence types (e.g. exclamations, all-given sentences, counter-assertions, etc.), which are by definition those on which least research has been done, will be recorded in isolation in order to provide a more representative sample of such phenomena. The details of how these examples have been generated are given below.

3.1. Corpus Materials
The detailed composition of the material making up the corpus is as follows:

A. 2 x 1,000 word passages of different styles, chosen from (a) *The Economist* and (b) *Cosmopolitan*, read as running text. The former is a technical article on Japanese corporate business, and contains quite complex syntactic and lexical constructions: the latter is a general-interest biographical article on a film starlet of the 1920s. These two articles have been chosen so as to include numerous examples of anaphora, lists, parentheticals, epentheticals, tags, counter-assertions, and other phenomena of prosodic importance.

B. Each sentence from the passages in (A) spoken in isolation. (Some sentence fragments, e.g. those bounded by colons, should also be recorded in isolation as well as in the text sentence.) This will allow a comparison between sentence-level and paragraph-level prosodic effects.

C. Additional sentences containing less frequent phenomena such as questions, contrastive accent, exclamations, commands, vocatives, etc., as well as further examples of the phenomena mentioned in (A). All such phenomena should be represented by at least 10 occurrences in total. These additional sentences have been derived directly from the sentences in (A), as is discussed below.

D. Each word from the sentences in (A) and (C) spoken in isolation. This will allow a comparison between word-level and sentence-level prosodic effects.

3.2. Recording Procedure
All the above are to be spoken by professional RP speakers under standard recording conditions (anechoic chamber, 2-channel digital recording, good-quality microphones, synchronised larynx trace, etc.), with multiple repetitions and training sessions to ensure a natural reading style (i.e. relaxed, fluent, not list-intoned). As mentioned above, the words will be recorded before the sentences and the sentences before the running text. The sentences have been carefully randomised so as to bear no relation to their order in the text passages and to reduce any apparent coherence from one sentence to the next as much as possible. The impression of textual coherence has been reduced to a point where we feel that the use of additional techniques such as introducing unrelated sentences is unnecessary for this particular corpus: if we had been intending to record longer texts, however, some such device would have been necessary.

3.3. Samples of Corpus Material
As stated above, the two 1,000-word texts are taken from actual published material. The first paragraph of each text is given as an example of style and content.

The Economist, 10-16 February 1990 p. 75:
Japanese companies, accustomed to booming profits in recent years, are worrying about what to do now that profits have begun to sag. Companies' plans for diversifying into higher value-added products, streamlining operations at home, building additional factories abroad and acquiring yet more foreign assets all depend on a continuation of the healthy earnings growth of the past few years. But recent forecasts suggest the profit stream that has propelled Japanese companies since the high-yen recession of 1985-86 is slowing -- and more abruptly than anyone expected.

DESIGN OF A PROSODY CORPUS

> *Cosmopolitan*, February 1990 p. 106:
> The meteoric rise and fall of Louise Brooks astonished even herself. She was a symbol of the Twenties, a lithe spitfire with a jet-black bob and electrifying eyes. It is impossible to name another film actress who made such an erotic and lasting impact in such a small number of films.

There are obviously very few examples of rarer prosodically-important text phenomena (such as those mentioned at (C) above) in any 2,000 words of real text: it is therefore necessary to create additional examples of such phenomena. This has been done in as principled manner as possible, however, by taking a sentence from the original text and "transforming" it to include a particular rare phenomenon. For instance, to produce an additional example of a particular type of prosodically-marked parenthetical we took the sentence

> (1) By the time Louise was 10, she was already a child-prodigy dancer.

and inserted the phrase *you know* into it. There are three relatively natural insertion sites in this sentence, illustrated by sentences (2a-c):

> (2a) You know, by the time Louise was 10, she was already a child-prodigy dancer.
>
> (2b) By the time Louise was 10, you know, she was already a child-prodigy dancer.
>
> (2c) By the time Louise was 10, she was already a child-prodigy dancer, you know.

Sentence (2b) corresponded to the embedded version which we were aiming to create, and it is clear that it has been derived from (1) with the minimum of unnaturalness and uncontrolled variation being introduced. With both sentence (1) and sentence (2b) spoken in isolation in the corpus, any prosodic difference between them can confidently be attributed to the presence of the embedded parenthetical construction *you know*.

The other major body of text to be incorporated in the corpus, then, is the isolated sentences from both texts, suitably randomised, with the addition of the supplementary sentences derived as above. There are 120 isolated sentences in total: the first few of those randomised from the *Economist* article are given as an example here:

> The best solution is for Japanese firms to sell unwanted subsidiaries and buy firms which fit with their core business, even if this requires a hostile takeover.
> A further increase in interest rates is on the cards once the general election on February 18th is out of the way.
> Asahi Breweries recently "unbundled" its Nikka Whisky subsidiary in this way.

Finally, there are 945 isolated words taken from the isolated sentences, including letters and digits pronounced in isolation.

3.4. Corpus Annotation & Use

The spoken corpus will be transcribed at various levels. The syntactic and pragmatic description will consist of manual parses and annotations. The prosodic transcription will include at least two sources of information: experts' annotations of pitch accents, and naive listeners' perceptions of boundaries. The former will be as theory-independent and as detailed as possible, e.g. marking an accent-lending full rise rather than a H* accent or just a rising pitch: the latter (suggested by Lou Boves of Nijmegen University) is intended to avoid prejudging the nature of prosodic domains and to provide data on what listeners perceive rather than on what linguists expect. The phonetic transcription, providing both segmental and durational information, will be performed by CSTR's HMM-based automatic segmentation software [25].

