*Artificial Intelligence Review* **10**: 253–273, 1996.


# Visualization of Rhythm, Time and Metre

NEIL P. McANGUS TODD[1] and GUY J. BROWN[2]

[1] *Department of Music, University of Sheffield, Sheffield S10 2TN, U.K. (Current address: Department of Psychology, University of Manchester, Manchester M13 9PL, U.K.; E-mail: todd@hera.psy.man.ac.uk);* [2] *Department of Computer Science, University of Sheffield, Sheffield S10 STN, U.K.; E-mail: g.brown@dcs.shef.ac.uk*

**Abstract.** Developments in the theory of auditory processing of rhythmic signals have enabled the construction of a robust algorithm for recovery of rhythmic grouping structure. This algorithm appears to be effective for both speech and music signals. The theory upon which the algorithm was based was inspired by the theory of edge detection in vision. The output of the algorithm can be visualised in the form of a "rhythmogram", examples of which are shown for a variety of speech signals. The relationship between rhythm, time perception and metre is discussed in the light of a recent "auditory-motor" theory of beat induction.

**Key words:** rhythm, time perception, metre, beat induction, temporal grouping, prosody

## INTRODUCTION

Prosody plays a vital role in speech comprehension since it can assist in the segmentation of utterances into phrases and syllables, and in resolving syntactic or semantic ambiguities. A computational model for the recovery of prosodic structure from a speech signal would therefore be of great value for the goals of comprehension and synthesis. Such a model, in the form of a multi-scale analysis, has recently been developed (Todd 1994; Todd and Brown 1994). This model was inspired by the theory of edge detection in low-level vision (Marr 1982). The basic idea, in the case of vision, is that following retinal transduction the optical signal is blurred over a range of spatial scales by a number Gaussian low-pass filters. Edges are detected by looking for zero-crossings in the second derivative of the Gaussian which has a characteristic "Mexican hat" shape. Information from each of these spatial channels can be combined to form a higher representation.

In the case of hearing it is possible to do a similar computation on a simulation of the auditory nerve response. The essential difference is that the low-pass filtering is done in time rather than space. The result of such a computation can be represented in a number of ways. One way is to plot the output of each low-pass filter to form an "energy flux surface" in a three dimensional (energy, time, time-constant) space. Another way is to plot zero-crossing points of the derivatives of the low-pass response. This can be done either in the energy/time plane or the time-constant/time plane. Such a representation, referred to as a

"rhythmogram", in case of the time-constant/time projection, resembles the trees used in the phonological analysis of prosody (Liberman and Prince 1977; Selkirk 1984). In this paper we describe this algorithm and its relationship to a more general "auditory-motor" theory of rhythm and metre.

## I. RHYTHM AND METRE IN MUSIC AND SPEECH

It is generally agreed that rhythm in music can be described by two hierarchical structures – *grouping*, the organisation of events into well-defined sections, phrases, motifs, etc. and *metre*, a regularly repeating pattern of strong and weak beats. It is no accident that some theories of rhythm in speech (Geigerich 1985; Selkirk 1984) similarly require two complementary components (although there are some differences in the way they are defined), since they are based on the premise that speech rhythm is analogous to musical rhythm. According to Selkirk (1984) a phonological representation is formed by a *prosodic constituent structure*, which refers to the grouping of linguistic units such as the syllable into higher level units including the foot and phonological phrase, and a *metrical grid*, which represents a "hierarchy of temporal periodicities". (In practise though, metrical grids in speech are measures of relative stress, rather than periodicity.)

Whilst both grouping and metre are important for the description of rhythm, computational models of rhythm perception in the case of music have concentrated, almost exclusively, on the metrical component (Longuet-Higgins and Steedman 1971; Longuet-Higgins and Lee 1982, 1984; Lee 1991; Povel and Essens 1985; Desain 1992; Rosenthal 1992; Large 1994; Parncutt 1994). A further characteristic of all these models is that they start with a very abstract, symbolic representation of the musical signal, which has been already segmented into discrete events. One danger of this approach is to fall into the trap of believing that rhythmic beats are isochronous.

Another, more empirical approach has been to measure the subtle variations in timing during rhythmic performance. For example Sloboda (1983) has shown that musicians communicate metre by making tones falling on intended strong beats, longer, louder and more legato. Further, it is well known in performance analysis (Seashore 1938; Shaffer 1981; Todd 1985, 1989, 1992; Clarke 1988; Repp 1990, 1992) that expressive performance involves large modulations of tempo and that the modulations themselves carry important information about structure. Most metrical models simply cannot deal with such large tempo variations (but see Todd (1995) for further discussion of expressive timing and dance music).

