MARK TATHAM and KATHERINE MORTON

COMPUTATIONAL MODELLING OF SPEECH PRODUCTION: ENGLISH RHYTHM

1. Introduction

In this paper we examine the treatment of English rhythm from both a theoretical and an experimental perspective. There are major shortcomings in the way not just rhythm but also prosodics in general is modelled; this is all too clear in various applications of the theory, particularly in computational areas such as speech synthesis (Keller and Keller 2002). Our objective is to begin characterising rhythm computationally within a hierarchically based model in which prosody is the framework for speech production in general. Such a model tightly integrates suprasegmental and segmental properties of speech in a well defined and well motivated binding process.

Since the overall framework is the prosody the phonetic rendering of utterances depends on their prosodic structure; this includes the detail of the structure of the syllable – a unit of the prosody. The idea is not novel [see Firth (1948) on the prosodic structure, and Kahn (1976) and Gussenhoven (1986) on the structure of syllables, in particular the phenomenon of ambisyllabicity and detailed phonetic rendering], and it formed the basis of the conceptual design of our SPRUCE computational model (Lewis and Tatham 1991). The prosodic framework approach within SPRUCE is expressed in such a way that (a) it incorporates hooks for phonetic rendering with expressive content, and (b) it is transferable to a speech synthesis system for the kind of model testing associated with utterance synthesis. SPRUCE itself is not a piece of 'speech technology', rather it is a speech production model which, because of its computational nature, lends itself to being the basis of a speech synthesiser proper.

2. Some details of the model

The model is arranged on planes in two tiers (Figure 1), the logically prior and more abstract of the tiers being a prosodic/phonological characterisation of potential utterances. The double plane architecture formally separates a static conventional phonology from a similar but dynamic usable phonology. The lower, less abstract tier on each plane models the phonetic rendering of utterances. Thus, following linguistic convention we maintain a hierarchical 'separation of components' – though our theory of speech production defocuses the formal distinction between phonology and phonetics, certainly on the dynamic plane, when we are dealing with the cognitively mediated management of phonetic rendering (Tatham 1994). All this means is that in our view phonetic rendering is dynamically managed or supervised and that this is a complex cognitively sourced activity rather than a simple autonomous physical activity. We speak of the components as having two planes, and of each plane (in this fragment) as having two major components or tiers.

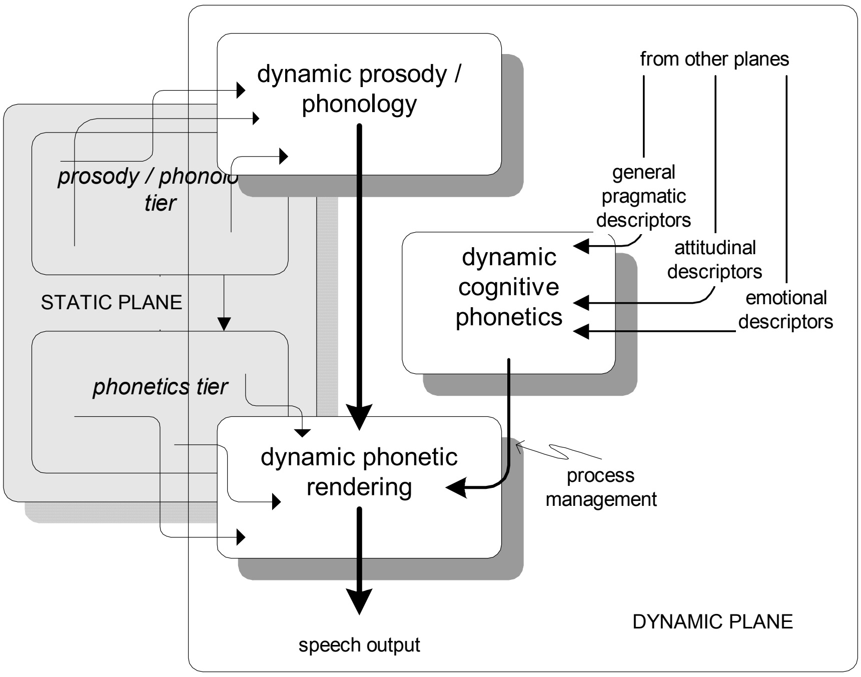

Figure 1: A fragment of the speech production model. Two planes are visible: to the back is the 'static plane' which holds characterisations of prosody/phonology and phonetics; to the front is the 'dynamic plane' holding the rendering flow pathway from dynamic prosody/phonology to phonetics to the final speech output. The rendering process is managed by dynamic cognitive phonetic processes which are mediated by supplementary descriptor inputs.

One plane of our phonology holds a description of the prosody (along with other non-prosodic phonological phenomena) of the language in general, whereas the other plane is designed to mirror but extend to characterising the dynamic planning of the prosody of particular utterances. For this reason we refer to this plane as 'derivative'. In computational terms the phonological model can be run to provide a dynamic (or real-time linked) expression of the prosodic structure of particular utterances, in which case only a subset of the processes it characterises for the language are called. This dynamic application of phonological processes is what leads directly to phonetic processes – also expressed on two planes in much the same way. The phonological and phonetic models therefore each exist on a dual plane structure – the static one characterising the system's potential without regard to any particular utterances, and the dynamic one characterising, in effect, grouped or unique instantiations of the first plane. All possible instantiations on this dynamic plane would, of course, in an abstract sense be equivalent to the potential characterised on the static plane – the planes are kept separate for formal reasons.

In our overall dynamic speech production model the planes co-exist and are simultaneously active. The difference is that one plane knows about (that is, characterises) all utterances and little or nothing of the uniqueness or dynamism of individual utterances, whereas the other knows only to activate those processes associated with the dynamic rendering of particular utterances or 'group types' of utterance. The static plane can exist and be inspected independently to reveal the general prosodic structure of the language, but this is not true of the dynamic plane – the dynamic plane only 'exists' when it is run: the activation flow as it runs is its existence.

If we rotate Figure 1 such that our two planes are horizontal for each of phonology and phonetics, then communication between them is vertical, enabling or inhibiting information flow to drop or climb between them. Communication is between points or nodes on the two planes: only a subset of nodes can communicate vertically – the sources of constraint which determine which nodes can be invoked are both internal and external to the prosodic/phonological system. These sources include the same kind of constraint between the nodes horizontally on the static plane itself – constraints which might, among other things, reveal something of the structure of mind, a general pursuit of linguistics. This concept is no different from the general aim of descriptive linguistics to focus on constraints and their sources. An example of a constraint might be the convergence of [+voice] and [-voice] plosives in word final position in German (*bund* vs. *bunt*) or their near convergence in English (*dog* vs. *dock*) – though here with 'contrastive feature transference' to the vowel offglide to ensure morphemic distinction.

