Comparison of algorithms for predicting accent placement in English speech synthesis

Alan W Black
ATR Interpreting Telecommunications Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN
awb@itl.atr.co.jp

Introduction

Accent placement in English speech synthesis is important in producing natural sounding speech. This work compares three different techniques for predicting where accents may be placed: a heuristic-based algorithm, a prosodic phrase theory-based algorithm, and automatically learned decision trees. Their predictions are compared with respect to the Boston University FM Radio database. Although a simple scoring measurement places the heuristic-based and decision tree algorithms equally ahead, after further classifying, the "errors" made by the heuristic-based algorithm are considered to be less serious.

Accents in English speech

In order to synthesize natural sounding speech, it is necessary that it contain varied prosodic events in appropriate places. This paper concentrates on one part of the synthesis of prosodic variation. Syllables in English may be *accented*. By *accented*, in this paper, we mean that there is a distinct change in the fundamental frequency contour (F_0) around that syllable. It may be a rise, or a fall, or a fall and a rise, or most typically a rise followed by a fall. In the work presented here, we are interested in predicting if such an event exists or not, rather than its type or size. The type or size of an accent is considered as a separate problem and we have developed alternative algorithms for predicting those variations.

The algorithms were tested with respect to Boston University's FM Radio database. The data used consists of around 45 minutes of radio news messages from one female speaker of American English. The speaker is a professional speaker and uses a characteristic "news announcer" style. The data has been automatically phonetically labelled and part of speech and break levels have been hand assigned. Most importantly, for this work each syllable has been hand assigned a label "prominent" or "non-prominent" which we have interpreted to mean accented or not. It is those "prominence" markers that we are trying to model.

Three algorithms were tested. The object was to see if an automatically learned model would be significantly worse or better than either a set of heuristics or a more carefully designed theoretically-based set of rules.

Hirschberg

This is a heuristic-based algorithm based on the work of [3]. It assigns one of four features to each word, *emphatic*, *accented*, *deaccented* or *cliticised*. For the purposes of this test we assume *emphatic* and *accented* to imply accented and *deaccented* and *cliticised* to imply not-accented.

Initial algorithms, based primarily on part of speech, label key words with labels which directly predict one of the four classes. Special heuristics are used for proper nouns, numbers and complex nominals. In addition to these general heuristics there are a number of fine tuned heuristics for specific words such as "*not*", "*but*", "*first*" etc.

This algorithm although originally implemented elsewhere, at ATR was first used on different data (see [1]). Although some tuning was necessary, in applying it to the Radio database, the algorithm was largely unchanged.

Monaghan

Unlike the Hirschberg algorithm the Monaghan algorithm [4] has an explicit notion of prosodic phrase, and of its internal accent structure. Each phrase must contain one and only one major accent, no accents may follow it within that phrase, but secondary accents may precede it. After initial accent assignment the *Rhythm Rule* ensures a well-formed-ness condition on accents, basically disallowing two accents to be on adjoining words. In addition to these general conditions there are a number of specific heuristics for special words, although not exactly the same as the special conditions in the Hirschberg algorithm their similarity is too obvious to ignore. Again they cover special cases of individual words such as "*not*" and "*but*".

Decision Trees

The third method used in the comparison is decision trees automatically learned from word feature vectors. The actual technique used was the Quinlan C4 (an extension of the standard classification

A. W.
(ATR)

and regression trees (CART) [2]), with 10% withheld cyclicly for cross-validation.

Three trees were used in the test based on varying numbers of features.

The simplest case `dtree1` predicts accent by a tree learned from four features: part of speech, previous part of speech, previous previous part of speech and boundary type after current word. The learned tree itself has only one condition, which predicts accents on content words and no accents on function words. This is a naive heuristic, but gives surprisingly good results. `dtree2` uses one more feature, the following word's part of speech (effectively offering a window of 4 part of speech tags plus the boundary type after the current word). The tree is much more complex but some heuristics are still obvious from it. `dtree3` adds a further feature looking ahead one more word. The learned tree here is much simpler than `dtree2`.

Although it may be useful to increase the number of features used in the decision trees, the learning software does have limitations (both in implementation and in theory). As in the case of the Hirschberg and Monaghan algorithms special rules were used for particular words. Unfortunately adding the word forms as features themselves, did not help in accuracy of the trees as there are too many words and not enough instances of the "interesting" examples to allow learning to occur. Even with a larger data set it is unclear if it would be able to learn these special cases without some form of semantic classification.

Comparison of Results

The results presented here are direct numerical values. It should be noted that not all errors are the same, some errors are actually worse when perceived by native listeners though that has not been taken into account in the measurements.

The data consisted of 8451 words, 4371 labelled as prominent, 4080 as non-prominent. The high degree of prominent words is due to the speaking style used in the database. The style is news announcer speech which is more accented and more stylized than normal speech, but this should make it easier to model.

The following table shows over-prediction (prediction of prominence when not prominent) and under-prediction (prediction of non-prominence when prominent) and the overall percentage of words that were predicted correctly.

Strategy	Over		Under		all
	prediction	%	prediction	%	%
Hirschberg	991/4080	24	890/4371	20	22
Monaghan	684/4080	16	1561/4371	35	26
dtree1	1503/4080	36	294/4371	6	21
dtree2	1159/4080	28	449/4371	10	19
dtree3	1295/4080	31	397/4371	9	20

Discussion

Although `dtree1` (simply accenting all content words) has a low overall error rate, it massively over-predicts. In fact all three decision trees over-predict to a greater degree than under-predict.

A slight change to the Monaghan algorithm (in the interpretation of the Rhythm Rule), not mentioned in the thesis, improved its results to close to the Hirschberg algorithm. Its tendency to under-predict may be due to the style of the data which is more accented than normal spoken English.

There is the issue of how bad the errors are to the human listener. The decision trees make isolated predictions based on static context, ignoring neighbouring predictions. This often causes unacceptable accents on adjacent words (and long periods with no accents). The Hirschberg algorithm and more so the Monaghan algorithm specifically avoid too many accents on adjacent words (as in complex nominals), thus giving a more even distribution of over- and under-prediction. Hence, in limited perception tests they were found to sound better than the apparently better scoring decision trees. Also, the "mismatches" made by the Hirschberg algorithm (and Monaghan) are often "acceptable" though different from what the speaker in the database chose.

From this work we can draw the following conclusions. First, much care must be taken that score functions are actually scoring the desirable properties. Although crude measurements may be easy to specify they should not be used for fine tuning. Second, the decision tree method actually fails to build an appropriate model because it does not take into neighbouring predictions. A more complex interpretation has yet to be tried.

References

[1] A. W. Black and P. Taylor. Assigning intonation elements and prosodic phrasing for English speech synthesis from high level linguistic input. In *ICSLP94*, Vol 2, pp 715–718, Yokohama, 1994.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* newblock Wadsworth & Brooks, Pacific Grove, CA., 1984.

[3] J. Hirschberg. Using discourse content to guide pitch accent decisions in synthetic speech. In G. Bailly and C. Benoit, editors, *Talking Machines*, pages 367–376. North-Holland, 1992.

[4] A. Monaghan. *Intonation in a text-to-speech conversion system.* PhD thesis, University of Edinburgh, 1991.