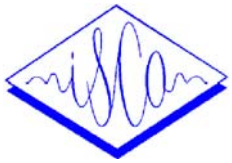

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

CART-BASED DURATION MODELING USING A NOVEL METHOD OF EXTRACTING PROSODIC FEATURES

Paul Deans, Andrew Breen and Peter Jackson
BT labs, Martlesham Heath,
Ipswich, Suffolk, England.
{paul.b.deans, andrew.breen, peter.2.jackson}@bt.com
http://www.labs.bt.com

ABSTRACT

The prediction of accurate segmental durations remains a difficult problem when synthesising speech from text. Inaccurate durations are often perceptually prominent and detract from the naturalness of the quality of speech.

For a concatenative system, a statistical approach is an excellent way of predicting segmental durations. More specifically a CART (Classification And Regression Tree) method is appropriate [1], but only if it has been correctly trained with data that reflects a phoneme's characteristics. A *feature-set* is used to describe the flavour of a phoneme in the process of building of CART trees.

We describe a novel method where BT's Laureate Text-to-Speech system (TTS) is used to automatically donate the prosodic information required to make up the feature-set, ultimately being used as training data for building a CART tree. This tree, in turn, is used to predict segmental durations.

The extraction of *salience* (derived from a metrical analysis of the text) and the other prosodic and segmental features in this way, is a novel concept. CART trees consistently show that this salience feature, in particular, has a large effect on the duration of a phoneme.

The paper describes in detail this concept and shows the importance of salience.

An evaluation of the effectiveness of CART-based duration modelling against the rule-based Laureate TTS method is given in the results.

Keywords = Synthesis, Text-To-Speech, Duration

1. THE FEATURE SET

Natural speech clearly has a duration. It exists as an acoustic signal for a finite length of time. The rate at which it is produced may vary from slow to fast. When a speaker wishes to emphasise a given word or phrase, changing the duration of the speech signal is one of the techniques they employ. Natural speech has a particular rhythm to which humans appear very sensitive. Clearly, synthetic speech, if it is to sound natural, must accurately model the duration of an utterance.
It is hoped that by using a CART tree to predict these natural durations that the speech synthesis will be improved.
One phoneme may differ from another in terms of its characteristics. These characteristics or features need to be chosen carefully as they and their values are required to describe the phoneme and its prosodic properties, context and perceptual prominence, all of which may affect its segmental duration.
These features need to be assessed to see if they warrant inclusion. The first of these is Salience.

1.1 Salience

The Salience of each phoneme is an obvious candidate for inclusion within the feature set due to the strong link between word stress and segmental duration.
Although the general aim is to obtain accurate segmental durations, the resolution of stress marking in the BT Laureate system only offers stress fields at word level and syllable level (lexical – stressed or unstressed). Each phoneme is deemed to have the same stress as its parent 'the syllable' in which it belongs.
Metrical trees [2] are used to assign the various prominence of words and syllables within a phrase. A grid is created that describes the phrase focus and it is from this grid that values for word and

syllable salience and stress respectively are determined.

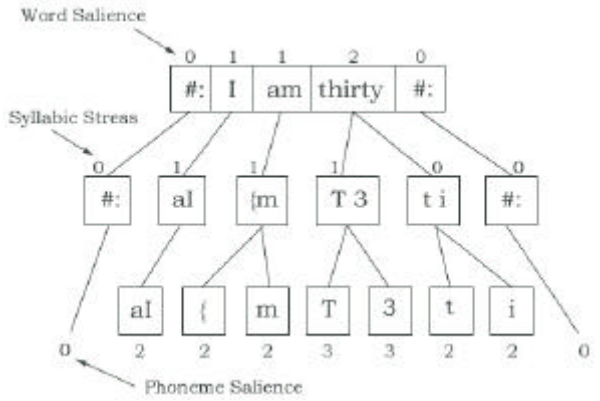

Figure 1. *Application of salience and stress*

Since this is the way that Laureate assigns salience and stress, it was decided to combine these two elements to give an overall phoneme salience.
This value of salience is achieved by assigning the salience of each word to each phoneme and adding one unit of salience to it for stressed syllables.
Figure 1 shows the values at each level.

1.2 Category, Manner and Place

Since the Laureate TTS system labels each phoneme in its textual annotation file, it is possible to reference each phoneme by way of three 'categorical' associations.
Category describes, by way of an associated number, whether a phoneme is a consonant, affricate, short vowel, long vowel or diphthong. It is felt that the 'Category' of a phoneme is strongly related to its duration.
Place is the location in the vocal tract where the articulators form a constriction, and manner describes the manner of this closure. A table of manner and place, derived from a table in Ladefoged[3] was employed and converted to numbered categorical variables which the CART interprets.