We invoke here two planes for each of phonology and phonetics, but in fact our model is more complex that this, allowing for *n* planes, in rotation vertically linked *via* hooks to capture processing for the likes of expressive (stylistic, emotional and intentional) content (Morton 1992, Morton and Tatham 1995, Tams and Tatham 1995). Contained within the hierarchy there are additional tiers on one axis (equivalent to the phonological and phonetic tiers) and planes on the other axis which express the semantics and pragmatics, for example, of the language. These are also linked, node to node, rather than simply plane to plane. From these planes are derived the constraint information needed to attach expressive content to utterances. Thus a 'cubic' architecture for speech production proceeds *via* a series of tiers on the one axis and planes on the other axis with both inter-tier and inter-plane communicative activation.

The theoretical position which leads to this model makes a number of points. We highlight here those relevant to the present study:

- all utterances are rendered within a prosodic framework from which they cannot be divorced (Lewis and Tatham 1991)
- the prosodic framework is hierarchically organised (Morton and Tatham 1995)
- phonetic rendering is dependent on a simultaneous review of the entire underlying linguistic structure (Tatham and Lewis 1992, Morton et al. 1999)
- the rendering process is managed within a wider 'semantic delivery' system (Tatham 1991, Tatham and Morton 1995) operating within a specific scenario, a supervisory process overseeing critical areas of the rendering (Tatham 1994, 1995).

To illustrate, here is a fragment, in traditional form, of the general prosodic markup, instantiated to the utterance *'Better be safe than sorry'*. A traditional phonetic transcription – | 'bɛ–tə–bɪ | "seɪf–ðən | 'sɒ–ri | – shows some allophonic detail (that is, it is not an abstract phonemic transcription), including marking of rhythmic unit boundaries (… –ðən | 'sɒ …), syllable boundaries within rhythmic units (… –tə–bɪ …), primary-stressed syllables (…'sɒ …), and nuclear sentence stress on one syllable (… "seɪf …). The prosodic markers used here focus on rhythm rather than intonation.

We can set up a hierarchical characterisation of this utterance (on a 'specific' plane) and illustrate this in Table 1, which is a fragment of the entire prosodic characterisation of the utterance. For illustrative purposes a transcription of the utterance is shown in the lowest row – though in practice its phonetic detail has yet to be added: this transcription is at best 'extrinsic allophonic' in terms of detail. Upper rows showing intonational phrase and accent group have been omitted, but the general hierarchical relationship is:

> *IP → AG (AG ...) → (...ρ) ρ (ρ ...) → σ (σ ...) → (o) r → n (c)*
> [where *IP* = intonational phrase, AG = accent group, ρ = foot (instantiated as 'rhythm unit'), σ = syllable. Syllables take the traditional onset + rhyme (→ *nucleus* + *coda)* form. Note the arguable syllable sharing (ambisyllabicity) of a couple of segment units (shaded cells).]

<table>
<tr><td colspan="9">ρ</td><td colspan="6">ρ</td><td colspan="6">ρ</td></tr>
<tr><td colspan="3">σ</td><td colspan="3">σ</td><td colspan="3">σ</td><td colspan="3">σ</td><td colspan="3">σ</td><td colspan="3">σ</td><td colspan="3">σ</td></tr>
<tr><td rowspan="2">o</td><td colspan="2">r</td><td rowspan="2">o</td><td colspan="2">r</td><td rowspan="2">o</td><td colspan="2">r</td><td rowspan="2">o</td><td colspan="2">r</td><td rowspan="2">o</td><td colspan="2">r</td><td rowspan="2">o</td><td colspan="2">r</td><td rowspan="2">o</td><td colspan="2">r</td></tr>
<tr><td>n</td><td>c</td><td>n</td><td>c</td><td>n</td><td>c</td><td>n</td><td>c</td><td>n</td><td>c</td><td>n</td><td>c</td><td>n</td><td>c</td></tr>
<tr><td>b</td><td>ɛ</td><td colspan="2">t</td><td>ə</td><td></td><td>b</td><td>ɪ</td><td></td><td>s</td><td>eɪ</td><td>f</td><td>ð</td><td>ə</td><td>n</td><td>s</td><td>ɒ</td><td colspan="2">r</td><td>i</td><td></td></tr>
</table>

Table 1: Notional graph of the hierarchical relationship between identified prosodic units

For several decades notions of syllable structure have referred to phonological ambisyllabicity but seldom highlighted its real influence on the detail of phonetic structure rendered from the phonological prosodic representation. The researcher who first brought this out clearly seems to have been Kahn (1976) who investigated ambisyllabic constraints on low level allophonic processes, including, for example, the systematic occurrence or not of aspiration. This is in effect the supervised management of phonetic rendering – a concept developed by Tatham within the Cognitive Phonetics paradigm (Tatham 1995) implemented in the SPRUCE computational model. Kahn's groundbreaking ideas (supplemented by those of Gussenhoven 1986) resurface also in the YorkTalk synthesiser (Local and Ogden 1997) and later in Ogden et al. (2000), though, unlike SPRUCE, the focus here is on speech synthesis rather than a computational model of human speech production.

3. Prosodics of the acoustic signal – moving toward synthesis

Whatever theoretical position is taken it is probably the case that we lack detailed knowledge of the correlates of prosodic phenomena in the acoustic signals, or indeed in the actual articulations and aerodynamics which produce the acoustic signals. In this paper we confine the discussion to the acoustic signal – since acoustic models are those which are used most in synthesis, certainly in the synthesis we use to test our models. By looking closely at the acoustic signal we expect to find what others have found before us, of course, but also we hope to be able to highlight aspects of the signal which might perhaps have been dismissed as random variability. In earlier discussions of variability we warned against the view that variability which is not predictable from the phonology is not linguistically significant. It may be that some of the variability in the acoustics of speech is more predictable than is generally taken to be the case. We are particularly interested in finding out more about naturalness – humanness – for synthetic speech and the ability of a listener to declare that samples of acoustic signal are from the same speaker. To put the last point another way, we are interested to discover what it is that causes the a person listening to synthetic speech to doubt the integrity of the signal: it has a temporal jerkiness (usually this is what is meant by integrity), but it should also have something that human speech has: integrity of speaker.