The corpus will be processed and accessed via APS, an acoustic-phonetic software environment developed at CSTR [26]. The corpus statistics will be used for basic research on prosodic events and also for developing statistical rules for assigning prosody to English text in the POLYGLOT speech output system.

4. CONCLUSIONS

We are confident that this procedure will produce a corpus containing a representative sample of all the interesting prosodic phenomena in standard British English which are relevant to read text, and moreover that it will do this without unnecessary repetition of previous work. Such a corpus will represent a concentration of prosodic data which is ideally suited to linguistic analysis and which to the best of our knowledge fills a persistent and crucial gap in speech corpora. It will therefore permit analyses of factors not currently considered or incorporated in text-to-speech systems. We believe that, compared with either collecting material along the lines of the LOB corpus or carefully constructing artificial texts, this is a much more efficient and effective approach to collecting a prosody corpus for English and indeed the only practicable way for small projects such as POLYGLOT to do so.

5. REFERENCES

[1] R KINGDON, *The Groundwork of English Intonation.* Longmans, London (1958).
[2] D CRYSTAL, *Studies in the Prosodic Features of Educated Spoken British English, with Special Reference to Intonation.* Ph.D. Thesis, University of London (1966).
[3] M A K HALLIDAY, *Intonation and Grammar in British English.* Mouton, The Hague (1967).
[4] K L PIKE, *The Intonation of American English.* University of Michigan, Ann Arbor Mi. (1945).
[5] D BOLINGER, 'A Theory of Pitch Accent in English', Word 14 pp. 109-149 (1958).
[6] G L TRAGER & H L SMITH, *An Outline of English Structure.* Battenburg Press, Norman Ok. (1951).
[7] C GUSSENHOVEN, *On the Grammar and Semantics of Sentence Accents.* Foris, Dordrecht (1984).
[8] J M B TERKEN, *Use and Function of Intonation: Some Experiments.* Ph.D. Thesis, University of Leiden (1985).
[9] S G NOOTEBOOM & J G KRUYT, 'Accents, Focus Distribution, and the Perceived Distribution of Given and New Information: An Experiment', JASA 70 pp. 1512-1524 (1987).
[10] J BAART, *Focus, Syntax and Accent Placement.* Ph.D. Dissertation, University of Leiden (1987).
[11] E O SELKIRK, *Phonology and Syntax: The Relation between Sound and Structure.* MIT Press, Cambridge Mass. (1984).
[12] M NESPOR & I VOGEL, *Prosodic Phonology.* Foris, Dordrecht (1986).
[13] N WILLEMS, R COLLIER & J 't HART, 'A Synthesis Scheme for British English Intonation', JASA 84 pp. 1250-1261 (1988).
[14] S J EADY & W E COOPER, 'Speech Intonation and Focus Location in Matched Statements & Questions', JASA 80 pp. 402-415 (1986).
[15] K SILVERMAN, *The Structure and Processing of Fo Contours.* Ph.D. Thesis, University of Cambridge (1987).
[16] J 't HART & R COLLIER, 'Integrating Different Levels of Intonation Analysis', Journal of Phonetics 3 pp. 235-55 (1975).
[17] 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).
[18] H QUENE & R KAGER, 'Automatic Accentuation and Prosodic Phrasing for Dutch Text-to-Speech Conversion', in J. P. Tubach & J. J. Mariani (eds), *Proceedings of the European Conference on Speech Technology and Communication*, vol. 1 pp. 214-217. CEP, Edinburgh (1989).
[19] A D BELL, 'Towards Assigning Prosodic Patterns in Speech Synthesis', in J. Laver & M. Jack (eds), *Proceedings of the European Conference on Speech Technology*, vol. 2 pp. 169-172. CEP, Edinburgh (1987).

[20] W N CAMPBELL 'Syllable-Level Duration Determination', in J. P. Tubach & J. J. Mariani (eds), *Proceedings of the European Conference on Speech Technology and Communication*, vol. 2 pp. 698-701. CEP, Edinburgh (1989).
[21] E FITZPATRICK & J BACHENKO, 'Parsing for Prosody: What a Text-to-Speech System Needs from Syntax', in *Proceedings of the Annual AI Systems in Government Conference*, pp. 188-194. IEEE Computer Society Press, Washington DC (1989).
[22] J B PIERREHUMBERT, 'Synthesising Intonation', JASA 70 pp. 985-995 (1981).
[23] J LOCAL, 'Modelling Assimilation in Non-Segmental, Rule-Free Synthesis', paper given at the Second Conference on Laboratory Phonology, Edinburgh, July 1989.
[24] J E BLAUERT & E SCHAFFERT, *Automatische Sprachein- u. Ausgabe*. Bundesanstalt f u.br [25] F R McINNES, Y ARIKI & A A WRENCH, 'Enhancement and Optimisation of a Speech Recognition Front End Based on Hidden Markov Models', in J. P. Tubach & J. J. Mariani (eds), *Proceedings of the European Conference on Speech Technology and Communication*, vol. 2 pp. 461-464. CEP, Edinburgh (1989).
[26] G S WATSON, 'APS: An Environment for Acoustic Phonetic Research', in J. P. Tubach & J. J. Mariani (eds), *Proceedings of the European Conference on Speech Technology and Communication*, vol. 2 pp. 300-303. CEP, Edinburgh (1989).