In some ways this division, between abstract models of metre induction and empirical analysis of music performance, parallels the historical division of speech science into phonology on the one hand, and phonetics on the other (Cutler and Ladd 1983). The phonological approach has been concerned with the place of prosody in linguistic structure and defines as prosodic "any phenomena that involve phonological organisation above the segment" (Cutler and Ladd 1983). The phonetic approach defines prosody in more physical terms, as "those

phenomena that involve the acoustic parameters of pitch, duration and intensity".

It has been realised for some time though, that a proper understanding of speech prosody requires an integration of the phonological and phonetic views (Kingston and Beckman 1990). Similarly in music research, attempts are being made to bring together the abstract and the more psychoacoustic approaches (Leman 1994). The model of rhythm perception advocated here falls into this category (Todd 1994a; Todd and Lee 1994).

## II. A MULTI-SCALE MODEL OF RHYTHM PERCEPTION

As discussed above the inspiration for the model was the work of David Marr and co-workers in the theory of low-level vision. Before describing the model then, the theory of vision is briefly reviewed.

### II.1. *Review of The Theory of Low-Level Vision*

In the theory of vision as proposed by Marr the first stage in the detection of edges involves blurring the image $I$ by convolution with a Gaussian low-pass filter $g$. A 2-D Gaussian function is given by

$$(1) \qquad g(r, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-r^2/2\sigma^2],$$

where $r$ is *radius* and $\sigma$ is a *space-constant*. Since it is not possible to capture all information about intensity changes in a single convolution operation this blurring or smoothing process is carried out over a range of scales. In other words the image is convolved with a number of Gaussian functions with a range of space-constants $\{\sigma_k | k = 0 \ldots N - 1\}$.

The next stage in edge detection involves differentiating the smoothed versions of the image and looking for either peaks in the first derivative or zero-crossings of the second derivative which in the 2-D case is given by the Laplacian operator $\nabla^2$. Since this is a linear operation it can be taken inside the convolution so that effectively the image is convolved with

$$(2) \qquad \nabla^2 g(r) = \frac{-1}{\pi\sigma^4}\left(1 - \frac{r^2}{2\sigma^2}\right) \exp[-r^2/2\sigma^2],$$

which has a characteristic "Mexican hat" shape.

A symbolic description of the image, which Marr refers to as the *primal sketch*, could then be parsed from sets of zero-crossings from different channels over a range of scales. The actual rules for combining different channels are quite complicated. However, the basic principle is that if different channels agree on the spatial position of a zero-crossing segment then they can be combined to form a description of a single phenomenon. This principle is referred to as the *spatial coincidence* assumption.

### II.2. *A Temporal Analog to Low-Level Vision*

The simplest way of applying the visual analogy is to substitute *time t* for radius and *time-constant* $\tau$ for space-constant to obtain a function $g(t, \tau)$ which has the properties of a Gaussian low-pass impulse response. Then substitute *signal power* $|x(t)|^2$ for image intensity. In this way a smoothed or "windowed" measure of the *power density* $S(t)$ may be obtained by convolution of the impulse function with the signal energy, i.e.

$$(3) \qquad S_g(t) = g(t, \tau) * |x(t)|^2 = \int_0^\infty g(t', \tau)\, |x(t - t')|^2 \, dt'.$$

Like the visual case it is not possible to capture all information about energy changes in a single channel. This smoothing operation is therefore carried out over a range of time-constants $\{\tau_k | k = 0 \ldots N - 1\}$ and thus we obtain a measure of the *power density spectrum* $S_g(t, \tau_k)$. Finally, to complete the analogy, the direct temporal analogue of the $\nabla^2$ operator used for edge detection (Marr 1982), i.e. $d^2/dt^2$, can be used involving the condition

$$(4) \qquad \frac{d^2}{dt^2} S_g(t, \tau_k) = 0 \quad \text{and} \quad \frac{d}{dt} S_g(t, \tau_k) > 0.$$

The product of such an analysis, if carried out over a suitable range of time-constants, is a single hierarchical structure whose terminal elements are the onsets of individual events. As the time-constants are gradually increased, more and more detail is smoothed out so that eventually the sequence will be "reduced" to a single event.

There are two fundamental reasons why the simple analogue of visual processing cannot be a good model of auditory processing. The first is that the auditory system is causal. It is not possible to physically realise the Gaussian ideal as a causal system. The second reason is that the peripheral auditory system first carries out a frequency analysis via the basilar membrane followed by transduction by hair-cells which are subject to adaptation. A realistic model of rhythm perception must therefore take into account the effect of this transformation on any incoming signal (Figure 1).