In several papers we have addressed the question of intonation and discussed how we currently handle this in the SPRUCE computational model (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.

4. Rhythm – preliminaries

There is a basic assumption in the literature that speech is rhythmic; this assumption precedes any discussion of isochrony. We define rhythm as the patterned temporal occurrence of events. But precisely what physical events contribute to the acceptance of rhythm as a feature of speech is not obvious and there is by no means general agreement. 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). Given this perceptual focus the research question might be reformulated as:

1. Can we show not that speech is rhythmic [an assertion at the acoustic level] but that speakers intend rhythm and that listeners can always perceive rhythm? – and
2. Can the acoustic signal and the results of its perception be directly related?

Another question we might ask: Is rhythm in speech inevitable? If it is, then it cannot be said to be planned (except in the sense that if its inclusion is inevitable, it must be allowed for or accommodated). And if rhythm is intrinsic to speech then systematic variations for expressive effect might be modelled as modulations on this 'carrier' – not as 'fresh' patterns generated anew each time. Put another way we need to discover whether rhythm is planned as part of the general linguistic phonological planning of speech or whether it is simply an intrinsic property of speech – part of how speech is. Another way of putting this in the terms of Figure 1 is:

1. On which plane and in which tier do we place the origin of rhythm? – and then
2. How does rhythm permeate the model?

These are not trivial questions. Even if rhythm is an intrinsic property of speech we would want to know whether it is an intrinsic cognitive property or an intrinsic physical one. The fact that all speech is rhythmic in some sense does not throw light on the origin of this rhythmic property, whether it be intended, perceived, intrinsic-cognitive or intrinsic-physical.

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", writes Abercrombie (1967, p.97). It's suggested that the units of rhythm for some languages, English included, 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, is the concomitant claim. These researchers and many others more recently have observed that for whichever type of language the temporal phenomenon that raises the status of rhythm to stressed syllable timing in speech is isochrony – that is, the equal timing of these units. So, in English for example, there might 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 either in the acoustic rendering or even sometimes as a perceived property.

Dauer (1987 and in other papers) points out that in fact the distinction between different types of language is best modelled as bimodal rather than 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 other languages 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 can shift the isochrony along the scale as, for example, when some expressive pause is inserted before a word:

> 'The e |***vent*** took |***place*** at the |***com***pany's ^^ Los |***An***geles ^^ fa |***ci***lity.'
>
> [The marking conventions here are: stressed syllable – bold, italic; ^^ = 'special' pauses; | - rhythm unit boundary marker.]

– in which pauses before and after the sequence *Los Angeles* serve as highlight markers for contrastive emphasis, i.e. 'not their San Francisco facility'. This is a good example of interference, or interaction, between the rhythm unit sequence (which ignores word boundaries) and expressive marking, which does not (cf. for example Ogden et al.'s [1999] view [shared with others] that phonological marking does not generally need to include word boundaries).

5. Isochrony as a perceptual phenomenon

Isochrony as a feature of speech rhythm is generally viewed as 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). Because of the difficulties of matching acoustic signals with the results of their perception 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 within a well established paradigm which seeks to identify the acoustic correlates of perceptual units in general.

There are, of course, several possibilities for making the measurements on the acoustic signal. Do we measure from the start of a portion of the acoustic signal judged to correlate with a stressed syllable to the start (or end) of the next stressed syllable? Do we measure from the point of peak amplitude in the stressed syllable to the point of peak amplitude in the next stressed syllable? Even finding the syllable itself can be difficult, since the notion 'syllable' is, of course, an abstraction borrowed from phonology: we are really measuring an acoustic signal which correlates with a stressed phonological syllable. [As an example consider the differences in certainty in locating the acoustically correlating start point of the stressed phonological syllable in 'pea |***stalks***' compared with its start in 'peace |***talks***' – the onset of a voiceless fricative [s] being easier to spot than the onset of a voiceless plosive [t].]

In attempts to combine both perceptual and objective approaches some researchers (e.g. Cummins and Port [1996]) have hypothesised what they call the p-centre ('perceptual centre). This is a point, determined by experiment, in the acoustic signal which marks the perceived temporal mid- or centre-point of a stressed syllable – perhaps it may be thought of as the perceptual centre of gravity of a syllable. As such, this point is a candidate marker for the stressed beat of perceived isochrony: so for these authors measurements of rhythmic units on the acoustic signal are from p-centre to p-centre rather than from some other point.

Some researchers have tried various transformations of the data to try to build an isochrony model at the acoustic surface. Thus we find early work by Hill et al. (1978) trying to determine an index based on some intrinsic period in rhythmic unit repetition. And again, Jassem et al. (1984) use a relatively elaborate statistical technique in their quest for finding hidden isochrony in the acoustic signal. Later, 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 their measurements were exhaustive – their statistical analysis revealed a very slight, significant tendency towards isochrony in the measurements.

6. The 'acoustic correlates' paradigm

Investigating the relationship between acoustic measurements and cognitive phenomena has a long tradition in experimental speech studies, underlining clear recognition of the abstraction differences between cognitively assigned labels like 'syllable', 'intonation', 'stress' and sections of a physical acoustic wave. It is not the case, however, that the acoustic correlates paradigm is universally understood: there is a great temptation to believe that syllables, for example, are being measured directly and have some physical status.

In general research using the acoustic correlates paradigm focuses on the speaker, and is in line with the traditional approach in linguistics that underlying the acoustic signal (or any other phonetic signal) is a cognitive representation. An object *X* in the cognitive representation correlates with an object *X'* in the physical signal. It is quickly clear that the correlation between *X* and *X'* is often non-linear and/or often not one-to-one. But strictly we need a slightly different paradigm if we cannot establish independently that isochrony is part of the plan: we are after the acoustic correlates not of the underlying plan, but of the final percept. The percept may be a reproduction of the original plan, but it may not be. So, we must be clear: are we looking for the acoustic correlates of planned isochrony or of perceived isochrony? This seems like splitting hairs, but does make sense in the light of uncertainty as to whether isochrony is planned: researchers seem to agree that it is perceived and tend not to discuss whether or not it is planned.

