INTRINSIC AND ADJUSTED UNIT LENGTH IN ENGLISH RHYTHM SYNTHESIS

Mark Tatham University of Essex, Dept. Language and Linguistics
Katherine Morton University of Essex, Dept. Language and Linguistics

1. INTRODUCTION

In the design of high-level synthesisers (Tatham *et al*. 2000) nothing is more pressing than accurate modelling of the various parameters of prosody. The parameters we will consider are: sentence stress, rhythm and intonation. Improvements to prosody modelling is urgent for two reasons:

1. prosodic, or 'suprasegmental' quality is a major contributor to judgements of naturalness in the output speech – a poor judgement here negates the achievement of reasonable accuracy in rendering the 'segmental' quality of speech units such as syllables or words;

2. models of prosody are not very good at the moment – they regularly produce undesirable adverse judgements from listeners.

It follows from these observations that more research is called for in how prosody might be more successfully modelled for the purposes of spoken language engineering, in both synthesis and recognition, for it is becoming equally clear that accuracy of automatic speech recognition devices could be also improved in this area.

In several papers we have addressed the question of intonation and discussed how we are currently modelling this in the SPRUCE high-level synthesis system (Tatham *et al*. 2001). We believe we now have a reasonable working model of intonation and want to turn our attention to the other parameters. In this paper we consider rhythm.

2. RHYTHM – PRELIMINARIES

It is accepted that speech exhibits rhythm – that is, the patterned temporal occurrence of events. But precisely what physical events contribute to the acceptance of rhythm as a feature of speech is curiously elusive. This has led to the widely held view that rhythm is a perceived effect which may or may not have reliable acoustic correlates (Hay and Diehl 1999, among others).

Pike (1945) and Abercrombie (1967) are often cited as putting forward the idea that some languages, like English, are 'stress timed' and that others, like French, are 'syllable timed'. 'As far as is known, every language in the world is spoken with one kind of rhythm or with the other' Abercrombie (1967, p.97). The suggestion here is that the units of rhythm for English are the time intervals from stressed syllable to stressed syllable, but in French they are just the time intervals from syllable to syllable. English has both stressed and unstressed syllables, whereas French has only stressed syllables, it is claimed. These researchers and many others had observed that for

whichever type of language the temporal pattern that raises the status of rhythm to syllable timing in speech is *isochrony* – that is, the equal timing of these units. So, in English for example, there would be equal timing from the start of a stressed syllable to the start of another stressed syllable, whereas in French there would be equal timing from the start of any syllable to the start of the next syllable. These and other researchers accept that due to stylistic and other effects such as hesitation phenomena the isochrony may not be perfect.

Dauer (1987 and in other papers) points out that the distinction between different types of language is not bimodal but scalar. English and French may be fairly near the extremes of this scale, but languages like Iberian Spanish and Catalan are consistent in falling at a point between the two extremes, and some others seem sometimes to be at one point of the scale and at other times on another point. We ourselves have observed that the French of Montréal, for example, has isochronic units closer to those of English than to those of Metropolitan French (Tatham and Morton, *in preparation*), and that style changes within one speaker shift the isochrony along the scale.

3. ISOCHRONY AS A PERCEPTUAL PHENOMENON

Observations of isochrony as a dominant feature of speech rhythm turn out to be a matter of perceptual reality rather than of physical fact (Donovan and Darwin 1979). Many researchers have investigated the acoustic signal of a number of languages in the hope of finding some measurable parameter which might be responsible for triggering the perception of regular rhythm. An early, but definitive study is that of Lehiste (1977), who concluded that the effect is a perceptual phenomenon, with listeners latching onto stressed syllables, which may carry a higher semantic load (Buxton 1983). Benguerel and d'Arcy (1986) discouraged measuring the acoustic signal to look for perceived regularity, despite the fact that such work has been carried out in a well-established paradigm which seeks to identify the acoustic correlates of perceptual units in general.

Some researchers have had recourse to various transformations of the data to try to come up with an isochrony model at the acoustic surface. Thus we find early work of Hill, Jassem and Witten (1978) trying to find an index based on some intrinsic period in rhythmic unit repetition. And again, Jassem, Hill and Witten (1984) use an elaborate statistical technique in their quest for finding hidden isochrony in the acoustic signal. Williams and Hiller (1994) tried delimiting the rhythmic unit in different ways. For example, the stressed syllable is usually taken as being the first syllable in the unit, but perhaps the stressed syllable should fall somewhere else in the unit, say, at the end. Williams and Hiller were painstaking and exhaustive – their statistical analysis revealed a very slight, significant tendency towards isochrony in the measurements.

