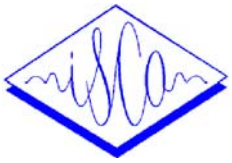

4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

Word Class Driven Synthesis of Prosodic Annotations

Simon Arnfield
(S.C.Arnfield@Reading.ac.uk)

Department of Linguistic Science, The University of Reading
Whiteknights, Reading, UK. RG6 6AA

ABSTRACT

Prosody is an important aspect of speech that current text to speech synthesis systems fail to mimic in a convincing or natural way[1, 2, 3, 4]. This paper describes research on a partial system for prosodic synthesis using easily derived low level syntactic information.

A computer program has been developed that can annotate unseen text with prosodic stress and tone marks using the sequence of part of speech tags previously assigned to each word by a tagging system.

Training and testing material was taken from the Lancaster/IBM Spoken English Corpus (SEC). Co-occurrence measures were calculated relating stress and tone mark annotations to the word class annotation information. A model was developed around the statistical information which calculates a score for all possible mappings between a given part of speech sequence and all the potential stress/tone annotations. The highest scoring pattern is selected as that which is the most likely "baseline" annotation, according to the model. Performance figures attain up to 91% agreement with the original corpus annotations.

1. INTRODUCTION

Few speech synthesis systems model prosody in a convincing way. Those systems that do have a prosodic element tend to use a devised rule–based system as opposed to any model of prosody derived empirically from real–life data.

The research described in this paper is a first step towards a model of prosody that is statistically based and derived from real–life corpus data.

2. CORPUS PROCESSING

Statistical measures were derived from a corpus of prosodically and part of speech annotated spoken English. These measures were used to drive calculations for a model that predicts prosodic annotations.

2.1. The Spoken English Corpus

The Lancaster/IBM Spoken English Corpus (SEC)[5, 6] was compiled between 1984 and 1985 at the Unit for Computer Research on the English Language (UCREL), University of Lancaster, and the Speech Research Group at IBM UK Scientific Centre, Winchester. The corpus was collected mainly from BBC Radio 4 broadcasts and is available as lexicographically transcribed texts (with and without punctuation), as part of speech annotated texts, and as prosodically annotated texts.

The part of speech annotations in the corpus were assigned at the University of Lancaster using their CLAWS[7, 8] tagging program which was first developed between 1981 and 1983 at the Universities of Lancaster, Oslo and Bergen.

The prosodic annotation scheme used in the SEC is often referred to as "standard British prosodic annotation" which is a variation on that described by O'Connor and Arnold[9]. Prosodic annotation was performed by two expert transcribers.

2.2. Statistical Measures

If a number of "measures" of the relationship between parts of speech and prosodic annotations can be collected we can combine these measures together. Each differing measure of likelihood of relationship forms a constituent of the overall measure of relationship. Using a number of such constituents to relate one entity (such as part of speech) to another (prosody) is what Atwell[8] has called constituent likelihood.

The prosodic and part of speech versions of the SEC were combined by specially written software that allowed for variations in representation between the versions. For reasons of annotation some parts of the corpus were adjusted. For example "19" might be changed to "nineteen" to allow for stress annotations to be placed on the correct syllable — or in the part of speech annotation enclitics are treated as more than one word, for example "won't" becomes "will n't". These alterations were not carried over to the other versions

of the corpus.

The combination of this data allowed calculations to be made to show the total number of occurrences of a given part of speech with a given prosodic annotation. Due to the large number of tags used for parts of speech in the CLAWS system it is unsurprising to discover that some tags occur very infrequently — or not at all. This leads to gaps in the table of co–occurrences. However, for the majority of tags it is possible to estimate likelihoods for co-occurrence.

Bi–gram frequencies were also counted and likelihoods estimated for sequences of a part of speech tag with a given prosodic annotation followed by another part of speech tag with another given prosodic annotation.

A shortage of corpus data and hence subsequent poor estimates for likelihoods (for those elements that were very infrequent) forced the decision to group together some of the annotations. In the most extreme case this resulted in just two prosodic annotation groups dubbed *stressed* and *unstressed*. In other cases some part of speech tags were grouped.

3. PROSODIC MODEL

Two models were developed. Both models function in a similar way. Input data takes the form of sequences of part of speech tags which are associated with actual words. Output takes the form of annotations on a word by word basis (there is a general assumption that the syllable receiving annotation in a word can be identified by rule and lexicon lookup). The difference between the models is in the extent to which the prosodic annotations are specified.

The first model classifies output annotations as either stressed or unstressed (stressed refers to any type of prominence). The second model attempts to further specify stress by classifying a number of tone directions (rise, fall, fall–rise, and level).

3.1. Performance Measures

It is possible for a sentence to be stressed in different ways in different texts (contexts). A prediction based on sentence–syntax, without any model of "text grammar" or inter–sentential cohesion cannot hope to work perfectly. This leads to a problem of evaluation if the predicted stress is different from that in the corpus — it need not necessarily be wrong. Consider the phrase "John isn't here". Depending upon the stress pattern a number of meanings are possible. If "John" were stressed it might mean everyone in a group were present except John. If "isn't" were stressed the utterance might be used to correct someone's misconception about the presence of John, and so on.