It seems to us that in producing utterances 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 isochrony is indeed planned. An alternative explanation mentioned above, 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 pointing this out, even if we reserve judgement on agreeing with it, 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. Non-awareness of a speech property is usually taken to mean that it is non-manipulable as part of the linguistic encoding process. But see the arguments to the contrary adduced in the general theory of Cognitive Phonetics (Tatham 1991) and the specific theory of Supervised Rendering (Tatham 1994 and 1995).

But even if speakers and listeners are aware of isochrony, why would it be planned? One perhaps weak suggestion is that it may form a baseline rhythm against which to measure stylistic or expressive departures from such a rhythm. But it does not need to be planned for this: if an isochronic rhythm were intrinsic it could be used as the baseline – just as some coarticulatory effects are intrinsic and used as a baseline for stylistic or accent effects. For example, the New York City cognitively derived and managed enhancement of intrinsic oral vowel nasalisation between nasal consonants is often cited in the Cognitive Phonetics literature (for example, Tatham 1986, 1990). If isochrony were planned it would appear in the prosodic tier on the static plane, if intrinsic it would likely appear in the phonetics tier since the phenomenon must be generally 'known' in order to be manipulated during rendering.

7. Synthesising isochrony

Certainly for the purpose of testing a production model, but also for the purpose of designing synthesis systems one might 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 finding 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. In other words, if it is true that listeners prefer isochrony it needs to be triggered by the acoustic signal irrespective of whether it is planned or not. The feeling of unease in those listening to synthetic speech which does not trigger the feeling of isochrony is, we hypothesise, similar to listeners' unease if intrinsic coarticulation is badly rendered in synthesis.

But to synthesise something which gives rise to the illusion of isochrony is a tall order since the literature is about looking for acoustic 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. That is, researchers have looked for ways to process the acoustic data they have gathered to somehow reveal a hidden isochrony – implying that the perceptual processing of the listener has similarly retrieved the isochrony. However, this is only the first of two ways of setting up a basic perceptual model – either

- perception processes the incoming data until the percept is found in the data, or
- the incoming data is used to trigger a percept that is already in the listener’s mind and which is then assigned to the data.

We feel that we need to discover what effects there are which might trigger the perceptual effect and do not require elaborate statistical processing arguably not at the disposal of the listener. 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? The experiments we report below reflect that we are asking the former question.

Why not simply synthesise an acoustic signal which actually does have physical isochrony in the hope that this will do? But however attractive the idea from a practical viewpoint (though not from a theoretical one), 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 adaptively adjust it as part of the perceptual process to something more normal. In other words although it might seem that physical isochrony would help the listener perceive isochrony, paradoxically the listener seems perplexed and reports the signal to be unnatural. The listener seems not to adapt instantly to the unusual signal. Clearly how and why this is the case needs proper investigation, and constitutes a research topic which we are pursuing.

Positive results might point 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 Kelso 1995 and Patel et al. 1999). A tentative conclusion here is that listeners (and speakers) are aware of the shortcomings of the phonetic rendering process and so used to them that they react unfavourably if they are presented with a signal which should be more helpful to them in the decoding process. Paradoxically they seem to reject it as wrong in some sense rather than accept it as helpful. This could be an indication that the whole idea of isochrony is wrong or it might be telling us that there are interesting non-linearities in the perceptual system which prompt further investigation.

8. The experimental investigation – isochrony

To construct a rhythm model testable by speech synthesis and of practical use we needed yet more data. Although many researchers had already investigated the problem and reported the results of 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. Thus 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 broadened out later if results proved promising. But in addition our practical speech synthesis system is designed to speak in a read speech manner (for example, in reciting retrieved information from a database) rather than in short sentences or in a conversational mode. Also, for these experiments, we used the speech of only one person, on the grounds that although we cannot dismiss inter-speaker variability as trivial and possibly therefore likely to colour our results, measuring the speech of one person removes the 'noise' of inter-speaker variability, and easily points the way to re-testing the hypotheses with new speakers. No one can deny that different speakers may do things differently but it seems a good idea to model one speaker before moving on to make the problem more complex by introducing inter-speaker variability: intra-speaker variability is hard enough to understand in prosodic phenomena. The usual argument against this single speaker approach is that we cannot guarantee that the speaker is behaving typically; perhaps not, but we *can* guarantee that the speaker is a successful and coherent speaker – so must be doing something appropriate.

9. Working definitions, data assembly, hypotheses

9.1. 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) and includes its entire onset. The rhythmic unit should not be confused with the 'foot' used by writers such as Abercrombie (1967, p.131). The foot is effectively a cognitive unit of planning or of perception, whereas the rhythmic unit is one which is physical and measurable.
- 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. In a surface linear characterisation, 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). The consonants preceding the vowel nucleus of the syllable are referred to as its onset, and those following the vowel (and included with it as the syllable's rhyme) are termed its coda. The hierarchical structure, put in the context of the overall model in the detail of Section 2 above, is:

  syllable → onset+rhyme → [onset+] nucleus+coda.

  For the purposes of measurement the physical acoustic 'syllable' is the unit. To measure the duration of rhythmic units we needed to accurately endpoint the acoustic syllable (see below in the section on measurement).

  > [Note: in our measurements on this occasion we have taken no account of possible ambisyllabicity – the phenomenon of consonant membership of two adjacent syllables: rightmost in the first (with coda status) and leftmost in the second (with onset status) – even across foot, and hence rhythmic unit, boundaries. Ambisyllabicity as a concept figures in our theoretical prosodic framework, even at the level of its engineering application (Tatham and Lewis 1999, Lewis and Tatham 2001), but rarely plays a role in the isochrony studies we are trying to replicate and extend (an exception is found in the temporal modelling strategy briefly outlined in Ogden et al. 1999).]

- A stressed syllable is one which carries primary phonological stress: that is, reflects planned prominence which the listener perceives from (importantly not in) 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). However, bear in mind that just

> as with rhythm in general we are not talking of stress found in the signal, but of something in the signal which triggers stress to be perceptually assigned.

9.2. Data assembly

One subject, a female speaker of the general accent of California, read out loud the front page of *The Los Angeles Times* for 25th December 2000. This consisted of half a dozen brief 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 'CoolEdit96' (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 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 beginning modelling a speaker's production of rhythmic structure. In our experimental paradigm we are reporting the 'proof of concept' stage. A single speaker was used partly because our aim was a coherent model of one speaker and partly to eliminate inter-speaker variability. One of the reasons why researchers have difficulty in finding anything in the acoustic signal indicative of regular rhythm may be that their statistics cannot cope with the 'noise' generated by inter-speaker variability. Wishing to characterise isochrony as a property of the language (rather than some idiolectal artefact) researchers will often pool data from more than one speaker. If we find isochrony or some other patterning in one speaker of the language, we can always go on to look for the same correlates in another speaker.

