Introducing Natural Sounding Rhythm into Synthetic Speech

Mark Tatham
Katherine Morton

Reproduced from *Proceedings of the Institute of Acoustics* (2002). St. Albans: Institute of Acoustics.

Introduction – the lack of naturalness in synthetic speech

There is an increasing need for speech generated by computer [synthetic speech] in many areas of communication between people and machines. A good example is an automated booking system over the telephone. However, although in some respects modern speech synthesis is very good it still has an overall unnatural quality which can be related not to the individual sounds of the speech but to how the overall utterances sound. One of the parameters of speech which contributes to a sense of naturalness is the way stressing and rhythm are related and it is in this kind of area that we lack an adequate model of the overall process of producing and perceiving speech.

Rhythm in human speech

All speech has rhythm: speakers and listeners alike feel the rhythm and can report it. We can see this most clearly in poetry where we feel there is a pattern to the way syllables are expressed. In fact it can be recognised and reproduced when written as the following general way: *di dah, di dah, di dah* or *dah diddy, dah diddy dah*. And most speakers of English will be able to produce these two sequences of sounds.

It would seem to follow that since we are aware of speech rhythm, we should be able to establish its acoustic correlates – we should be able to find in the acoustic signal what it is that is conveying the sense of rhythm. However, although there are many apparent regularities in the acoustic signal, it has not been possible to locate a patterned temporal occurrence of events or to establish a clear relationship between what is perceived and what occurs physically in the signal. In human speech there is a great deal of variability anyway, but even allowing for natural irregularities, tracking down rhythm in the acoustic signal has proved curiously elusive – even though everyone agrees it's there.

One perceived property of speech rhythm is called **isochrony**. This is an apparent regular temporal spacing of stressed syllables. A stressed syllable is any syllable which is perceived as being more prominent than adjacent syllables. For example, a word such as *Institute* has three syllables: in – sti – tute, and people report that they feel that the first syllable is more important or is more prominent: ***in*** – sti – tute. We are so used to stress patterns like this that it's actually quite difficult to say words with the 'wrong' pattern. Isochrony, then, is the term used to describe the perceptual effect of feeling that there is the same temporal distance between stressed syllables. For example: 'Happy New Year' contains 4 syllables, two of them are stressed. English speakers will say ***Hap*** – py – New – ***Year***, but Americans will say ***Hap*** – py – ***New*** – Year. But in either case we feel that the prominent syllables occur equidistantly in time. Formally we say that isochrony is the perception that utterances have temporal events which fall equidistantly in time.

Languages do differ when it comes to rhythm patterns. Some, like English, have this patterning of stressed and unstressed syllables, while others, like French, have a series of stressed ones only. But isochrony is independent of whether we're dealing with a sequence of just stressed syllables, or stress and unstressed syllables. For example, in English we have 'My – ***me*** – ssa – ges – are – ***ve*** – ry – ***in*** – tres – ting' – just two stressed syllables, but in

French they're all stressed: 'Mes – me – ssages – sont – trè – s in – té – re – ssants'. In both languages, the perception of isochrony is that of some equivalent distance in time between the stressed syllables, even if *all* syllables are stressed.

The perception of isochrony in speech

In trying to model what's going on we need to take into account the speaker as well as the listener. Speakers report that they feel they produce speech in an isochronic way. If so, the inference here is that it is reasonable to suggest that speakers actually *plan* isochrony. Planning in speech production is a very important cognitive process, and there's every reason to believe that phenomena like stressing and rhythm is just as planned as the actual sounds which make up the words.

How can we relate perception and production of speech to isochrony? We suggest the goal of planning in human speech production is to trigger particular percepts in the listener. The basis for this suggestion is that listeners report a clear perception of isochrony – so the speaker must have triggered this perception. For the listener the rhythm is patterned in a predictable way; it is not random. So, we conclude that speakers can create from the plan a suitably patterned acoustic signal which will trigger this perception. This sounds complicated, but all it means is that if listeners consistently report predictable rhythms in speech they must be responding to some temporal pattern in the acoustic signal produced by the speaker. And it is this acoustic pattern which is proving so difficult to find objectively.