### II.3. *A Multi-Scale Auditory Model*

In the first stage of the proposed model, the transfer function of the outer and middle ears is approximated by a simple high-pass filter. The frequency selective properties of the basilar membrane are modelled by a bank of gammatone filters (Patterson *et al.* 1988) spaced according to the ERB-rate scale of Glasberg and Moore (1990). Each of the cochlear channels are then processed by the Meddis (1986) inner hair-cell model which gives the auditory nerve firing probability (Brown 1992; Brown and Cooke 1994).

In the second stage, the auditory nerve response is pooled across frequency and passed to a multi-scale Gaussian low-pass filter system. The Gaussian filters were based on an analogue polynomial approximation to the Gaussian ideal

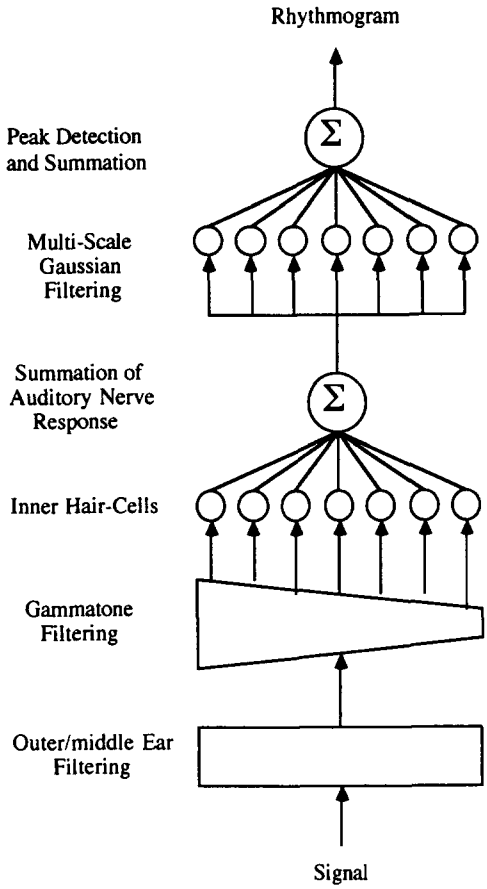


*Fig. 1.* The "rhythmogram" algorithm.

(Dishal 1957). These were then discretized by using a bilinear transformation to form an efficient digital IIR form. It turns out that these polynomial approximations are quite close to the gamma distribution

$$g(\tau;\, t,\, n) = \frac{t^{n-1}}{(n-1)\tau^n} \exp[-t/\tau], \tag{5}$$

where $n$ is the order of approximation. Thus, for example, when $n = 1$ the impulse response is just that of a 1st order low-pass filter or leaky integrator. Figure 2 shows the response of a bank of 8th order Gaussian approximation filters to a 50 ms pulse.

The last stage of the model involves looking for peaks in the low-pass response, or zero-crossings of the 1st derivative of the low-pass response (Figure 3) (equation (5)), i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t} S_g(t;\, \tau_k,\, n) = 0 \quad \text{and} \quad \frac{\mathrm{d}^2}{\mathrm{d}t^2} S_g(t;\, \tau_k,\, n) < 0. \tag{6}$$

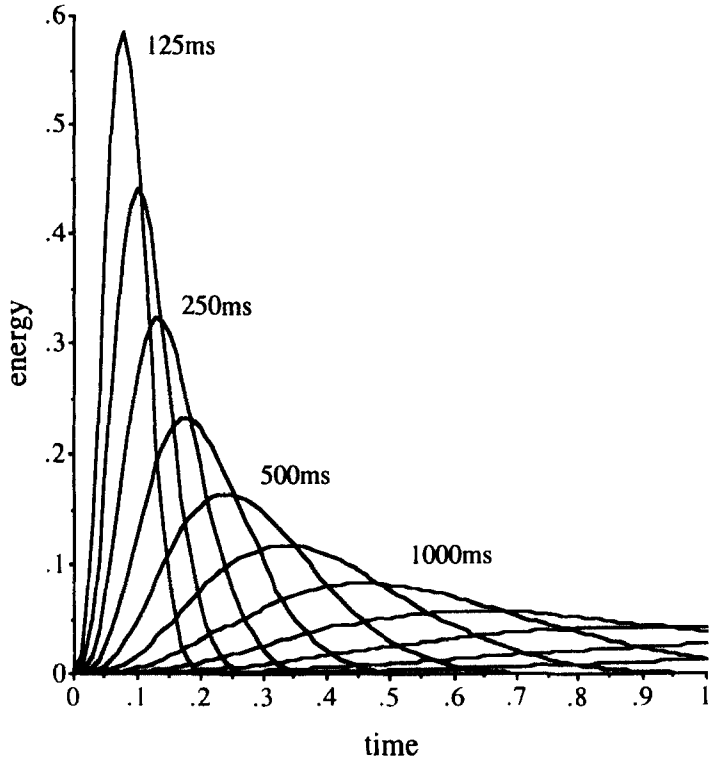


*Fig. 2.* The response of a Gaussain filter-bank to a 50 ms unit pulse. Each filter is an 8th order Taylor series approximation. The time-constants range from 125 to 400 ms. The sampling rate is 100 Hz.