4. THE 'ACOUSTIC CORRELATES' PARADIGM

Investigating the relationship between acoustic measurements and cognitive phenomena has a long tradition in experimental speech studies. One aspect of the approach attempts to discover regularity in the phonetic rendering of some underlying phonological plan: acoustic measurements are related to abstract phonological representations. In addition researchers have sought to discover acoustic events which give rise to predictable perceptual responses: correlates are again identified between physical events and abstract representations. What is there to say about a correlation between isochrony and the acoustic waveform?

It seems to us that speakers might even be aware of isochrony in their *own* speech, just as listeners report the perception of isochrony in the speech of others. If this is true it could follow that it is planned. An alternative explanation, though one less attractive to us, is that isochrony is a necessary physical property of speaking – of which speakers and listeners alike are aware. Our

reason for saying this is that there many such properties of speech, like coarticulation (see the collection of studies in Hardcastle and Hewlett 1999), of which speakers and listeners are generally *not* aware.

For the purposes of designing synthesis systems we want to include acoustic effects which would trigger perceived isochrony in the listener, for, if isochrony is an expected feature of human speech, the results will sound unnatural if the feeling of isochrony is lost. But since most of the literature does not report isochrony in the acoustic signal it would seem that we need to synthesise a rhythm which is not itself isochronic, but which gives rise to the perception of isochrony. This is a tall order since the literature is about looking for isochrony, failing (in general) to discover it, and then trying to manipulate the data in various ways to discover a hidden concomitant rendering of equal timing. We feel that we need to discover what effects there are which, without such elaborate statistical processing clearly not at the disposal of the listener, might trigger the perceptual effect. Indeed this philosophy is behind much of our work: how do we generate an acoustic signal to cause appropriate responses in the listener? – not: how do we generate the 'right' signal?

It is tempting to synthesise an acoustic signal which actually does have isochrony in the hope that this will do. But however attractive the idea, we have shown in an unreported pilot investigation that it does not work. It seems that such a scenario taxes too heavily the human ability to adapt to an unusual signal and perhaps perceptually adjust it to something more normal. Clearly how and why this is the case needs proper investigation, and constitutes a research topic which we are pursuing. The research might prove valuable because it points toward an understanding of the limits of temporal adjustment on the part of the listener. A corresponding strategy in non-prosodic aspects of synthesis might involve building speech as a conjoined string of idealised segments devoid of coarticulatory effects – this doesn't work either (Peterson and Shoup 1966; see also Patel *et al.* 1999 and Kelso 1995).

5. THE EXPERIMENTAL INVESTIGATION - ISOCHRONY

To construct a rhythm model for our speech synthesis we needed yet more data. Although many researchers had already investigated the problem and reported the results of, for example, statistical treatments of their data, we needed to have our own raw data to perform a range of analyses designed to throw light on a number of hypotheses – as well as drawing on the experience of others and *their* analyses. Our data would consist of read speech. We avoided the extremes of

a. short sentences or unnatural utterances within frames – these would tend to develop a rhythm of their own which might well approximate to isochronic repetition of stressed syllables, and

b. ordinary conversation – too many false starts and other pause or interruptive effects.

Read speech seemed a suitable compromise – to be widened later if results proved promising. But in addition our speech synthesis system is called upon more frequently to speak in a read speech manner (for example, in reciting retrieved information from a database) than in short sentences or in a conversational mode.

6. WORKING DEFINITIONS, DATA ASSEMBLY, HYPOTHESES

6.1 Definitions

- *Rhythm* is the patterned temporal occurrence of pre-defined rhythmic units.

- *A rhythmic unit* is the temporal interval from the start of a stressed syllable to the start of the next stressed syllable: that is, a rhythmic unit always begins with a stressed syllable (see Jassem 1952 for the use of the term).

- A *syllable* is a phonological unit which forms the basis of the prosodic parameters of rhythm, stress and intonation – it is defined in terms of its hierarchically organised structure based on its segmental (consonantal and vocalic) composition. Syllables must have one vowel as their nucleus with margins where, in English, from zero to three consonants precede the nucleus and from zero to four consonants follow the nucleus: $C_0^3VC_0^4$ (Gimson 1962; see also van der Hulst and Ritter 1999 for a collection of much wider discussions on the nature and structure of syllables).