Redefining the goals of synthetic speech

The goal in speech synthesis, as in human speech, is to elicit particular percepts in the listener, especially the impression that *human* speech is being perceived. In synthesis research, there are two different approaches. We can either

- produce an acoustic signal which is acoustically as close as possible to the signal produced by a human speaker, or
- produce an acoustic signal which optimally triggers the desired percept in the listener.

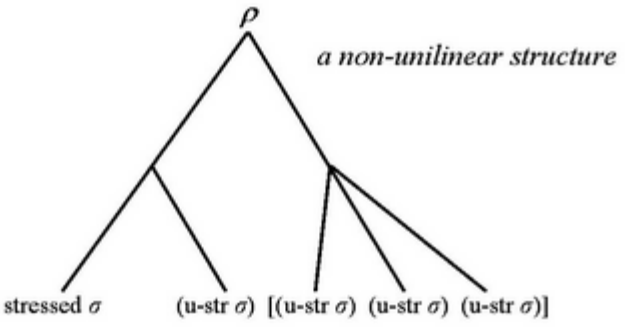
In our own work we take the second approach. This means that if the synthesiser conveys the impression of being good, then it *is* good, irrespective of the measurable objective quality of the audio. The advantage of this approach is that not all the details of the signal need to be reproduced; and not all of the details of the signal need to be described and modelled. Listeners have a remarkable ability to take what they hear, rapidly detect any errors and then repair those errors to their best advantage. We tap into this ability by contriving a synthesised acoustic signal which maximises the use of this ability.

The rhythmic unit

In our synthesis of rhythm, the concepts we have to work with are:

- ***rhythm:*** the patterned temporal occurrence of rhythmic units;
- the ***rhythmic unit:*** the time from the start of a stressed syllable to the start of the next stressed syllable;
- ***isochrony:*** the perceived and reported temporal occurrence of the rhythmic units.

Our task in synthesis of English is to produce speech characterised by a series of rhythmic units which will be perceived as isochronous.



$$\rho \rightarrow [\, s\text{-}\sigma + (u\text{-}\sigma)\,] + [\, u\text{-}\sigma_0 \ldots u\text{-}\sigma_3\,]$$

The overall formal structure of the rhythmic unit in English is shown in Figure 1. The rhythmic unit (rho) supports one syllable (sigma) which is stressed and the possibility of up to three or four following unstressed syllables. A hierarchal representation is used, rather than the simpler linear representation because it shows the underlying relationships between the surface elements (those at the bottom of the tree). For example, the sentence: 'Analysis of speech can be a frustrating activity' is not described as a simple series of stressed and unstressed syllables, but rather as a series of timed intervals – the rhythmic units – each representing a time unit that will correspond to the *perception* of rhythm. Note that the stressed intervals do not necessarily correspond to individual words in English: they can quite happily span word boundaries. This is the basis of a non-linear predictive model for synthesising rhythm.

Experimenting with rhythm

We conducted an experiment to confirm the findings of others that acoustic correlates of isochrony could not be established but that there might be a positive correlation between the duration of a rhythmic unit and the number of unstressed syllables it contains. Although other researchers had done similar work, the difference was that we approached the evaluation of the data from a slightly different point of view. That point of view focused on the temporal relationships within the abstract hierarchical model presented above.

Five newspaper articles were read, the utterances were segmented into rhythmic units, and durations measured of all syllable units were measured. Three hypotheses were tested, each formulated as a null hypothesis – one which can be rejected with confidence if the data falls out right:

1. Any pattern of the rhythmic units is isochronic. [In line with other researchers (e.g. Lehiste 1977) we expected to find *no* objectively measurable isochrony.]
2. There is no correlation between the duration of a rhythmic unit and the number of unstressed syllable. [In line with Jassem and Hill (1984) we expected to find a significant correlation between the duration of a rhythmic unit and the number of unstressed syllable in that unit.]
3. There is no trend for rhythmic units to increase in duration before syntactic boundaries, such as end of phrase, or end of sentence. [We expected to find durational increases in rhythmic units before pauses and sentence end.]

