Proceedings of the Institute of Acoustics

A METHOD OF ESTIMATING SEGMENTAL DURATIONS

Dr. A. P. Breen (1), Dr. M. G. Easton(2)

(1) B.T. Laboratories, Martlesham Heath
(2) Liverpool University, Dept. of Electronics and Electrical Engineering

1. INTRODUCTION

The rhythmic structure of an utterance is an important prosodic attribute, Synthetic speech produced with inappropriate rhythm can sound unnatural and laboured. In text to speech synthesis, the rhythm of speech has typically been modelled by modifying the durations of the constituent phonetic segments. The amount by which a given segment is modified is dictated by a predictor. There are two approaches commonly adopted in the prediction of segmental durations, non-parametric predictors and parametric predictors. Non-parametric predictors normally take the form of statistical methods, which are effectively brute force approaches and make very few assumptions about the underlying causes that lead to duration changes[1][2]. In contrast parametric predictors attempt to model the underlying causes that lead to duration changes[3][4]. Both approaches use contextual information in the generation of a prediction. Statistical decision trees, for example, use contextual information at decision nodes to determine which path should be taken at any given point, while parametric models use contextual information to from duration modification rules.

A new method of determining optimal predictor coefficients for a parametric duration predictor is presented. The method contains a multiplicative predictor model whose coefficients are automatically determined from a large corpora of annotated speech data. Predictor context rules are generated in a text file using a rule generation language. These rules are then compiled and "optimal" coefficients determined from the data. This method of rule generation allows a number of different context rules to be quickly assessed.

2. OVERVIEW

The aim of the work described in this paper was to develop a technique which could automatically determine an "optimal" set of predictor coefficients for a multiplicative duration model given a set of context sensitive rules and a database of annotated speech. This process is presented pictorially in figure 1. Each of the process boxes (Those shown without shading in figure 1) will be described in detail below.

3. CONTEXT SENSITIVE RULE GENERATION

Before a set of predictor coefficients can be determined, a rule file is generated which contains the context rules used to control the duration predictor. Each rule specifies an environment of interest, which is assumed to affect the duration of a segment. The rule file, generated as a text file, may contain up to forty separate rules, the syntax of which is as follows:

>> $R_0 / R_{-1}\, R_{-2} \ldots R_{-n} / R_1\, R_2 \ldots R_n / B /$
$R_0 / R_{-1}\, R_{-2} \ldots R_{-n} / R_1\, R_2 \ldots R_n / B /$

Where R_0 is a sub-rule describing the phoneme, $R_{-1}\, R_{-2} \ldots R_{-n}$ are sub-rules describing the previous phonemes (the most recent phoneme is listed first), $R_1\, R_2 \ldots R_n$ are sub-rules describing the following phonemes and B is a Boolean expression. 'n' is a context window, which specifies the number of phonemes over which the rule context

extends. The sub-rules describing the preceding phonemes, the following phonemes and the Boolean expression may be omitted if desired. Only the sub-rule R_0 must be specified. The start of a new rule is marked using the double chevrons. Multiple line rules can be constructed by omitting these double chevrons and are evaluated by 'ORing' the outcome of the results of each line. Comment lines may be inserted into the rules file by preceding the rule with a '!' symbol. Each sub-rule description R is of the form:

SPt
Where

S =	'	primary stress
	"	secondary stress
	*	Either primary or secondary stress
P =	V	any vowel
	C	any consonant
	D	any diphthong
	F	any Fricative
	A	any affricate
	b	any plosive
	N	any nasal
	S	any syllabic
	P	any phoneme
	#	end of phrase marker
[aI,z,h,tS,...,V]		any list of phonemes

t =	$	end of syllable
	_	end of word

Possible examples for sub-rules R include:

*V	any stressed vowel
PS	last phoneme of a syllable
C_	any word final consonant
"[aI,t]	Secondary stressed aI or t at the end of a word