When these points are plotted in a time-constant/time graph they give rise to a representation referred to as a rhythmogram, examples of which are shown in the next section.

It should be noted also that the rhythmograms obtained give a picture of the temporal grouping of identified segments. The actual boundaries between segments are obtained by looking for troughs, rather than peaks, in the sound energy flux. That is when

$$(7) \qquad \frac{\mathrm{d}}{\mathrm{d}t} S_g(t, \tau_k, n) = 0 \quad \text{and} \quad \frac{\mathrm{d}^2}{\mathrm{d}t^2} S_g(t, \tau_k, n) > 0,$$

over a range of time-constants $\{\tau_k | k = 0 \ldots N-1\}$. However, since the boundary structures have a one-one relationship with the temporal grouping structures they are omitted here (but see Todd (1994a) for examples).

One way of interpreting this mechanism is as a form of sensory memory. From this point of view the low-pass responses may be seen as representing the overall distribution of neural activity as one ascends the central auditory system. Thus, according to this view, very close to the periphery, after an input burst, the overall

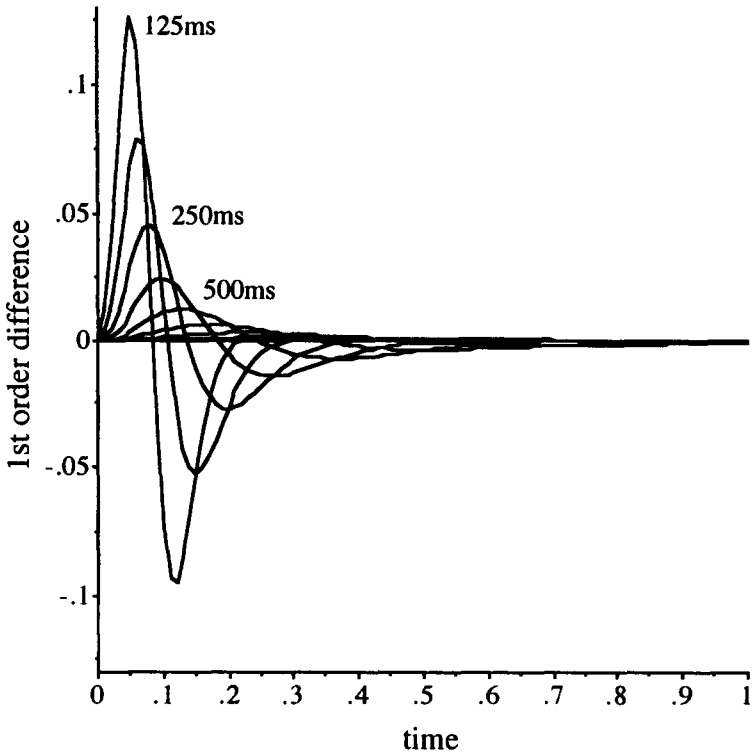


*Fig. 3.* The first derivative response of a Gaussain filter-bank as in Figure 2.

activity is very sharply defined in time. As activity spreads into the system though, the distribution becomes more diffuse. The increasing time-constant in this picture then, corresponds to the "distance" from the periphery at which a "snap-shot" is taken of the distribution of overall activity (Todd 1994b). In other words, the behaviour becomes more low-pass the further one ascends the system.

## III. RHYTHMOGRAM ANALYSIS OF PROSODIC STRUCTURE

In order to demonstrate the behaviour of the mutiscale mechanism in this section we look at a number of example analyses. The analyses are classified as large scale and small scale according to the time scale of interest. In the case of the large scale the smallest phonological constituent we are interested in is the phrase. For the small scale the smallest constituent is the syllable or phoneme.

### III.1. *Structure of Individual Words*

Since the work of Liberman and Prince (1977) it has become standard practise in metrical phonology to represent they rhythmic structure of an utterance in the form of a tree. In most cases these metrical trees are strictly binary so that

the relationship between a strong and a weak syllable may be either SW or WS (Figure 4 left). We refer to this representation as the binary tree (BT) approach. Lerdahl and Jackendoff (1983) have suggested an alternative representation refered to as a time-span reduction (TSR) (Figure 4 right). Although this was developed primarily to describe musical structure, Lerdahl and Jackendoff have suggested that it may also be useful for speech rhythm. In this section we show three examples of rhythmograms compared in each case to the Liberman and Prince and Lerdahl and Jackendoff representations.