9.3. Hypotheses

We formulated a number of hypotheses based on researchers' earlier findings. We do not expect to find anything different from earlier work, so we are able to formulate an expectation on which to base the hypotheses. There is nothing novel about these experiments except that the data is being collected and interpreted from a different theoretical perspective than hitherto. So, three expectation based hypotheses, couched to invite refutation (that is, as null hypotheses) follow:

> $\boldsymbol{H_1}$ Any pattern of rhythmic units observable in the data is isochronic – expectation: we shall find no statistically significant isochrony.
>
> $\boldsymbol{H_2}$ 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.
>
> $\boldsymbol{H_3}$ 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 almost all previous 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 there is 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 also 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 when suggested by our general prosodic framework (see Section 2 above).

10. Data measurement and analysis

10.1. 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. Figure 2 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.

> [Note: There are some problems in setting up measurement rules relating to mixing levels of abstraction. Syllables are phonological units, but acoustic events are actually what are being measured – and the assumption is of a one-to-one correlation (or at least a predictable correlation) between the two. Syllables comprise individual segments at the phonological level, and since much is known of the acoustic correlates of individual segments we aimed to use these as anchors these when looking for the 'boundaries' of syllables. This is strictly an unsound approach (the mixing of levels) from the point of view of the theory, but is the one most commonly pursued.]

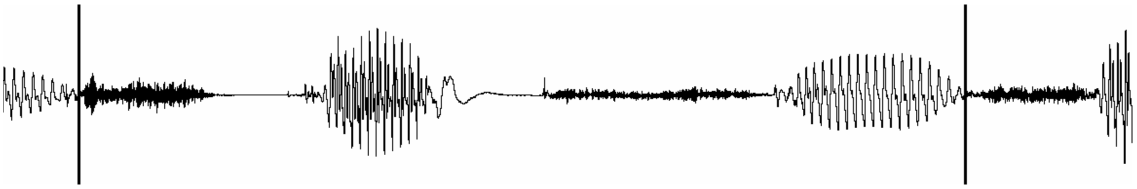

Figure 2a: Rhythmic unit | **steps** in | . At the start of the rhythm unit [s] overlaps a preceding [n], but is taken to start where [n] vocal cord vibration stops. At the end the rule is the same: stop the [n] of inside where vocal cord vibration stops despite slight overlap from the following [s].

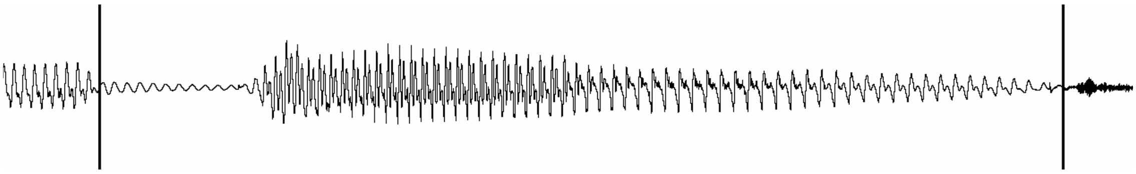

Figure 2b: | [b] + VC[ən] | – a rhythmic unit zoomed to show the moment of stop onset (closure) for the initial [b] of this unit. Note that this speaker carries the vocal cord vibration right through the 'voiced' [b]. The unit ends where the vocal cord vibration for the final [n] despite slight overlap with frication for a 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, look at just one sentence in Article 1, beginning: *"For | **her** and her | **friends** | ..."* which is a hanging unit (underlined) 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 certain 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 2a. (without pre-pause units) and 2b. (with pre-pause units). The corresponding sample graphs (Figure 3) show the speed reduction trend and this is shown in the tables by an increase in the mean unit durations.

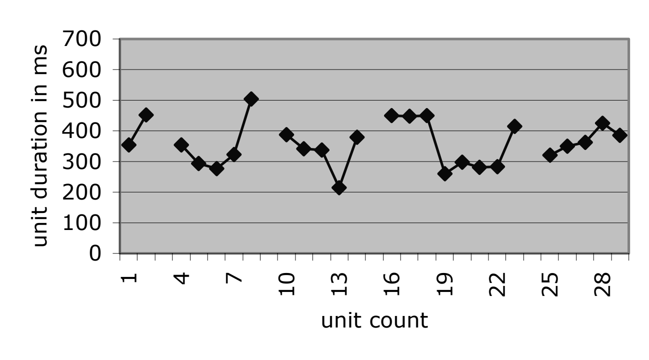

Fig.3a – Durations for each rhythm unit in paragraph one of Article 1

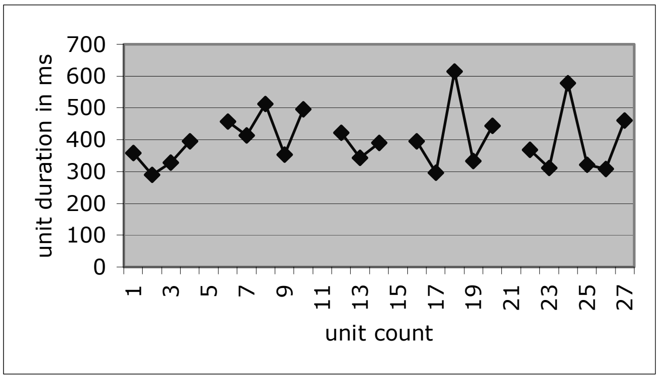

Fig.3b: Durations for each rhythm unit in paragraph three of Article 1

	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***

Table 2 – Durations of rhythm units in Article 1, by paragraph

Table 3 shows the statistical analysis for rhythm unit durations for Articles 1 - 4 (the complete data set). Table 3a gives the scores for rhythm unit durations without including pre-pause units and 3b gives the scores including pre-pause units. Mean unit durations are again greater in 3b, indicating that the slowing down effect towards the end of utterance 'blocks' is consistent across the entire data set.

	a. durations in ms without pre-pause units							**b. durations in ms with pre-pause units**						
article	***mean***	***med.***	***SD***	***min***	***max***	***count***	***v***	***mean***	***med.***	***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***

Table 3 – Durations of rhythm units for Articles 1 – 4

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.

