From: Proc. Institute of Acoustics Conf. Speech and Hearing, 1996

STATISTICAL ANALYSIS OF SYNTAX_PROSODY RELATIONSHIPS USING THE PROSICE CORPUS

M. Huckvale & A.Fang

Department of Phonetics and Lingustics, University College London

1. INTRODUCTION

PROSICE is a corpus of spoken material specifically designed for the study of the prosody of read English and for technological applications. The corpus uniquely combines high-quality signals with a detailed set of annotations. It combines the best ideas from a number of existing British English spoken corpora: the anechoic recording conditions of EUROM0 [1], the balanced source texts of the London-Lund corpus of spoken English [2], and the aligned syntactic descriptions of the MARSEC corpus [3]. In addition, the PROSICE corpus provides an accurate fundamental frequency contour generated from a simultaneous Laryngograph signal, a set of word alignments and pause annotations, and sufficient data from a single speaker to build statistical models. This paper briefly describes the design aims and content of the corpus, it describes the system used for gramatical annotation, it proposes a radically different method for the modelling of prosody for text-to-speech applications, and it gives preliminary results of some correlations between grammatical descriptions and prosodic phrasing found in one of the corpus texts.

2. CORPUS DESIGN

The study of prosody is assuming a greater importance in phonetics, phonology and speech technology than ever before. The recent past has seen novel views of the phonology of intonation [e.g. 4], a new interest in prosodic phrase structure and prominence [e.g. 5], and the rise of non-linear accounts of phonetic substance [e.g. 6]. In speech synthesis, the success of concatenative systems has meant that the key issues have shifted from segmental to suprasegmental quality [7]. In speech recognition, the increasing emphasis on dialogue has meant more research is taking place into the use of prosodic structure for the purposes of disambiguating utterances [e.g. 8].

Contemporaneously with this renewal of interest in prosody has been the increasing influence of corpus studies, driven by the success shown by speech recognition systems based on statistical modelling. For example, the prediction of segment durations in text-to-speech systems is now commonly made from regression analysis of transcribed speech [9].

Unfortunately, the currently available corpora for the study of the prosody of English all have

limitations which make them less than ideal for technological applications. The EUROM0 corpus has fine phonetic detail and good fundamental frequency information, but is very small and has no prosodic or grammatical analysis. The SCRIBE corpus contains spontaneous speech, but is not annotated at all. The LLC corpus has manually transcribed prosodic annotations, but no quality annotated signals. The SEC corpus does not have time-aligned annotations, while the MARSEC corpus, which is otherwise the most sophisticated, has prosodic and grammatical annotations, but is based on a rather limited syntactic analysis of a rather mixed group of speakers. The MARSEC signals are also only of radio broadcast quality. Perhaps of these, only the LLC corpus has been used extensively for prosodic research so far [10, 11].

Our objectives with the PROSICE corpus have been to combine the best features of the existing databases, to keep in mind what characteristics are required for technological applications of prosody, and to work within certain resource limitations. The design objectives can be summarised as: (i) high quality recordings, (ii) accurate fundamental frequency information, (iii) genuine spoken texts, (iv) annotations at word, wordclass and syntactic levels linked to signal, (v) relevant phonetic annotations, (vi) sufficient quantity for statistical modelling. However because of current resource constraints, we have been obliged to (i) re-use existing grammatical analyses, (ii) avoid all manual annotation, (iii) fit everything onto a single CD-ROM.

The impracticality of manual annotation has meant that we can not currently provide a phonological level of prosodic annotation. This limitation is potentially very serious in that it precludes the study of the relationships between the typical division of levels found in scientific research in prosody [e.g. 12] where syntactic information and phonetic information are each projected onto an intermediate phonological level. On the other hand, the imposition of a phonological level also brings with it difficulties for technological applications. An intermediate representation, such as the ToBI system of mark-up [13], means that the prediction of prosody in a text-to-speech system must fall into two stages, syntactic to phonological and phonological to phonetic. Since the phonological level is not derived automatically from the text or the signal it suffers from both theoretical and practical weaknesses: any phonological level supposes a model of prosody which may be inadequate or inflexible, and any set of manual annotations is prone to subjective error.

The problem of the intermediate phonological level of annotation has been solved in speech recognition. There, a phonological description of a word as a sequence of phones may be used to help identify the words in the lexicon, but without a requirement that training material be annotated at a phonological level. The statistical re-estimation methods used in speech recognition assume a lexical to phonological mapping that is used implicitly for the interpretation of the phonetic material: given that an utterance is supposed to have this phonological transcription, interpret the data as if this was so. In a sense, the training data is analysed according to a generative model, and then the parameters of this generative model are updated. We believe such an approach could also be exploited for prosody. In our work on PROSICE we have relied on the grammatical annotations and the phonetic annotations alone. We aim to predict the lowest-level characteristics as far as possible from the time-aligned grammatical description. Our preliminary results for phrasing are

detailed below. For intonation, we would expect to require a stylised fundamental frequency contour, but propose to use an automated system to obtain such a level of description.

The texts used in PROSICE so far have been taken from ICE-GB, the British English component of the International Corpus of English [13]. We have chosen the category of broadcast talks (S2B-021 to S2B-040) where speakers are the actual composers of their deliveries, where the prosodic phrasing and phrasal prominences have their intended meaning. We have made new recordings of 8 of these texts using a single speaker in a controlled acoustic environment. The speaker was not aiming to create a copy of the original, but to create a similar fluent style of presentation. All disfluencies and deviations from the original text were incorporated in the annotations.