According to the standard theory a core syllable (Sy) can be structured hierarchically into an onset (On) and a rhyme (Rh). The rhyme of a syllable further divides into the syllable peak (Pe) or nucleus and a coda (Co). The standard phonological approach represents this as in Figure 5a left (Giegerich 1985) whilst the time-span equivalent is shown in Figure 5a right.

Figure 5c shows a rhythmogram for the word "clamp". This has clearly segmented the onset, peak and coda. However, the rhythmogram looks more like the time-span reduction, but with one main essential difference. The outer stop consonants /k/ and /p/ have a greater prominence than /l/ and /m/ – a clear contradiction of the sonority principle.

This next example (Figure 6) looks at the tri-syllabic word "pa-me-la". The stress clearly falls on the first syllable. In this example the rhythmogram shows a striking resemblance to the time-span reduction. There is a very clear syllabification and the stressed syllable has been clearly identified.

The next level of phonological organisation above the syllable is the foot (Ft). The third example (Figure 7) is the word "re-con-ci-li-a-tion" which contains three feet according to conventional analysis. Each foot has two syllables the first of which is stressed.

Again we may see strong similarities between the rhythmogram and the time-span reduction but with two main differences. First, the particular speaker has divided the word into two feet of four and two syllables, rather than three feet of two syllables. Thus the two most prominent syllables are "re" and "a".

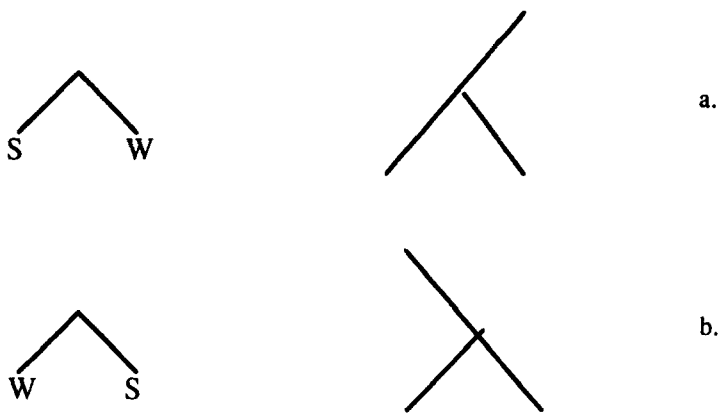


*Fig. 4.* Strong-weak vs weak-strong stress relationships. Liberman and Prince's representation (left) vs Lerdahl and Jackendoff's equivalent time-span representation (right).