10.2. 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 earlier literature to be measured. We emphasise the definition, for it may be that some other, as yet unformulated definition may yield the 'desired' result.

Article 2 *durations (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

Table 4: Rhythmic unit durations related to syllabic composition: 'str' = stressed, 'u-str' = unstressed

10.3. The predictive rhythm unit duration model

Most rhythmic units 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 finding 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 focuses 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 these formulae which describe not only the unit lengths, but also their relationship between each other, as the basis for a predictive model.

10.3.1. Testing the predictive rhythm unit duration model

To test the above 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 set. In a real situation we are free to instantiate *L* with any number we choose, provided we have adequate criteria for making the choice. The results are shown in Figure 4 where the predicted durations and the actual measured durations are both plotted.

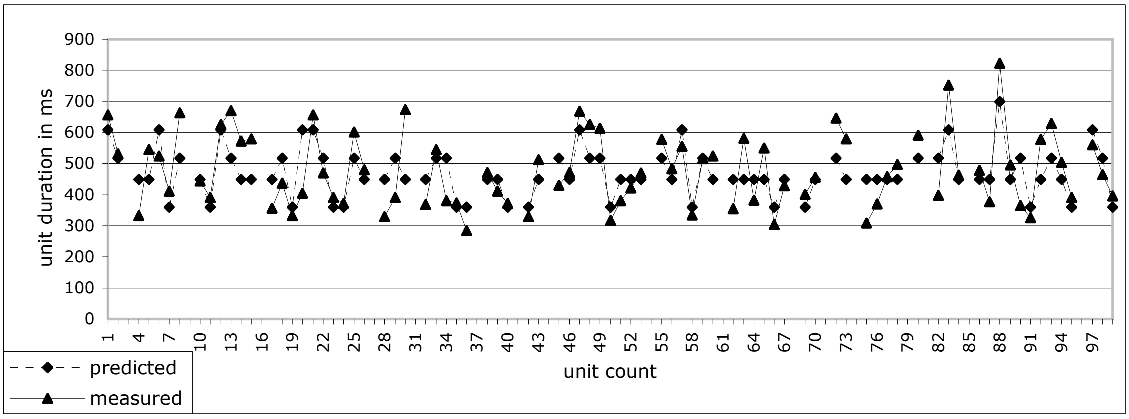

Figure 4: Predicted rhythm unit durations shown against measured unit durations in the test data (no utterance block end correction)

10.4. Hypothesis 3

In Figure 4 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 due to the variance between 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.

> [This description implies that pause effects are a surface phenomenon – perhaps an effect generated by the rendering process. An alternative explanation is that they are programmed by the underlying structure of the utterance: this explanation would imply that there are elements or combinations of elements in the underlying structure which can determine how and when the pausal effects occur. If this is the case, then it could be tested by altering the underlying structure, predicting how the surface pattern would change. If the result were a similar effect, whatever the underlying structure, there would be no change at the surface.]

An analogous situation occurs if contextual effects at the surface are said to be determined by choice of allophone at a deep level (or by reason of the deep structure). An analysis or transcription of the word initial consonants in the sentence *'A* **c***up of* **t***ea'* might produce $[_{a}k_{u}]$ and $[_{\#}t_{i}]$ respectively. In this analysis each consonant is indexed with its immediate left and immediate right contexts to produce 'context sensitive allophones' [Wickelgren 1969]. A deep explanation of context sensitive allophones (the one favoured by Wickelgren) would predict their intact metathesis in the accidentally generated sentence *'A* **t***up of* **k***ea'*. But in fact this does not happen: the consonants appear as: $[_{a}t_{u}]$ and $[_{\#}k_{i}]$, suggesting that the contextual effects have been added after the metathesis – and indeed are very low level coarticulatory effects. Note also that the local fundamental frequencies are not transposed either. In our general prosodic framework concept segments are dominated by the entire prosody of an utterance; this enables underlying structural constraints to manage the phonetic rendering process, and this is one way in which prosodic and segmental rendering can behave independently if necessary whilst being in principle closely linked.

To move some way toward incorporating these pausal 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 Figure 5, 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.

Although the model produces good results we are aware that it offers no real explanatory power: it succeeds in predicting the correct result when set against the measured results, but does not of itself indicate the origin of the effect – unless this is simply a surface rendering artefact.

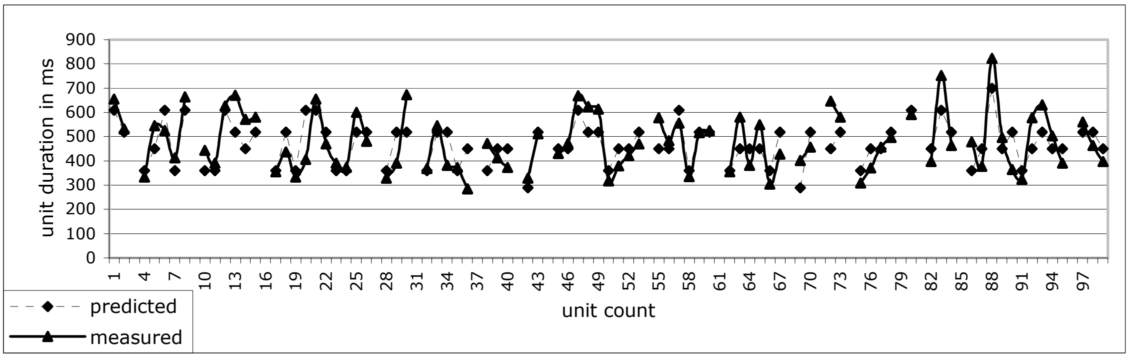

Figure 5: Predicted rhythm unit durations shown against measured unit durations in the test data with utterance block end correction before each pause

11. Discussion

If the perceived isochrony between successive feet in an utterance 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. Indeed interference with isochrony is a possible way of conveying a stylistic effect; our theory allows for this as an expressive use of rendering supervision (Tatham 1995 – and see Figure 1), a prosodic framework based version of the dialectal supervision which gives us heightened nasality in some U.S. accents of English (Tatham and Morton 1980, Tatham 1990). Whatever processes perception brings to the assignment of isochrony as a cognitive phenomenon we suggest the act is mediated by an acoustic signal to which the listener is demonstrably sensitive: listeners do readily report detecting errors of rhythm in synthetic speech, for example. If they feel a mistake has been made they must have an expectation which has been challenged – and that expectation is clearly predictive.