- A *stressed syllable* is one which bears phonological primary stress: that is, some kind of planned prominence which can also be perceived from the acoustic signal. The prominence distinguishes it from other, less prominent syllables. There is no fixed acoustic correlate of prominence, but it may be correlated with enhanced amplitude, increased duration or abrupt change of fundamental frequency – or all three in any combination (Fry 1958).

6.2 Data assembly

One subject, a female speaker of the general accent of Southern California, read out loud the front page of *The Los Angeles Times* for 25th December 2000. This consisted of half a dozen stories in marginally different journalistic styles. The material was recorded in a quiet room, directly onto the hard disk of an IBM ThinkPad computer using the shareware signal processing software 'CoolEdit 96' (Syntrillium Software Corporation, Phoenix – *syntrillium.com*), and this software was used for all editing and subsequent analysis. The recording was made in mono mode using a sampling rate of 16kHz with 16bit amplitude resolution. The microphone used was a Sony electret microphone, ECM-909A.

Four of the articles constituted the data for analysis and one of the remaining articles the data on which to test the derived model. The database was therefore quite small, but certainly sufficient in our view to establish trends and contribute to begin modelling a speaker's production of a rhythmic structure.

6.3 Hypotheses

We formulated a number of hypotheses:

***H*$_1$** Any pattern of rhythmic units observable in the data is isochronic – expectation: we shall find *no* statistically significant isochrony.

***H*$_2$** There is no correlation between the duration of a rhythmic unit and the number of unstressed syllables it contains – expectation: there *is* a statistically significant correlation.

***H*$_3$** There is no trend for rhythmic units to increase in duration before particular syntactic boundaries – expectation: there *is* a statistically significant durational increase.

Hypothesis 1 is designed to enable us to say whether or not the data has its rhythmic units arranged isochronically. Since the majority of researchers have not been able to find direct acoustic representation of equi-timed rhythmic units we expect to reject the hypothesis.

Hypothesis 2 investigates the degree of correlation between rhythmic unit duration and the number of syllables it contains: if just one syllable it will be a stressed syllable, if more than one the initial syllable will be stressed and the remainder will be unstressed. We expect to reject the hypothesis, finding a quite strong correlation between rhythmic unit duration and the number of syllables within the unit.

Hypothesis 3 is expected to be rejected: perceptually rhythm is known to slow down towards the end of sentences – though we are looking at phrase boundaries and other pauses as well.

7. DATA MEASUREMENT AND ANALYSIS

7.1 Hypothesis 1

Using all data from the first story the duration of each rhythmic unit was measured by hand. We developed a set rules to ensure consistent measuring of the data. So, for example, rhythmic units ending in a plosive were measured up to and including the release of the final syllable's closing plosive consonant. Rhythmic units beginning with a voiceless plosive 'stole' a stop associated silent interval prior to the release, or, if the plosive was voiced were deemed to have begun at a point were the stop phase began irrespective of any carry-over vocal cord vibration from a previous vowel or continuant, etc. Fig.1 illustrates one or two of these rules – though there were many, most of them fairly standard in the measurement of acoustic speech signals. Standardisation and consistency are what matters here, and particular attention was paid to these considerations.

Fig.1a. | **steps** in | – rhythmic unit excised from *…Beron steps inside a…* . At the start of the rhythm unit [s] overlaps the preceding [n], but is taken to start where [n] vocal cord vibration stops. At the end the rule is the same: stop the [n] of *inside* where vocal cord vibration stops despite slight overlap from the following [s].

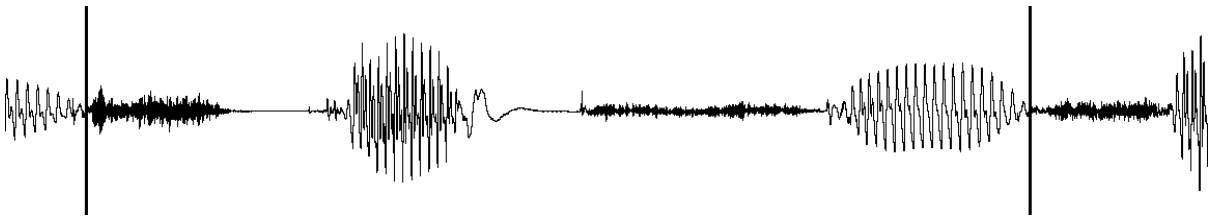
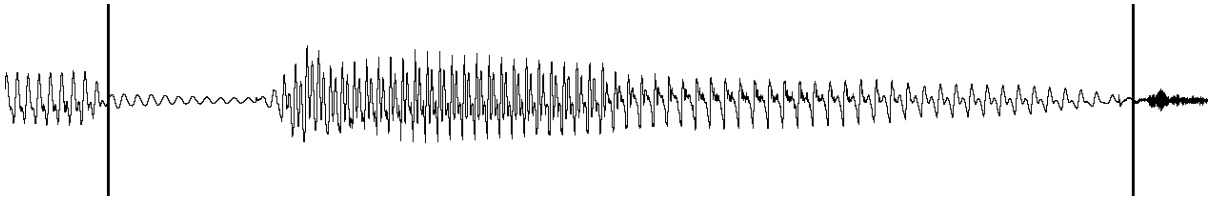