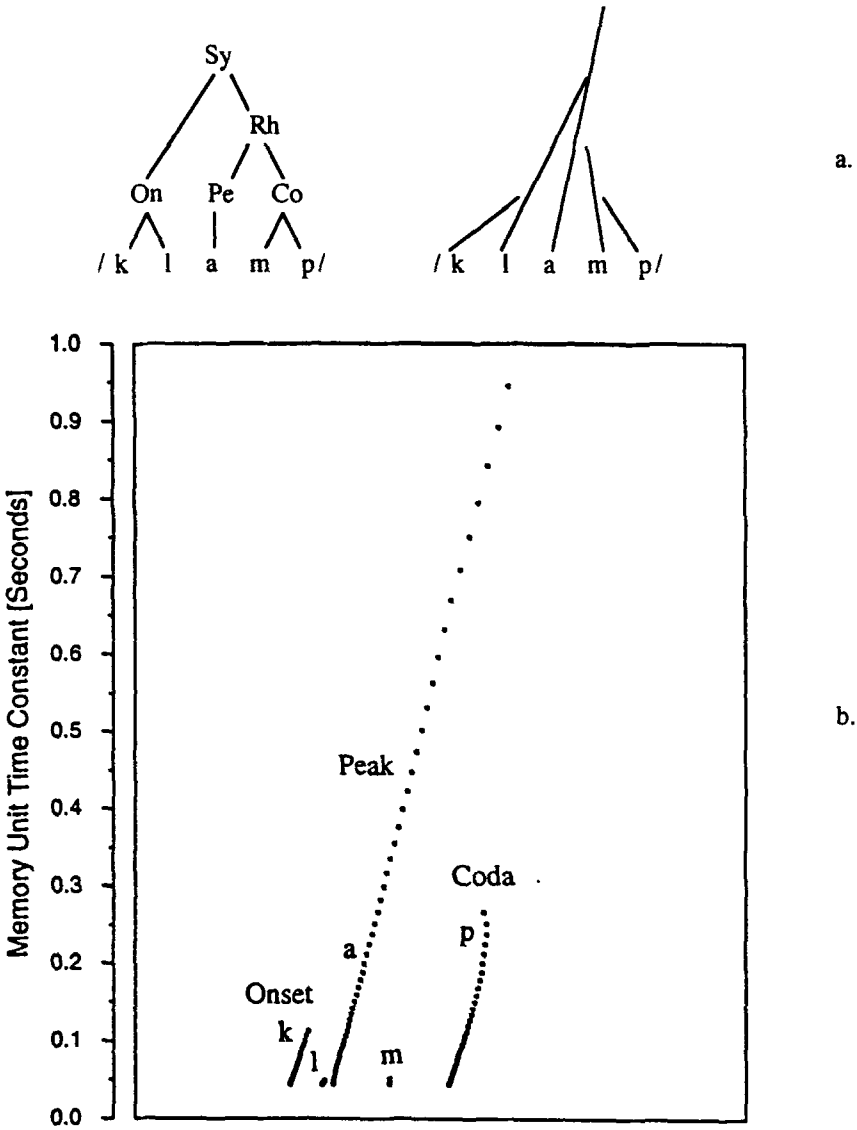


*Fig. 5.* (a) Conventional vs Lerdahl and Jackendoff representation of a syllable. (b) Rhythmogram of monosyllabic word "clamp".

Second, there are a number of individual phoneme segments which have the same, if not more prominence than some of the weak syllables, the last nasal /n/ is a clear example.

### III.2. *Structure of Complete Utterances*

The case of large scale phonology although in principle the same as small scale, in practise, because of the large time-scale, it is convenient to compute the

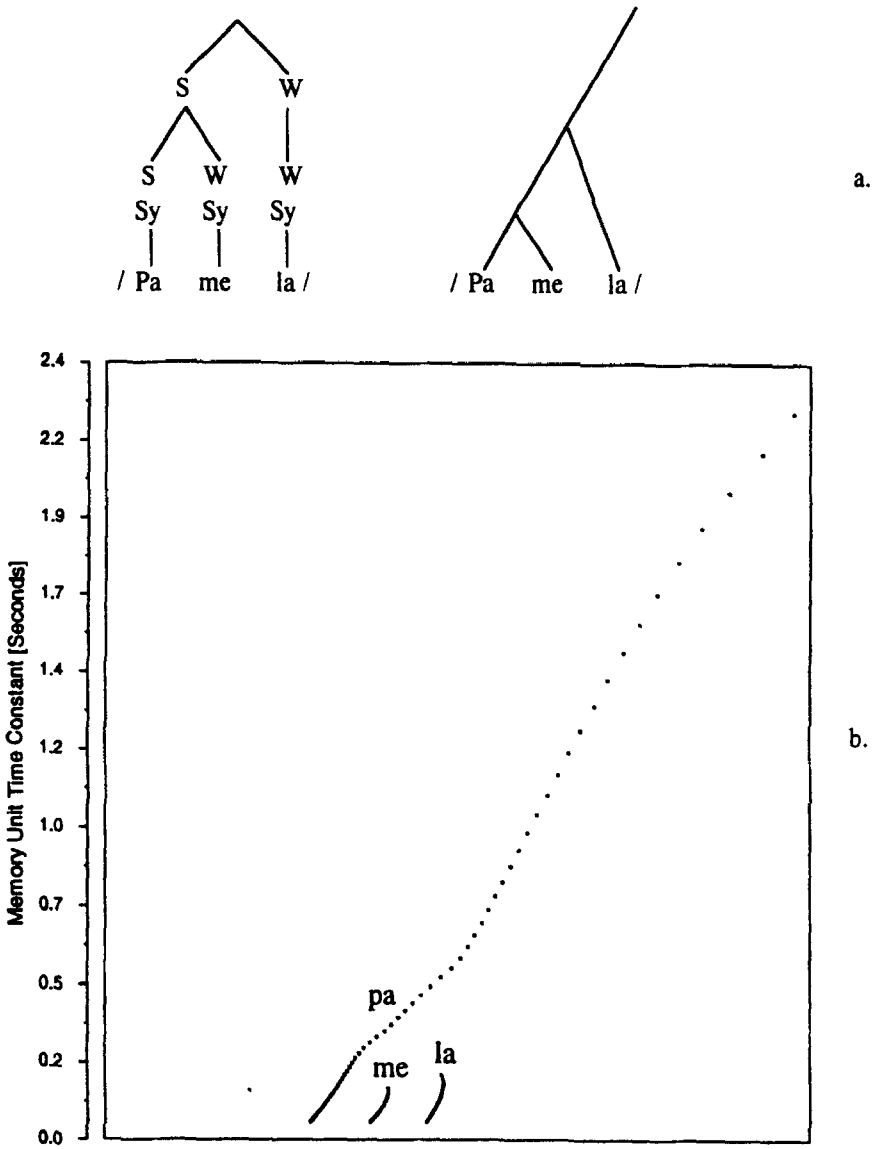


*Fig. 6.* (a) BT vs TSR for the tri-syllabic word "pamela". (b) Rhythmogram of word "pamela".