For us the task of practical speech synthesis is to manipulate the listener's perceptual system in such a way that they believe they are listening to coherent 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 and is coherent in the sense that it is coming from a single speaker and has been uttered on a continuous basis (i.e. has not been pieced together from a database). Whether researchers have measured physical isochrony or not, we want 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 and our underlying overall prosodic framework capable of passing constraints to the phonetic rendering processes. This is the way in which the required fine phonetic detail is assured.

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. This is the unit selection (Black and Campbell 1995) problem: This approach hypothesises the availability in principle of a suitable unit for a particular 'slot' in an utterance; it is not difficult to show, however, that to find every possible unit would require a database of infinite size. If the 'strict selection' principle is adhered to then the model fails (just as slot grammars of syntax must fail) unless the selected unit is allowed to be modified (the 'surface adjustment principle'). The other approach is to build an inventory of normalised units capable of modification: this approach guarantees the availability of the correct unit. There is an useful analogy here: suppose we were selecting units for rendering the words *port* and *sport* in English. A strict selection approach (which must retrieve the correct final rendering from a database of final renderings) needs to find [p^h] and [p] respectively, but the alternative would be to find /p/ which is correct for both – but is a more abstract representation. Subsequent rendering (surface constrained – as in early Transformational Grammar phonology or in Classical Phonetics; or deep constrained – as in hierarchical prosodic models of the kind we use in SPRUCE) derives the correct [p^h] or [p]. Apart from evaluation of these approaches on a theoretical basis, subjective evaluation by listener would focus on utterance integrity (or coherence as detailed above). An utterance with integrity stimulates listener confidence concerning the utterance source – one talker, one occasion.

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 listeners that they hear real speech and, importantly, that will explain (by virtue of correct perceptual triggering) phenomena such as perceived isochrony.

We incorporate these findings into our general model of speech production by characterising rhythm initially in the phonological tier and on the static plane. Here rhythm is isochronic and the positioning of isochrony at this points accounts for speaker and listener feelings that feet are 'equidistant in time'. In the phonetic tier on the dynamic plane the utterance is rendered under supervision to produce a measurable output of rhythmic units. The fact that we have been able to begin to develop a predictive model indicates that the rhythmic units are structured, for without inherent structure a predictive model would not be possible. This physical structure is not itself isochronic, but has been derived from an isochronic plan and leads to the correct perception of the speaker’s isochronic plan. Notice, though, that the physical mechanism has considerable inherent variability in many parameters. This detail is either discarded by the listener or is used to reconstruct a copy of the speaker's intended plan – that is there is close correlation between speaker plan and the perceiver's reconstructed plan, but not always an obvious relationship between these and the mediating acoustic signal.

12. Conclusion

Based on a simple statistical analysis of four short articles of read speech in slightly different styles and using one speaker, we confirmed general rejection of the standard isochrony hypothesis as far as physical timing was concerned. 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. Important in the thinking behind the model was 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. At the same time we had to ‘position’ the various stages of rhythm generation within the overall production model, at least to characterise speakers' feelings that their own speech has rhythm and to indicate the mechanism for deriving from this planned rhythm an actual physical pattern for listener use.

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. In particular we needed to maintain integrity of the signal to ensure listeners' feelings that the entire signal emanated from the same speaker on the same occasion. In this latter requirement we fell a little short, though we felt that much progress has been made and that the model marks forward progress nonetheless. There can be do doubt that there is scope for refinement of the model based on further data collection. It should be pointed out, however, that we do not know exactly how short our model fell of its objective, because an exact physical description of intrinsic rhythm is not available (since it is an abstract concept); it is possible that with proper 'expressive content' (emotional or attitudinal effects) modulated onto this intrinsic carrier some of the lack of integrity will disappear – that is, expressive content will improve naturalness. We are investigating this idea.

We have presented a generalised predictive model which gives us a first approximation to modelling rhythm in speech. 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. This model forms the basis of our modelling of rhythmic aspects of expressive content in speech.

References

ABERCROMBIE, D. (1967): Elements of General Phonetics. Edinburgh.

ALLEN, J. (1987): From Text to Speech: the MITalk System. Cambridge.

BENGUEREL, A.-P. / D'ARCY, J. (1986): Time-warping and the perception of rhythm in speech. In: Journal of Phonetics 14, pp. 231–246.

BLACK, A. W. / CAMPBELL, N. (1995): Optimising selection of units from speech databases for concatenative synthesis. In: Proceedings Eurospeech '95, Madrid, pp. 581-584.

BUXTON, H. (1983): Temporal predictability in the perception of English speech. In: CUTLER, A. / LADD, D. R. (eds.): Prosody: Models and Measurements. Berlin, pp. 111-121.

CUMMINS, F. / PORT, R. F. (1996): Rhythmic commonalities between hand gestures and speech. In: Proceedings of the Eighteenth Meeting of the Cognitive Science Society, Mahwah, NJ, pp. 415-419.

DAUER, R. M. (1987): Phonetic and phonological components of language rhythm. In: Proceedings of the XIth International Congress of Phonetic Sciences, Tallinn, vol. 5, pp. 447-450.

DONOVAN, A. / DARWIN, C. J. (1979): The perceived rhythm of speech. In: Proceedings of the IXth International Congress of Phonetic Sciences, vol. 1, pp. 268-274.

FIRTH, J. R. (1948): Sounds and Prosodies. In: Transactions of the Philological Society, pp. 127-152. [Reproduced in: JONES, W.E. / LAVER, J. (eds.): Phonetics in Linguistics: a Book of Readings. London.]

FRY, D. (1958): Experiments in the perception of stress. In: Language and Speech 1, pp. 126-152.

GUSSENHOVEN, C. (1986): English plosive allophones and ambisyllabicity. In: Gramma 10, pp. 119-141.

GIMSON, A. C. (^{1}1962): An Introduction to the Pronunciation of English. London.

HARDCASTLE, W. J. / HEWLETT, N. (eds.) (1999): Coarticulation – Theory, Data and Techniques. Cambridge.

HAY, J. / DIEHL, R. (1999): Effect of duration, intensity and f0 alternations on rhythmic grouping. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 245-248.