1.3 Pitch Accent Degree.

It can be seen that Pitch Accent Degree or prominence, a variable linked with pitch accent changes, is closely connected with salience and therefore with duration. That is to say, under certain circumstances a change in pitch can give rise to a change in duration. Whereas this may not always be the case, the argument is strong enough to include Pitch Accent Degree in the feature set.

1.4 Number of Accented Syllables

It is suggested that there is a correlation between the number of syllables in a sentence and speaking rate. The feature is described as Number of Accented Syllables and for the reasons given, merits a place in the feature set.

1.5 Offset from end of sentence and Offset from most prominent syllable

It is often the case in phrases that the amount of stress placed on a word increases towards the end of the sentence. A normalised figure representing the offset from the end of sentence is therefore included in the feature set.
Also, a speaker may speed up towards a prominent syllable or word and then slow down afterwards. This can be represented by a normalised figure between +1 and –1, as a phoneme may be described as preceding or following this prominent syllable.

2. EARLY EXPERIMENTS

Early experiments showed that the salience feature was considered highly significant by the CART tree. The *category* or *type* of a phoneme was the feature considered the most influential on phonemic duration by the CART tree though salience often came second.
The data used to train the CART tree in the early experiments comprised of phonetically rich, similar length sentences. The experiments showed that the CART tree successfully corrected known problem words and word groups. However, it fell short when predicting the duration of phonemes resident in smaller phrases, where it was trying to predict the segmental durations of unfamiliar phonemes. The CART had been trained using sentences ranging from between seven and thirty words in length. The training data was therefore felt to be insufficient in terms of small phrase information. This manifested in the unnatural effect of 'rushing' smaller phrases. A new corpus was required that would hopefully eradicate this problem.

2.1 New Corpus Design

The experimental CART trees showed weakness in terms of their ability to accurately predict the segmental durations in short phrases. In

comparison, the method currently employed by Laureate, uses a rules-based method of multiplying the durations by an amount directly related to the number of words in the major phrase [4]. Generally speaking, the fewer words in the major phrase, the bigger the multiplying factor.
It seems clear that the training data should consist of phrases that will 'teach' the CART about phonemes from various length sentences. It is also felt that the more sentences used, the better the CART would be at prediction. The early tests were executed with a CART trained using 250 phrases. The new CART will be trained with 2000 phrases. This constitutes to about 88000 phonemes.
It is vital that a suitable proportion of the sentences are short. The change in duration of phonemes is more noticeable in smaller phrases than large phrases. Thus, the phrases should be between 1 and 30 words in length with the majority at the lower end.
Consider what we want the TTS to deal with in terms of sentence length. If the CART is trained with short stories, It follows that the TTS system would be more capable of relaying similarly structured stories. If we want the TTS to read E-mail, we should train it with E-mail. If, however, we need a general text reader, then we should find many, different, corpora to train the CART. This is exactly what is required.
The style of the recorded speech is also of great importance. Assuming the style might filter through the system, it was felt that a relaxed newsreader style would be desirable. This would be clear, friendly without over-articulation, though more formal than a conversational style.
The corpus was recorded and the text and speech files used to obtain the training data.

3. METHOD

To build a CART tree that predicts segmental durations requires training data that represents a given phoneme in terms of its features plus the appropriate duration for that phoneme.
A novel way of obtaining these features is by using Laureate to donate them. Laureate calculates the prosodic information in order to produce speech output from text. This prosodic information is then taken and used to make training data.
The segmental duration, which will ultimately be the dependant variable for the training data, is extracted from the time-aligned, annotated speech files that Laureate normally employs.
These annotation files hold the duration pertinent to each phoneme.
The prosodic information, donated by Laureate, relates to each syllable, so an alignment process occurs that joins the syllable data from Laureate to the phoneme data from the annotation files.
The process of feature donation can be seen in Figure 2.