The Boolean expression B is of the form

^b1*^b2*^b3*...*^bi

Where ^ is a NOT operator and is optional,

* =	+	OR operator
	.	AND operator
b =	iw / fw	initial / final word of syllable
	is / fs	initial / final syllable of word
	ip / fp	initial / final phoneme of syllable

Some examples of possible rules are given below:

>> 'P / / / fs	any primary stressed phoneme on the final syllable of a word
>> C/C/C//	any consonant both preceded and followed by a consonant
>> *V$ / / / is.fs /	any syllable final stressed vowel in a monosyllabic word
>> C/ / V / /	
C / / C V / /	the two consonants following a vowel

The rule file with the annotation files are compiled into a set of truth data files which describe, for each phoneme in an annotation file, the number of context rules which are fired by the phoneme's environment. The format for each line of the truth data files is as follows:
P D t1,t2,..tm
where P is the phoneme name, D is its duration in seconds, $t_1,t_2,...,t_n$ are a list of the truths corresponding to each rule in the rules file. t=1 corresponds to true and t=0 corresponds to false.

4. PRE-PROCESSING THE ANNOTATION DATA

The database of speech annotations used in this work consist of files containing phonemically labelled segments. However the context rule compiler requires context information about a given phoneme, which is not easily accessible in such phonemic annotations strategies. Because of this the annotation files are pre-processed before entering the rule compiler. Each annotation is thus reformatted. Each line of the reformatted file contains all the information necessary to describe the context of every phoneme and is formatted as follows:
$D\ P_0\ P_{-1}\ P_{-2}\ ...\ P_{-n}\ P_1\ P_2\ ..\ P_n\ a\ b\ c\ d\ e$

where D is the duration of the phoneme in seconds.

P_0 is the description of the phoneme including stress and termination information.
$P_{-1}\ P_{-2}\ ..\ P_{-n}$ are the descriptions of the previous n phonemes. 'n' is the context window.
$P_1\ P_2\ ..\ P_n$ are the descriptions of the n following phonemes.
'a' is the position of the phoneme in the syllable while 'b' is the total number of phonemes in that syllable. 'c' is the position of the syllable containing P_0 in the word while 'd' is the total number of syllables in that word.
'e'='#' if the phoneme occurs in the final word of a clause, 'e'='f' if the phoneme occurs in the first word of a clause otherwise 'e'=' '.

5. CALCULATION OF PREDICTOR COEFFICIENTS

The following algorithm has been devised for predicting the lengths of phonetic segments in speech. The algorithm modifies the durations of particular phonemes based on a number of context sensitive rules described above. These rules are reflected in the algorithm as a number of coefficients which modify the duration of a given phonetic element.

Let the inherent and minimum durations of phoneme number p be given by inh_p and min_p. The duration is considered as a function of the inherent and minimum durations of the phoneme, the truths of each rule $t_{1p}, t_{2p}, ...,t_{Np}$ and the corresponding multipliers $c_1,c_2,...,c_N$ where N is the number of rules. Assume the duration to be approximated by the following predictor:

ESTIMATING SEGMENTAL DURATIONS

$$D_p = D(p, c_1, c_2, \ldots, c_N, t_{1p}, t_{2p}, \ldots, t_{Np}) = (inh_p - \min_p)\prod_{i=1} c_i + \min_p \qquad 1.1$$

where the product is calculated over all c_i when t_i=1. this product is an approximation to the actual duration modifier d given in equation (1.4).

Converting to logs

$$\ln((D_p - \min_p) / (inh_p - \min_p)) = \sum_{i=1}^{N} t_i \ln(c_i) \qquad 1.2$$

or

$$\ln d_p = t_p k \qquad 1.3$$

where the duration modifier $d_p = (D_p - \min_p)(inh_p - \min_p)$ 1.4

k is a column vector (of length N) of the logged prediction coefficients and t is a row vector (of length N) of the truths of the N rules corresponding to phoneme p. For example, if phoneme p satisfied rules 3 and 7, we would have

$$\ln d_p = k_3 + k_7$$

Let there be j possible rule combinations where $j<=2^N$. Let d'_i be the mean of all d_p corresponding to the ith possible rule combination. We now have a set of j simultaneous equations to solve in N unknowns.

$$d = kT \qquad 1.5$$

where d is a vector of logged mean duration modifiers; $d=[\ln d'_1, \ln d'_2, \ldots, \ln d'_j]^T$, and T is a (j x N) matrix of each truth combination; $T=[t1, t2, \ldots, Tj]^T$.