Fig.1b. | **ber**on | – rhythmic unit excised from 'Patty Beron steps…' Zoomed to show the moment of stop onset (closure) for the initial [b] of this unit. Note that this speaker carries the vocal cord vibration right through the 'voiced' [b]. The unit ends where the vocal cord vibration for the final [n] despite slight overlap with frication for the following [s].



Many boundaries exhibited end or pause effects. So, for example, at the start of a story, or following some kind of break (a pause, a phrase, sentence or paragraph boundary) there were sometimes 'hanging' rhythmic units – that is, units which did not begin with a stressed syllable. To illustrate this, examine just one sentence in Article 1, beginning

"For | ***her*** *and her* | ***friends*** | *…"*

which is a hanging unit followed by two complete units. Vertical bars are rhythmic unit boundaries, bolded syllables are the stressed ones. The rhythm unit immediately preceding these boundaries, while always complete in the sense that it always contained a stressed syllable, often exhibited an increased duration correlating with the slowing down effect before syntactic boundaries observed by, among others, Klatt (1975, 1979).

Article 1 was therefore analysed by paragraph – with all hung rhythmic units omitted from after paragraph, pause or other syntactic or stylistic pause boundaries. Two analyses were performed, one omitting all rhythmic units from before the above boundaries and one including them. The results appear in Table I a. (without pre-pause units) and b. (with pre-pause units). The corresponding sample graphs (Fig.2) show the speed reduction trend and this is shown in the tables by an increase in the mean unit durations.

Table I Durations of rhythm units in Article 1, by paragraph

	a. durations in ms without pre-pause units							**b. durations in ms with pre-pause units**						
paragraph	***mean***	***median***	***SD***	***min***	***max***	***count***	***v***	***mean***	***median***	***SD***	***min***	***max***	***count***	***v***
1	342.6	342	65.1	214	449	21	19	357.8	354	71.9	214	505	25	20.1
2	391.3	372	90.7	256	578	24	23.2	397.3	386.5	89.5	256	578	32	22.5
3	389.1	356.5	95.1	290	612	18	24.4	399.6	390	88.1	290	612	23	22
4	379.7	385	100.1	210	543	35	26.4	381	387.5	96.4	210	543	44	25.3
5	406.5	430	110.1	178	676	31	27.1	404.4	398	103.9	178	676	40	25.7
6	371.3	397	113.4	184	570	15	30.5	406.5	420	103	184	570	27	25.3
entire article	***382.3***	***368***	***97.8***	***178***	***676***	***144***	***25.6***	***391.4***	***390***	***94.3***	***178***	***676***	***191***	***24.1***