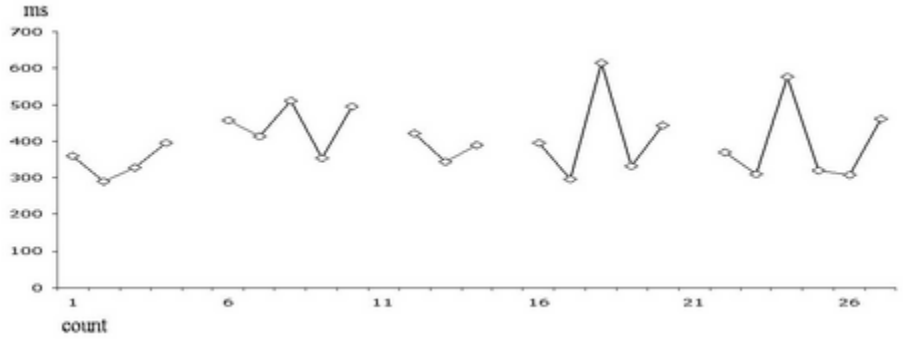


Figure 2 graphs a sample of the measured data – this is taken from one of the read passages. The horizontal axis shows that 27 rhythm units were measured, and the vertical axis gives the measurements for each. The gaps in the graph indicate pauses in the speaker's delivery; for example, at the end of a sentence or phrase. Note that before each of these pauses the final rhythm group is longer than the previous one, giving a slowing down effect. This is quite a normal pattern.

Modelling the details of rhythm

Table I shows the relationship between the duration of rhythmic units and the numbers of syllables in each unit. There is a clear correlation between length and syllable number. Variability is fairly even. The commonest unit was the type with one stressed syllable followed by a single unstressed syllable. This particular unit type, stressed syllable + single unstressed syllable, was taken as the ***basic rhythm unit*** and assigned a length *L*. Other types of unit relate to this basic unit according to the length ratio determined from the data.

article 2							
	mean	*median*	*standard deviation*	*min*	*max*	*number of examples*	*coefficient of variation*
stressed	355	366	112	177	673	63	31.5
stressed + (1 x u-stressed)	437	432	126	183	768	119	28.8
stressed + (2 x u-stressed)	497	488	220	267	781	74	22.2
stressed + (3 x u-stressed)	594	590	69	480	702	11	12

TABLE I Rhythmic unit durations in ms

Units before a pause or sentence end need to be lengthened. The model takes the next highest value in the ratio for these. So, if the last rhythm unit in a block as a stressed syllable followed by a single unstressed syllable, its duration is that normally assigned to one with two stressed syllables. The skeletal computational model is as follows:

```
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].

Testing the rhythm model

Testing the model produced a more acceptable perceived result in rhythm. The synthetic rhythm was not isochronic, but conformed to a more complex model as shown in Diagram page 44. We can conclude, pending further work, that a rhythmic pattern *can* be determined, but it would appear that objectively there are *no* equal durations between stressed syllables, although there *are* predictable ratios of stressed and unstressed patterning. The synthesised pattern produced a *perceived* isochrony in much the same way as real speech does. We tentatively assume that speakers plan this patterning and that listeners use it to reconstruct the percept of isochrony.

Conclusion

We are trying to improve the naturalness of synthetic speech to make it more acceptable. Advances in the general synthesis model means that the sounds which make up utterances are of excellent quality, but there nevertheless remains a very artificial quality to such things as stress and rhythm. Rather than try to base our synthesis on the idea that good synthetic speech produces a waveform identical to that produced by a human being we have adopted an alternative model which claims that good synthetic speech is one which makes the listener *believe* they are hearing natural human speech by pulling into the overall model the fact that listeners are able to repair damaged speech. We are aiming to trigger this listener ability to our advantage to make up for gaps in our understanding of the precise nature of the acoustic signal generated by human beings.

References

Jassem, W., Hill, D.R. and Whitten, 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

Lehiste, I. (1977) Isochrony reconsidered. *Journal of Phonetics* 5, 253-263