HILL, D. R. / JASSEM, .W / 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 DER HULST, H. / RITTER, N. (eds.) (1999): The Syllable: Views and Facts. Berlin.

JASSEM, W. (1952): Stress in Modern English. In: Bulletin de la Société Linguistique Polonaise XII, pp. 189-194.

JASSEM, W. / HILL, D. R. / WITTEN, I. H. (1984): Isochrony in English speech: its statistical validity and linguistic relevance. In: GIBBON, D. / RICHTER, H. (eds.): Intonation, Accent and Rhythm, Berlin, pp. 203-225.

KAHN, D. (1976): Syllable-based generalizations in English phonology. PhD dissertation MIT. [Also New York 1980].

KELLER, B. Z. / KELLER, E. (2002): Representing speech rhythm. In: KELLER, E. / BAILLY, G. / MONAGHAN, A. / TERKEN, J. / HUCKVALE, M. (eds.): Improvements in Speech Synthesis. Chichester, pp. 154-164.

KELSO, J.A.S (1995): Dynamic Patterns. Cambridge, MA.

KLATT, D. (1975): Vowel lengthening is syntactically determined in a connected discourse. In: Journal of Phonetics 3, pp. 129-140.

KLATT, D. (1979): Synthesis by rule of segmental durations in English sentences. In LINDBLOM, B. / OHMAN, S. (eds.): Frontiers of Speech Communications Research. New York.

LEHISTE, I. (1977): Isochrony reconsidered. In: Journal of Phonetics 5, pp. 253–263.

LEWIS, E. / TATHAM, M. (1991): SPRUCE - a new text-to-speech synthesis system. In: Proceedings of EuroSpeech '91, Genova, pp. 976-981.

LEWIS, E. / TATHAM, M. (2001): Automatic segmentation of recorded speech into syllables for speech synthesis. In: Proceedings of EuroSpeech '01, Aalborg, pp. 1703-1707.

LOCAL, J. / OGDEN, R. (1997): A model of timing for nonsegmental phonological structure. In: VAN SANTEN, J. / SPROAT, R. / OLIVE, J. / HIRSHBERG, J. (eds.): Progress in Speech Synthesis. New York, pp. 109-122.

MORAIS, E. / TAYLOR, P. / VIOLARO, F. (2000): Concatenative text-to-speech synthesis based on prototype waveform interpolation (a time frequency approach). In: Proceedings of the International Conference on Spoken Language Processing, Beijing. (CD-ROM).

MORTON, K. (1992): Pragmatic phonetics. In: AINSWORTH, W.A. (ed.): Advances in Speech, Hearing and Language Processing, London, pp. 17-55.

MORTON, K. / TATHAM, M. (1995): Pragmatic effects in speech synthesis. In: Proceedings of EuroSpeech '95, Madrid, pp. 1819-1822.

MORTON, K. / TATHAM, M. / LEWIS, E. (1999): A New Intonation Model for Text-to-Speech Synthesis. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 85-88.

OGDEN, R. / HAWKINS, S. / HOUSE, J. / HUCKVALE, M. / LOCAL, J. / CARTER, P. / DANKOVIČOVÁ, J. / HEID, S. (2000): ProSynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis. In: Computer Speech and Language 14, pp. 177-210.

OGDEN, R. / LOCAL, J. / CARTER, P. (1999): Temporal interpretation in ProSynth, a prosodic speech synthesis system. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 1059-1062.

PATEL, A. / LÖFQVIST, A. / NAITO, W. (1999): The acoustics and kinematics of regularly timed speech: a database and method for the study of the p-center problem. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, pp. 405-408.

PETERSON, G.E. / SHOUP, J.E. (1966): An acoustic theory of phonetics. In: Journal of Speech and Hearing Research 9, pp. 5-67.

PIKE, K. (1945): Intonation of American English. Ann Arbor.

TAMS, A. / TATHAM, M. (1995): Describing speech styles using prosody. In: Proceedings of EuroSpeech '95, Madrid, pp. 2081-2084.

TATHAM, M. (1986): Towards a cognitive phonetics. In: Journal of Phonetics 12, pp. 37-47.

TATHAM, M. (1990): Cognitive phonetics. In: AINSWORTH, W.A. (ed.): Advances in Speech, Hearing and Language Processing 1, London, pp. 193-218.

TATHAM, M. (1994): The supervision of speech production: an issue in speech theory. In: R. LAWRENCE (ed.): Proceedings of the Institute of Acoustics 16, St. Albans, pp. 171-182.

TATHAM, M. (1995): The supervision of speech production. In SORIN, C. / MARIANI, J. / MELONI, H. / SCHOENTGEN, J. (eds.): Levels in Speech Communication – Relations and Interactions. Amsterdam, pp. 115-125.

TATHAM, M. / LEWIS, E. (1992): Prosodic assignment in SPRUCE text-to-speech synthesis. In: LAWRENCE, R. (ed.): Proceedings of the Institute of Acoustics 14, St. Albans, pp. 447-454.

TATHAM, M. / LEWIS, E. (1999): Syllable reconstruction in concatenated waveform speech synthesis. In: Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, vol. 3, pp. 2303-2306.

TATHAM, M. / MORTON, K. (1980): Precision. Occasional Papers, Department of Language and Linguistics, Essex University, 23, pp. 107-116.

TATHAM, M. / MORTON, K. (1995): Speech synthesis in dialogue systems. In: DALSGAARD, P. (ed.): Spoken Dialogue Systems. Visgo, pp. 221-225.

TATHAM, M. / MORTON, K. (in preparation): Speaking rhythm in Montréal French.

TATHAM, M. / MORTON, K. / LEWIS, E. (2000): SPRUCE: speech synthesis for dialogue systems. In: TAYLOR, M. M. / NÉEL, F. / BOUWHUIS, D. G. (eds.): The Structure of Multimodal Dialogue II. Amsterdam, pp. 271-292.

TATHAM, M. / MORTON, K. / LEWIS, E. (2001): Re-engineering intonation in the synthesis of prosody. Proceedings of the Institute of Acoustics – WISP 2001, St Albans, pp. 219-229.

WICKELGREN, W. (1969): Context-sensitive coding, associative memory, and serial order in (speech) behavior. In: Psychological Review 76, pp. 1-15.

WILLIAMS, B. / HILLER, S. M (1994): The question of randomness in English foot timing: a control experiment. In: Journal of Phonetics 22, pp. 423–439.