The word alignments were performed automatically using a whole-word template recogniser (all details in [14]). The final recordings are presented as 16-bit 20,000 samples/sec signals on CD-ROM, and occupy approximately 2 hours of speech. Associated laryngograph recordings and fundamental frequency contours are also included. The grammatical annotation scheme is described in the next section.

3. GRAMMATICAL ANNOTATIONS

The necessity of combining grammatical analysis with phonetic analysis for the study of prosody has been repeatedly voiced in the past 50 years. Numerous investigations have been carried out to identify, for example, the correlation between canonical grammatical categories and prosodic consituents [e.g. 15, 16, 17, 18, 19]. However failings in the classification of linguistically relevant properties has meant that even such essential aspects as grammatical juncture have not been properly defined nor systematically distinguished [20]. For example, when investigating pauses as linguistic demarcators, Stenström [21] classifies pauses only in terms of whether they are found between sentences, clauses, clause elements or phrase elements. The inadequacies lie in the facts that clause elements subsume both clauses and phrases, and that simply identifying the number of pauses between verb and object ignores variety conditioned on sub-divisions of noun or clause objects.

It is thus desirable to annotate prosody corpora with a descriptive grammar formalism that can be consistently applied to the kinds of authentic data included in the corpus. Since the analysis of phrase structure alone is known to be inadequate for the explanation of prosody, the formalism should also provide syntactic functions such as subject, verb and complement as well as categories of word class, phrase and clause. We believe that a more successful attempt at modelling prosody will arise from a more sophisticated grammatical analysis which would include explicit indications about the syntactic functions of each phrase.

The formalism used for PROSICE has been empirically tested and modified through application in the million-word ICE-GB corpus. It is also the basis for a computational parsing system developed

by one of the authors [22]. The possibility of producing the grammatical formalism automatically, makes applications of the results of studying the corpus immediately practicable. The formalism is based on the wordclass and syntactic parsing scheme developed by the TOSCA research group at Nijmegen University. [23,24]. In the course of tagging and parsing the ICE-GB corpus it was substantially modified, and so we now refer to it as the ICE annotation scheme. Details may be found in [25].

The ICE scheme has 20 basic word class tags, complemented with features which extend the total number of types to over 250. Phrases are analysed into five categories: noun, verb, adjective, adverb and prepositional. Clauses are identified as subject, verb, object, complement or adverbial; they are further described by features of coordination, verb form, verb type, mood, pragmatic function, subordination, tense and voice (among others). We give an example annotation for a short section of the aligned corpus:

Figure 8 from SHL paper

4. PRELIMINARY STUDY

We have investigated one of the PROSICE texts for the correspondence between major (>0.5s) and minor (<0.5s) pauses with grammatical categories.

The experiment was based on the first recorded text S2B-025, which is entitled 'For he is an Englishman' broadcast on Radio 4 on 5 November 1990. The text has 2014 words in 132 parsing units (approximately 'sentences'). Our re-recording lasted 11 minutes, representing a rate of 180 words per minute - a rather rapid and fluent production. There were 309 identifiable pauses found as a byproduct of the automatic annotation scheme. The co-occurrence of pauses with the start of syntactic constituents were thene xamined. The outcome is a remarkable correspondence between prosodic phrasing and syntactic phrasing for this material; only 5.8% of pauses did not co-occur with one of the major formal categories. The table below lists the frequencies of the categories and any related pauses.

Category	Frequency	Pauses		Correlation
Sentence	150	150	48%	100%
Clause	106	38	12%	36%
Verb Phrase	273	38	12%	14%
Prepositional Phrase	197	20	6%	10%

STATISTICAL ANALYSIS OF PROSODY WITH PROSICE

Adverb Phrase	109	9	3%	8%
Adjective Phrase	161	8	3%	5%
Noun Phrase	581	28	9%	5%
Other		18	6%	
Total	1577	309	100%	

The first column lists the type of syntactic categories, and the second gives the observed frequency in the single text. The third column is the observed frequency of pauses co-occurring with the start of that category. The final column expresses the percentage of that category which had co-occuring pauses.

From the data one can observe that all sentences were preceded by a pause, compared to only 36% of clauses. Of the phrases with initiating pauses, there were some intriguing characteristics which need to be better tested on the whole corpus. For example, many of the pauses initiating verb phrases were preceded by a heavy subject, as in:

":GAP: A two-thirds majority but no appeal :GAP: was needed for a lamp post in a street."

Prepositional phrases with initiating pauses were typically those that can be termed as complex prepositions and complex conjunctions, e.g.

":GAP: The latter feared the boroughs :GAP: because of their masses of urban voters :GAP: and were looking for allies."

":GAP: These ways were left out of the act :GAP: with the result that astute gravel merchants and estate developers :GAP: exploit loopholes in the law :GAP: to discommon and ravage."

The relatively high proportion of adverb phrases with pauses could be conveniently explained by the use of adverbs as sentential adjuncts.

These correlations are only possible to determine because of the availability of the syntactic function categorisation in the ICE annotation scheme. Without introducing such functions, we could only expressed our correlations in terms of 'heavy' NP. ???

In summary, the preliminary study confirms the view that pauses are reliable demarcations of sentence structure in fluent read speech, and that the ICE annotation scheme may be of particular utility.

5. CONCLUSIONS

The findings of even the preliminary study have already been found to be useful within a computational linguistic application. The correlation of heavy subjects, complex prepositions and adverbials with pauses has contributed to the design of the SpeechMaker software - an automated system that proposes prosodic phrasing for foreign readers of English [26].

6. REFERENCES