rhythmograms from the total energy rather than the nerve simulation. The input to the system is an analog signal $x(t)$ which is full-wave rectified $|x(t)|$ then integrated with a user-selectable cut-off frequency. Since rhythmic phenomena typically have a range of frequencies from 10–0.01 Hz this cut-off is also usually low, about 50 Hz. The output of the analog front-end is then sampled, forming the input to a digital filter-bank. Since the frequency content of the input signal is low we can sample at a low rate, typically about 100 Hz, thus saving massively on computation time and storage space.

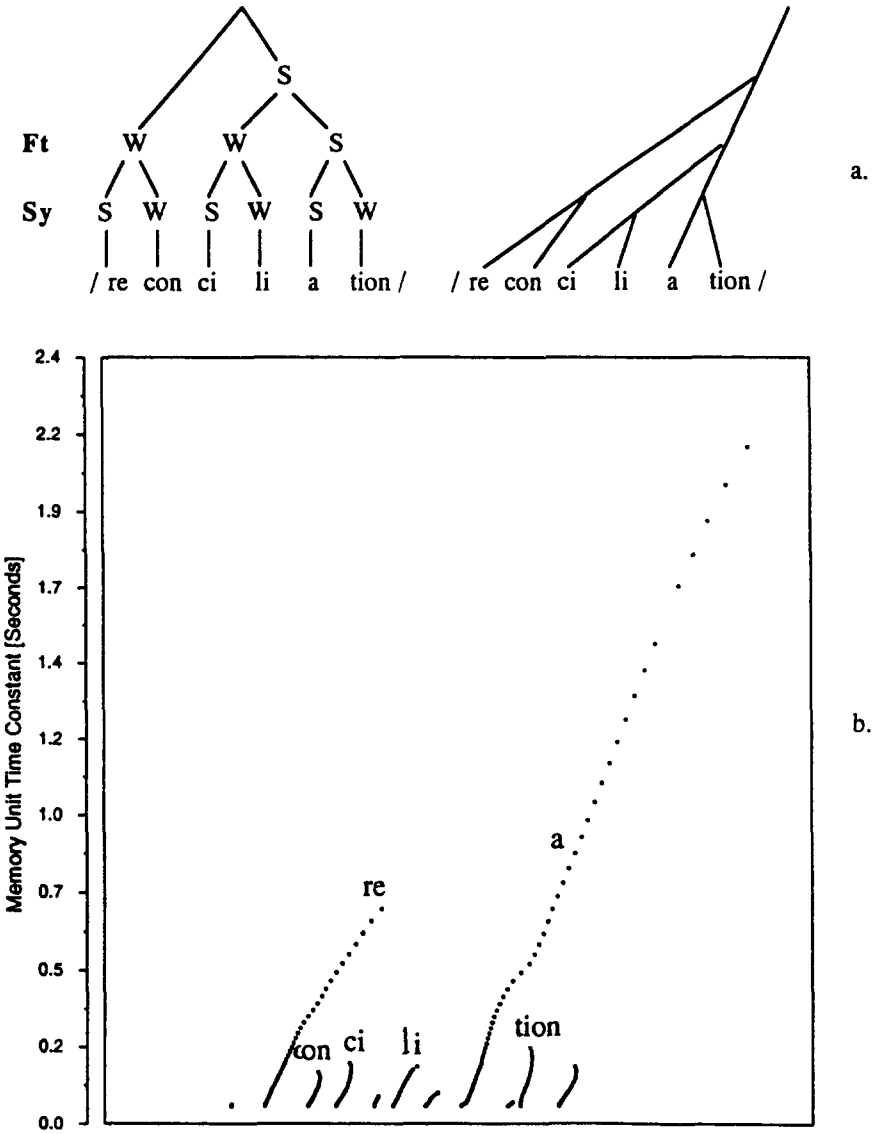


*Fig. 7.* (a) BT vs TSR for the multi-foot word "reconciliation". (b) Rhythmogram of "reconciliation".

In this section two examples are shown of large time-scale prosody. The first example is a recital of a poem by Thomas Hardy by a male professional actor, the second is a performance the Chopin Prelude Op. 28 No. 7 by a concert pianist.