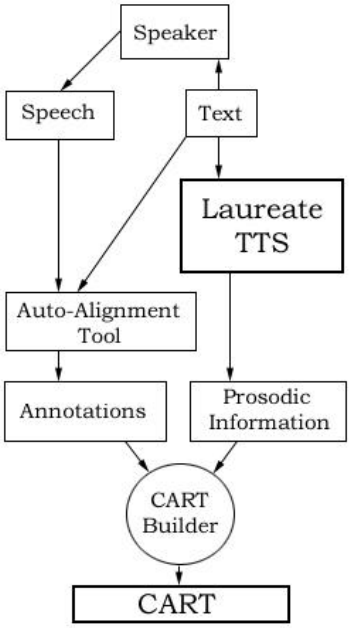

Figure 2*. Prosodic information donation.*

In summary, each phrase from a script is individually processed by Laureate and the prosodic information collected in a file. The durations from that phrases' annotation file are then added to make the total training data for that phrase.
This occurs for each phrase until a large training data file has been created.

Once created, the file is used to create a segmental duration-predicting CART which can then be used to predict durations dynamically. That is, when Laureate reads in text and outputs speech, it relies upon the CART to predict the segmental durations for each phoneme.

3. INFORMAL EVALUATION

A full comparison of the system using each method is a lengthy process as there are many variables involved. However, it is possible to reach a conclusion fairly quickly regarding Laureates behaviour under each duration method.

As mentioned previously, the human ear is sensitive to unnatural durations and it is this human quality that allows us to draw a swift conclusion after a listening to just a few sentences. To aid the informal evaluation, a small subjective test was composed.

3.1 Subjective Evaluation

Twenty phrases of various lengths were used in this informal evaluation. TTS was used to create these twenty phrases by using the rules-based method and twenty using the CART method.
The listener was offered ten phrases, randomly selected from the twenty phrases. They were offered two phrases at a time, one from each system. Both the rules and CART phrases were played first an equal amount of times. The listeners were asked their preference after each pair.
Also, the listener was never made aware of which system they were listening to.
The CART-based method was expected to lengthen segmental durations in smaller phrases in a way that was more naturally required, rather than the almost indiscriminate multiplying that occurs in the rule-based method.

4. RESULTS

Only a few subjects were used for this test but the results show that only 20% of the phrases created using the CART method were preferred over the rule-based method. This, although disappointing, was understandable as careful listening exposed many of the CART phrases as less natural sounding.
Also, the longer sentences seemed to fair better than smaller phrases. It was also observed that some words seemed inappropriately long and some inappropriately short. Short phrases, whose durations from early CART work were perceptually fast, still seemed unnaturally rapid.
Also, some signal processing error could clearly be heard in the speech units of some phrases.

The Salience feature in previous CART trees was deemed to have a high impact on the segmental duration. However, salience was *not* considered as important by the new CART. It was placed about halfway up the tree. Instead, Offset from end of Sentence became the most important feature.

5. CONCLUSION

The performance of the CART-based method is currently inferior to that of the rule-based method when assessed by subjective tests.
The CART may predict a suitable segmental duration for a given phoneme, with its particular characteristics. However, if the intrinsic duration of the chosen speech unit is substantially different from the predicted duration, a problem arises in terms of signal processing artefacts.
Close inspection of the speech waveform in
Figure 3, clearly shows this signal-processing problem.

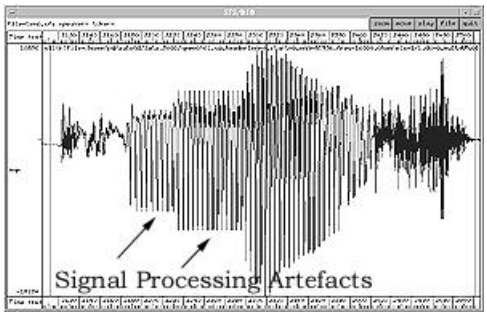

Figure 3*. Signal-processing error*

The rule-based model takes this into account and assigns a duration that does not require extensive signal lengthening.
One explanation for the generally poor results could be because the data was automatically produced instead of being hand annotated.

The rule-based method works well and though there is room for improvement in terms of naturalness, it is still a mature method that copes well with a variety of text.

6. REFERENCES

[1] Riley, M. (1992), Tree-based modelling of segmental durations. *Talking Machines. Theories, Models and designs. North-Holland.*

[2] Edgington, M; Lowry, A; Jackson, P; Breen, A. P and Minnis, S. Overview of Current text-to-speech techniques: Part II prosody and speech generation. *BT Technol J Vol 14 No 1 January 1996.*

[3] Ladefoged, P. (1982). A Course in Phonetics, *Harcourt Brace Jovanovich*. 1982.

[4] Breen, A P. (1995). A simple method for predicting the duration of syllables. *Proc. Eurospeech '95, Madrid, pp 595-598.*