Fig. 2a. Durations for each rhythm unit in paragraph one of Article 1

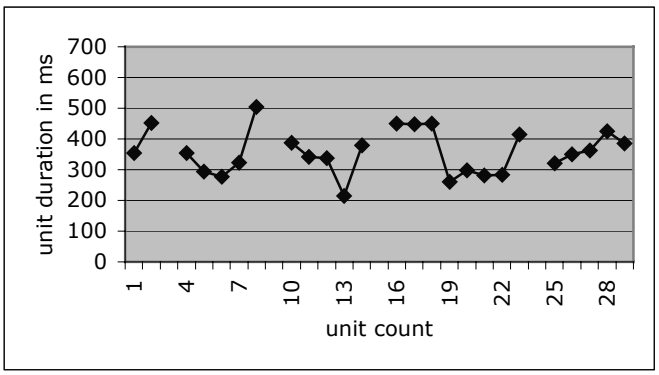


Fig. 2b. Durations for each rhythm unit in paragraph three of Article 1

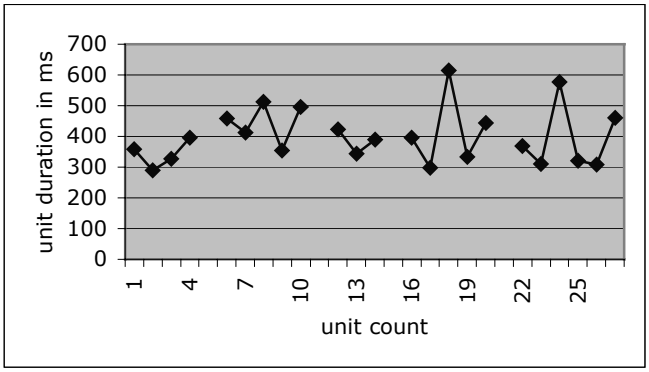


Table II shows the statistical analysis for rhythm unit durations for Articles 1 - 4 (the complete data set). Table II a. gives the scores for rhythm unit durations without including pre-pause units and II b. gives the scores including pre-pause units. Mean unit durations are again greater in b., indicating that the slowing down effect towards the end of utterance 'blocks' is consistent across the entire data set.

Table II Durations of rhythm units for Articles 1 - 4

	a. durations in ms without pre-pause units							**b. durations in ms with pre-pause units**						
article	***mean***	***median***	***SD***	***min***	***max***	***count***	***v***	***mean***	***median***	***SD***	***min***	***max***	***count***	***v***
1	382.3	368	97.8	178	676	144	25.6	391.4	390	94.3	178	676	191	24.1
2	428.6	421	122.4	177	781	211	28.6	444.8	433	129.2	177	781	259	29
3	430	405	142.7	143	847	118	33.2	453.1	427	148.7	143	852	131	32.8
4	429	410	122.7	146	716	223	28.6	433	418	122.4	146	716	257	25.3
1 - 4	***419.3***	***405***	***122.8***	***143***	***847***	***699***	***29.3***	***430.1***	***418.5***	***125.2***	***143***	***852***	***854***	***29.1***

These data sets clearly reject Hypothesis 1 – rhythmic unit variation is just too wide to claim isochrony as defined in the literature. But there is stability of a kind, for it is equally clear that rhythm unit duration is not random. The between paragraph results reveal this – the variation, though wide, is remarkably consistent. We might speculate that perhaps because of this it can be neutralised easily by the perceptual system – leading to perceived isochrony.

7.2 Hypothesis 2

For Article 2 in the series a correlation test was run comparing rhythm unit duration with the number of syllables within the unit. The result was a correlation coefficient of +0.54 (95% confidence) – a fair positive correlation. As rhythm units increase their number of syllables (in the data from this Article, from one stressed to one stressed with up to four unstressed syllables) their duration increased in a regular way. We interpret this as a clear indicator of no isochrony *as defined in the literature* to be measured. We emphasise the definition, for it may be that some other, as yet unformulated definition may yield the 'desired' result.

Table III Rhythmic unit durations related to syllabic composition

Article 2	***durations in ms***						
	mean	**median**	**SD**	**min**	**max**	**count**	**v**
str	354.5	366	111.5	177	673	63	31.5
str + (1 x u-str)	436.7	432	125.7	183	768	119	28.8
str + (2 x u-str)	497.3	487.5	110.4	267	781	74	22.2
str + (3 x u-str)	594	590	69.2	480	702	11	11.6

'str' = stressed, 'u-str' = unstressed

7.3 The predictive rhythm unit duration model

Most syllables in the data were of the type *stressed + unstressed* (i.e. two syllables) and the mean duration for this type was 436.7ms. Using this as our starting point we are now in a position to begin building a simple predictive model of rhythm, and we use this stressed + unstressed unit type as the ***basic rhythm unit***. Our model is based on rhythm unit ratios, and calculates the following rhythm unit durations from the starting point of a basic rhythm unit to which is assigned a value *L*:

```
basic_rhythm_unit = L;
        {
                if one_syllable_unit then L = L - (L*20/100);
                if two_syllable_unit then L = L;
                If three_syllable_unit then L = L + (L*15/100);
                if four_syllable_unit then L = L + (L*35/100);
                if five_syllable_unit then L = L + (L*55/100);
        }
```

That is, the ratio is:

```
        [62.4] : 81.2 : 100 : 113.9 : 136.1 : [155]
```