Given an utterance of length n and x possible annotations there are x^n possible annotation sequences. In the measures of performance used here only those results that matched annotations given in the corpus were considered correct. The results are therefore conservative.

3.2. Utterance "Scoring"

The models assign values (called scores) to each of the possible sequences for a given sequence using the following formula

$$score = \prod_{n=1}^{w} S(w_n, a_n) \times \prod_{m=2}^{w} B(w_{m-1}, a_{m-1}, w_m, a_m) \quad (1)$$

Where $S(p, q)$ is the likelihood measure that word class p would have annotation q, w_n is word class for word n, and a_n is the annotation of word n. $B(p, q, p', q')$ is the likelihood measure of the bigram of word class p followed by word class p' where p has an annotation q and p' has an annotation q'.

As an example Figure 1 shows a three word utterance with examples of the appropriate values for the sequence "at FORD MOTORS". Here S(II,unstressed) = 0.88, S(NP1,stressed) = 0.91, S(NN2, stressed) = 0.92, B(II,unstressed,NP1,stressed) = 0.39, and B(NP1,stressed,NN2,stressed) = 0.23. The product of these values gives the score for that sequence. Once all score values have been calculated the "most likely" sequence is taken to be that with the highest score.

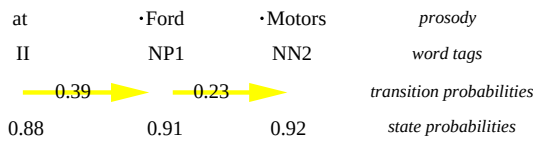

Figure 1: Example values used in model calculations.

For the second model there are five prosodic annotations but it was found that the likelihoods for the "untressed" words overpowered those of the different stress tones meaning that the model performed poorly on the stressed/unstressed distinction. The second model was therefore altered to boot–strap from the results of the first. That is, unstressed words were identified by the first model and constrained to remain unstressed whilst the values for the type of stress on the stressed words was calculated.

4. RESULTS

The first model, that predicted stress pattern sequences, performed well with figures of 91% agreement with the original corpus annotations. Table 1 shows results for a number of categories of data in the SEC. More detailed results are given in Arnfield[10].

The second model, which attempted to predict the actual tone marks on stressed words, did not perform as well at 65% agreement with corpus annotations. The distinction between some tone marks (fall, rise and fall–rise) was very

Category	%
Commentary	91
News Broadcasts	92
Lecture(general)	90
Lecture(specialist)	93
Magazine Reporting	90

Table 1: Performance statistics for stress prediction model. Percentage of words which are correctly stressed/unstressed.

poor. Fall tones were often predicted in place of rise and fall–rise tones.

The model had particularly poor tone prediction performance with nouns, adjectives, lexical verbs, adverbs and determiners but worked well with articles, prepositions, conjunctions, pronouns and non–lexical verbs. This is hardly surprising since the majority of semantics in a sentence is probably carried by the first group of words. This model alone cannot hope to encapsulate any semantic information.

Analysis of the performance of these models on a word class tag basis has shown that the performance for similar word class tags is very close suggesting that the fine level of distinction in CLAWS tags can be significantly reduced. This would allow likelihood figures to be improved as data from different classes could be combined and hence would help to alleviate the problem of there not being enough data in the corpus.

The results did indicate that a model that attempted to predict just three classifications (unstressed, level stress, and stress with tone movement) might perform well and this is the focus of a current research project.

5. REFERENCES

1. Alex Waibel. Prosodic knowledge sources for word hypothesization in a continuous speech recognition system. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 534–537. Morgan Kaufmann, 1990.
2. Denis H. Klatt. Scriber and lafs: Two new approaches to speech analysis. In Wayne A. Lea, editor, *Trends in Speech Recognition*, chapter 25. Prentice–Hall, 1980.
3. Denis H. Klatt. Review of the arpa speech understanding project. In Alex Waibel and Kai-Fu Lee, editors, *Readings in Speech Recognition*, pages 554–575. Morgan Kaufmann, 1990.
4. Wayne A. Lea. Prosodic aids to speech recognition. In Wayne A. Lea, editor, *Trends in Speech Recognition*, chapter 8. Prentice–Hall, 1980.
5. Gerry Knowles. The spoken english corpus: A progress report. *ICAME Journal of the International Computer Archive of Modern English*, 1988.
6. Gerry Knowles and Lita Taylor. *A Manual of Information to Acompany the SEC Corpus*. UCREL, The University of Lancaster, 1988.
7. Roger Garside, Geoffrey Leech, and Geoffrey Sampson, editors. *The Computation Analysis of English: A Corpus–Based Approach*. Longman, 1987.
8. Eric Atwell. Constituent-likelihood grammar. *ICAME Journal of the International Computer Archive of Modern English*, 7:34–67, 1983.
9. J. D. O'Connor and G. F. Arnold. *Intonation of Colloquial English*. Longman, London, second edition, 1961.
10. Simon Arnfield. *Prosody and Syntax in Corpus Based Analysis of Spoken English*. PhD thesis, The School of Computer Studies, The University of Leeds, 1994.