We wish to solve for k to minimise the mean squared error ε given by,

$$\begin{aligned}\varepsilon &= (kT - d)^T(kT - d) \\ &= k^T T^T Tk - 2Td^T k + d^T d \\ &= k^T Ak - 2bk + d^T d \end{aligned} \qquad 1.6$$

where

$$A = T^T T \text{ and } b = Td^T \qquad 1.7$$

To minimise, set $d\varepsilon/dk$ to zero, $Ak_{opt} = b$ 1.8

This set of N simultaneous equations in N unknowns can be solved for k. Then the prediction coefficients 'c' will be given by,

$$c_i = \exp(k_i)$$

The normalised minimal mean squared error is then,

$$e_{min} = 1 - (bk_{opt} / d^T d) \quad 1.9$$

The NAG utility F04ARF for solving simultaneous equations was used to solve the above.

6. PRELIMINARY RESULTS

This section demonstrates some initial investigations undertaken using this technique. Two simple examples were considered and one more complex example.

6.1. Investigation into the effect of shortening in clusters

The coefficients were optimised over a database of approximately 200 short sentences with the following simple rule specified:

! Rule 1. Consonant preceded by a consonant.
>> C / C / /
! Rule 2. Consonant followed by a consonant.
>> C / / C / /
! Rule 3. Consonant preceded and followed by a consonant.
>> C / C /C / /

The solved coefficients were as follows:

c1 = 0.989
c2 = 0.907
c3 = 0.787
with a normalised MSE=0.0115

Truths	No. of occurrences	d' mean duration	Prediction
010	662	0.907	0.907
100	662	0.989	0.989
111	54	0.705	0.705
000	8048	1.04	1.00

Table 1.

Table 1 shows the truth combinations the number of occurrences of each context in the database the mean duration modifier d' as defined in the previous section and the product of the corresponding prediction coefficients. The model is minimising the error between the last two columns of the table.

From the values of the coefficients there is evidence of shortening in clusters.

6.2. Investigation into the effects of stress on phoneme duration

As in the previous example the predictor coefficients were optimised using a database of approximately 200 short sentences. The context rules were as follows.

!Rule 1. Primary stressed vowel
>> 'V / / / /
!Rule 2. Primary stressed consonant
>> 'C / / / /
!Rule 3. Primary stressed diphthong
>> 'D / / / /
!Rule 4. Primary stressed plosive
>> 'b / / / /
!Rule 5. Primary stressed fricative or affricate

These somewhat arbitrary rules are were designed to demonstrate the over multiplicative nature of the predictor. It is possible to define rule sets which state requirements in both very broad and very specific terms. i.e. rule 1 and rule 3.

The predicted coefficients were as follows
c1 = 1.151
c2 = 1.177
c3 = 1.030
c4 = 1.002
c5 = 0.979

with a normalised MSE = 0.0158

Table 2 shows how give a break down of information for each rule

Truths	No. of occurrences	d' mean duration	Prediction
01000	618	1.177	1.177
10000	726	1.151	1.151
01010	678	1.179	1.179
00000	6366	0.957	1.000
01001	653	1.153	1.153
10100	385	1.186	1.186

Table 2.