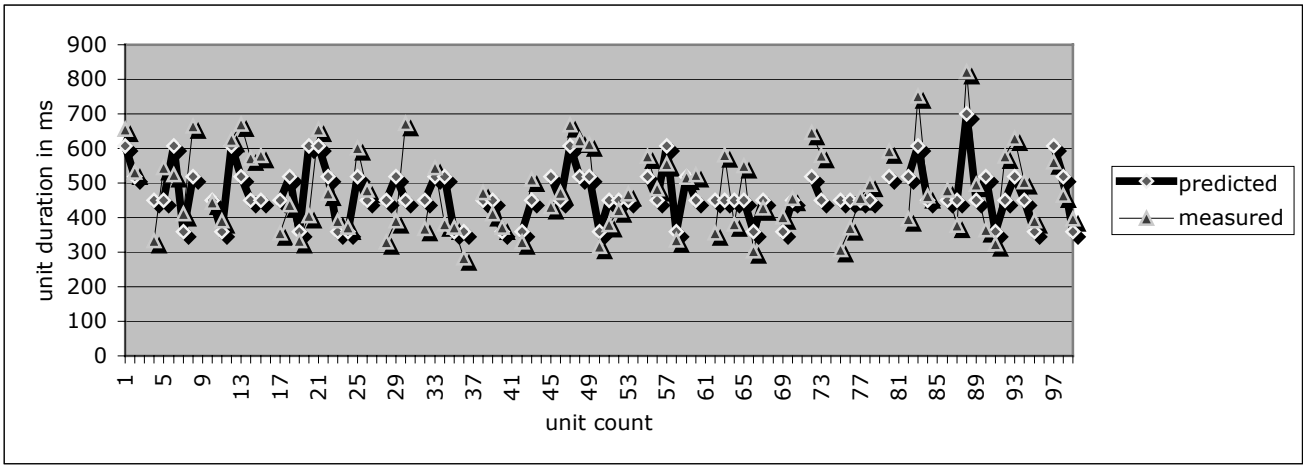
or, simplified:

```
        [62] : 80 : 100 : 115 : 135 : [155]
```

The bracketed values are to allow for end effects (see below). We shall use this as the basis for a predictive model.

To test the model we took the mean rhythm unit duration of the test data (excluding 'hanging' units) – which was 450ms – and calculated the durations of all units according to the above procedure. We used 450ms because this would automatically align our basic rhythm unit along the centre of the *y-axis* of the graphed data as determined by the data in the measured data. In a real situation we are free to instantiate *L* with any number we choose, provided our choice criteria are adequate. The results are shown in Fig.3 where the predicted durations and the actual durations are both plotted.

Fig.3 Predicted rhythm unit durations shown against measured unit durations in the test data (no utterance block end correction)

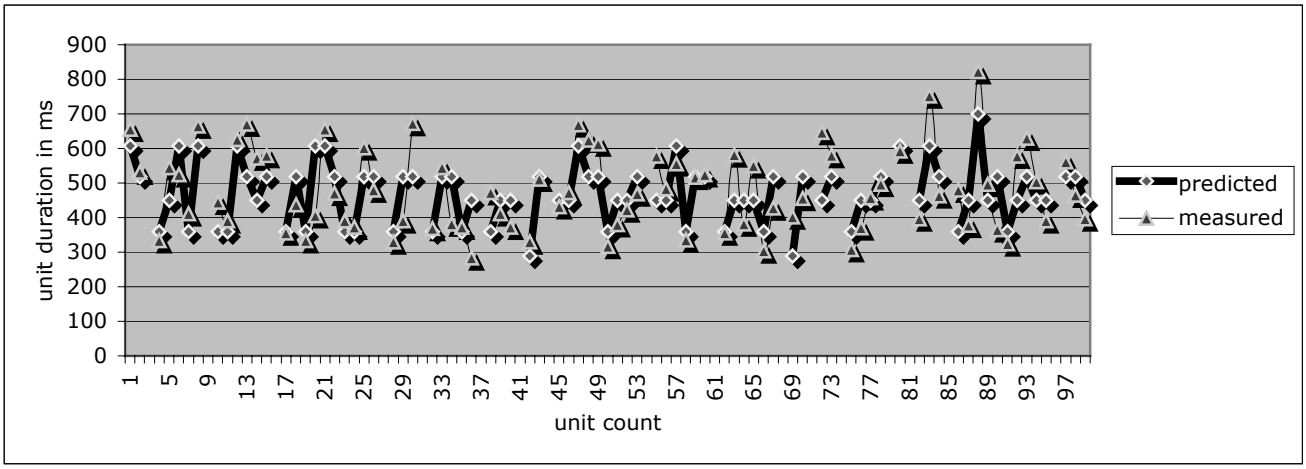


7.4 Hypothesis 3

In Fig.3 pauses of various types are show by gaps in the continuity of the data series. Note that two areas of poor fit between the predicted and measured data occurs before and after each pause. This was because the above procedure does not take into account the slowing down effect mentioned earlier, or an apparent slightly shorter rhythm unit duration immediately after a pause. Hypothesis 3 predicted no such effects, but the analysed data forces rejection of the hypothesis: pre-pause (or utterance block final) rhythm units tended on average to be around 20% greater in duration, whatever their syllabic composition. We examined, in this experiment, no further than this utterance block final unit – but we would expect that on closer examination the slowing down effect begins earlier in the block. However the noise cause by the variance among scores precluded reliable findings earlier than the final unit. Post-pause rhythm units showed a less consistent effect: they were *often* shorter by around 20%, though once again the experiment did not consider acceleration and deceleration effects in detail.

To move some way toward incorporating these pause effects in our predictive model, but only as a first approximation, we adopted the following end correction (applied after the above procedure): if the unit immediately follows a pause (and is not a hanging unit) use a value of *L* which corresponds to a unit with one fewer syllable. This effectively shortens the initial rhythm unit by around 20% of its duration. Similarly, if the unit immediately precedes an utterance block boundary use a value of *L* which corresponds to a unit with one more syllable. The results of applying these utterance block end correction are shown in Fig.4, where it is clear that the curve fit is improved. However, we feel further data and finer analysis might enable a yet more detailed set of algorithms to be devised that we can apply to synthesis.

Fig.4 Predicted rhythm unit durations shown against measured unit durations in the test data with utterance block end correction before each pause



8. DISCUSSION

If perceived isochrony has a physical correlate it is not to be found as temporally equidistant rhythm units – at least not in the data presented here. Yet listeners are sure about isochrony, can report it consistently and can report isochrony errors. Whatever perception is bringing to the cognitive assignment of isochrony we suggest the act is mediated by an acoustic signal to which the listener is demonstrably sensitive – as evidenced by the detection of errors.

For us, the task of speech synthesis is to manipulate the listener's perceptual system in such a way that they believe they are listening to real human speech, and we referred earlier to some of our other prosodics work which attempts this for intonation. In this paper rhythm is our concern, and by the same token as before we are seeking to create synthetic speech which manipulates the perceptual system to believe the rhythm has been human generated. Whether researchers have measured physical isochrony or not we have to create an acoustic signal which a listener will judge as isochronic and which will have detail of rhythm which is judged to be natural. Hence our predictive model.

In synthesis systems the segments of prosodics – syllables – are represented in some temporal format which is not going to be adequate for all required uses. In phoneme-based systems (Allen 1987, Holmes 1988) syllables have to be constructed and then assigned timing (Klatt 1979), the same is true of diphone-based systems. In syllable based systems (Tatham *et al*. 2000) of the 'stored normalised exemplar' type duration also needs to be calculated. In syllable or word based systems of the *unit selection* type (Morais *et al.* 2000) there is just a chance that the right duration syllable may be found, but in general this is not the case and recalculation here is also necessary. All these systems need a rhythm prediction model which will convince the listener that they hear real speech and, importantly, that will explain (by virtue of correct perceptual triggering) phenomena such as perceived isochrony.

9. CONCLUSION

Based on a simple statistical analysis of four short articles of read speech in slightly different styles, but using just one speaker, we confirmed general rejection of the standard isochrony hypothesis. We were able to refute the hypothesis that the number of syllables in each rhythm unit did not correlate with the unit's length. So, there was no isochrony, but there was syllable number correlation: this is in broad agreement with the earlier researchers (Lehiste 1977, Jassem *et al*. 1984).

Our task was to use these findings to build a generalised model of rhythm assignment which could be tested in a speech synthesis environment. A mainstay of the thinking behind the model was to be the need to 'explain' listeners' reactions to speech signals in respect of rhythm by predicting an acoustic signal which would trigger those same reactions – this was in line with the general strategy of our own synthesis system.

The predictive model was given a preliminary testing on a further passage of test data – a reserved portion of the original experimental data which was not used in any statistics or calculations on which the model was based. Results were promising in that natural rhythm trends were quite well tracked and the model exhibited the means to deal with utterance block 'start-up' and 'wind-down' effects.