As expected there is clearly evidence of phonemes lengthening when in stressed environments

6.3. Investigation of Klatt type context rules

The examples given above were all very simple. As a more interesting test a set of rules similar to those proposed by Dennis Klatt [3] were generated. The rule set is given below:

! Rule 1. Clause final lengthening.
>> V / / / fs.fw /
C / V / / fs.fw /
C / C V / / fs.fw /
!Rule 2. Non-phrase final shortening.

```
>> P / / / ^fs+^fw /
!Rule 3. Non word final shortening
>> P / / /^fs /
!Rule 5. Polysyllabic shortening
>> P / / / ^is+fs /
!Rule 6. Non initial consonant shortening
>> C / / / ^ip+^is /
!Rule 7. Lengthening for emphasis
>> 'V / / / /
!Rules 9. - 18 Post-vocalic context of vowels
>> V_ / / / /
>> V / / [v,D,z,Z] / /
>> V / / [b,d,k] / /
>> V / / [m,n,N] / /
>> V / / [p,t,k] / /
>> V_ / / fs,fw /
>> V / / [v,D,z,Z] / fw.fs /
>> V / / [b,d,k] / fw.fs /
>> V / / [m,n,N] / fw.fs /
>> V / / [p,t,k] / fw.fs /
!Rules 19 - 23. Shortening in clusters
>> V / / V / /
>> V / V / /
>> C / C / / /
>> C / / C / /
>> C / C / C / /
!rule 24. Lengthening due to plosive aspiration
>> *V / [p,t,k] / / /
```

The coefficients calculated using the rules given above seem to be producing reasonable results as exemplified by the coefficient for the clause final lengthening rule. The value for this coefficient was calculated as c1=1.899 which is similar to the value of approximately 1.4 calculated by Klatt. However, perhaps surprisingly there is little evidence of polysyllabic shortening as the coefficient value of c4=0.988 is very close to unity.

The overall mean magnitude error using the Klatt rules was found to be 20.6ms which does not compare too favourably to the mean magnitude error of 22.4ms using no prediction at all. The model has only improved the mean error by 2ms. However, this effect may be due to the predictive power of the model being swamped by the noise inherent in the annotation data.

The mean magnitude error calculated using the exact duration modifiers was found to be 18.8ms, roughly a further 2ms improvement. This indicates that a small increase in performance is possible by not using the coefficient model but retaining the actual rules. Even so most of the prediction error is due to the rules not being able to closely model the real data.

7. CONCLUSIONS

The work above suggests that the problem with predicting accurate segmental durations is not so much in the type of predictor models used, but in the assumptions made by the context sensitive rules. Such prediction methods are predicated on the fact that the rules adequately describe the major contributing factors effecting segmental durations. If this is not the case then it is hardly surprising to find that the models used to predict durations do not perform well.

8. REFERENCES

1. Breen, A.P., "A comparison of Statistical and Rule Based Methods of Determining segmental Durations", Proceedings of International Conference on Spoken Language Processing, Banff, pp. 1199-1203, 1992.
2. Riley, M.D., "Tree-based Modelling of Segmental Durations", *Talking Machines: Theories, Models, and Designs*, Ed: G. Bailly, C. Benoit, Norht-Holland, 1992.
3. Allen, J. Hunnicutt, M.S., Klatt, D., *From text to speech: The MITalk system*, Cambridge University Press. 1987.
4. van Santen, J., Olive, P., "The Analysis of Contextual Effects on Segmental durations", Computer Speech and Language, 4, pp. 359-390, 1990.

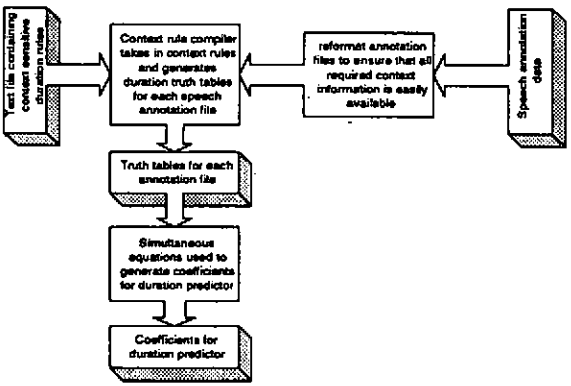

Figure 1