We have presented a generalised predictive model which gives us a first approximation to solving the task. Based on the notion of 'basic rhythm unit' – a unit with one stressed syllable followed by an unstressed syllable – our model computes the general cases of units with other possible syllabic structures in English. By using a relative formulation in the model we shall be able to use it in a variety of different rate environments: indeed the next stage is to test the model in this way and begin a programme of systematic improvement on its basic structure.

10. REFERENCES

Abercrombie, D. (1967) *Elements of General Phonetics*, Edinburgh: Edinburgh University Press

Allen, J. (1987) *From Text to Speech: the MITalk System*. Cambridge: Cambridge University Press

Benguerel, A-P. and D'Arcy, J. (1986) Time-warping and the perception of rhythm in speech. *Journal of Phonetic*s 14, 231–246

Buxton, H (1983) Temporal predictability in the perception of English speech. In A. Cutler and D.R. Ladd (eds.) *Prosody: Models and Measurements*. Berlin: Springer-Verlag 111-121

Cummins, F. and Port, R.F. (1996) Rhythmic commonalities between hand gestures and speech. In *Proceedings of the Eighteenth Meeting of the Cognitive Science Societ*y

Donovan, A. and Darwin, C.J. (1979) The perceived rhythm of speech*. Proc. ICPhS IX*, I 268-274

Dauer, R.M. (1987) Phonetic and phonological components of language rhythm'. *Proc. XI ICPhS, Tallinn* 5, 447-450

Fry, D. (1958) Experiments in the perception of stress. *Language and Speech* 1, 126-152

Gimson, A.C. (1962 – first edition) *An Introduction to the Pronunciation of English*. London: Arnold

Hardcastle, W.J. and Hewlett, N. (1999) *Coarticulation – Theory, Data and Techniques*. Cambridge: Cambridge University Press

Hay, J. and Diehl, R. (1999) Effect of duration, intensity and f0 alternations on rhythmic grouping. *ICPhs99, San Francisco,* 245-248

Hill, D.R., Jassem, W and Witten, I. (1978) A statistical approach to the problem of isochrony in spoken British English. *Computer Science Technical Report* 1978-27-6, University of Calgary

Holmes, J.N. (1988) *Speech synthesis and recognition*. Wokingham: van Nostrand Reinhold

van der Hulst, H. and N. Ritter (1999) (eds.) *The Syllable: Views and Facts*. Berlin: Walter de Gruyter

Jassem, W. (1952) Stress in Modern English. *Bulletin de la Société Linguistique Polonaise* XII, 189-194

Jassem, W., Hill, D.R. and Witten, I.H. (1984) Isochrony in English speech: its statistical validity and linguistic relevance. In D. Gibbon and H. Richter (eds.) *Intonation, Accent and Rhythm*. Berlin: Walter de Gruyter, 203-225

Kelso, J.A.S (1995) *Dynamic Pattern*s. Cambridge, MA: MIT Press

Klatt, D. (1975) Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics* 3, 129-140

Klatt, D. (1979) Synthesis by rule of segmental durations in English sentences. In B. Lindblom and S. Ohman (eds.) *Frontiers of Speech Communications Research.* New York: Academic Press

Lehiste, I. (1977) Isochrony reconsidered. *Journal of Phonetic*s 5, 253–263

Morais, E., Taylor, P and Violaro, F. (2000). Concatenative text-to-speech synthesis based on prototype waveform interpolation (a time frequency approach). *Proc. ICSLP 2000*

Patel, A., Löfqvist, A. and Naito, W. (1999) The acoustics and kinematics of regularly timed speech: a database and method for the study of the p-center problem. *ICPhS99, San Francisco*, 405-408

Peterson, G.E. and Shoup, J.E. (1966) An acoustic theory of phonetics. *J. Speech and Hearing Research* 9, Washington D.C.: ASHA, 5-67

Pike, K. (1945) *Intonation of American English*. Ann Arbor: University of Michigan Press

Tatham, M. and Morton, K. (*in preparation*) Speaking rhythm in Montréal French.

Tatham, M., Morton, K. and Lewis, E. (2000) SPRUCE: speech synthesis for dialogue systems. In M.M. Taylor, F. Néel and D.G. Bouwhuis (eds.) *The Structure of Multimodal Dialogue II.* Amsterdam: John Benjamins, 271-292

Tatham, M., Morton, K. and Lewis, E. (2001) Re-engineering intonation in the synthesis of prosody. *Proc. Institute of Acoustics WISP 2001* – this volume

Williams, B. and Hiller, S.M (1994) The question of randomness in English foot timing: a control experiment. *Journal of Phonetic*s 22, 423–439