*Hardy Poem "To the Moon"*

The poem (Hardy 1923) can be analysed to consist of four iambic anisometric stanzas of seven lines (Figure 8) which are denoted by $V_1$, $V_2$, $V_3$ and $V_4$. Each

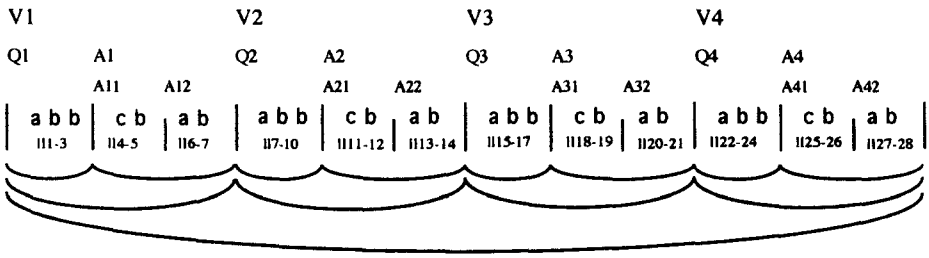


*Fig. 8.* A conventional metrical analysis of the Hardy Poem "To the Moon".

stanza is neatly sub-divided into a question and answer part which take three and four lines respectively, ($Q$ ? $A$). There are two rhyme patterns per stanza, $a$ and $b$, which occur in lines with an irregular number of stressed syllables or feet (denoted by superscript). The answer section seems also to divide naturally into two parts ($A_1$, $A_2$) with the $b^2$ forming a cadence. The rhymes are distributed as $Q \rightarrow (a^3, b^2, b^3)?$, $A_1 \rightarrow (c^4\ b^2)$, $A_2 \rightarrow (a^4, b^2)$. This subdivision is further strengthened by the fact that there is no syntactic break at the end of $c^4$.

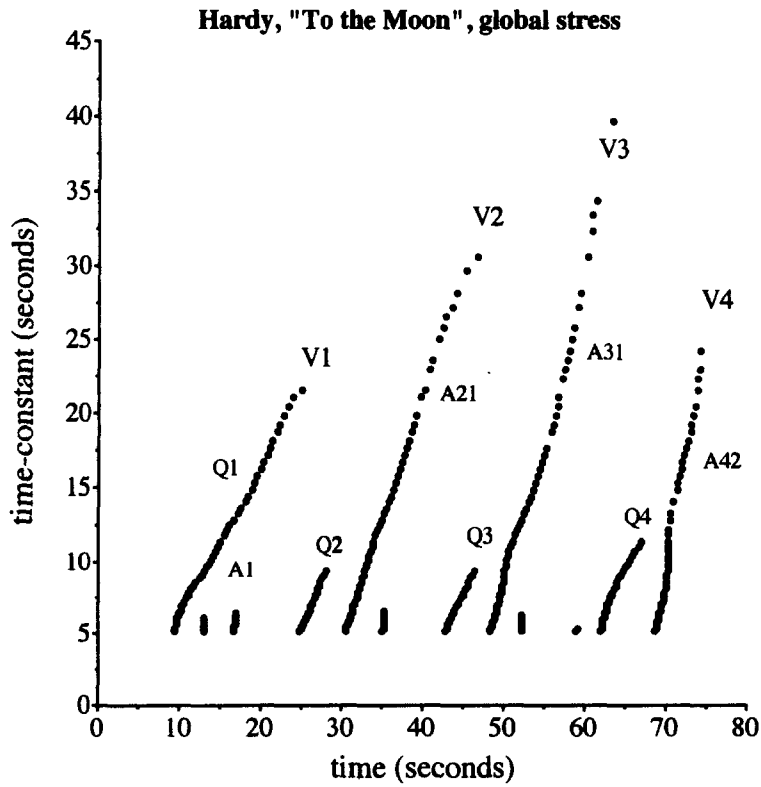


*Fig. 9.* Rhythmogram showing the high level prosodic structure from a performance of the Hardy Poem "To the Moon".

The rhythmogram has clearly resolved the global prosody of the poem. The rhythmogram also suggests a clear pattern of relative importance between the $Q$ and $A$ sections.

*Chopin Prelude Op. 28*

Figure 10 shows a conventional analysis of the Prelude Op. 28, No. 7. At the highest level it consists of two 8-bar sections $A_1$ and $A_2$. Each of the sections contain four 2-bar phrases $p_1$, $p_2$, $p_3$, $p_4$. In the first section there is one harmonic change per phrase consisting of V-I-V-I. Whilst the second section parallels the first $p_1$*, $p_2$*, $p_3$*, $p_4$*, the harmonic rhythm is doubled in the inner two phrases, $p_2$* and $p_3$*, to allow insertion of a "tonicization" of the supertonic so that the harmony consists of V-I (V/ii-ii) V-I (see Lerdahl and Jackendoff 1983).

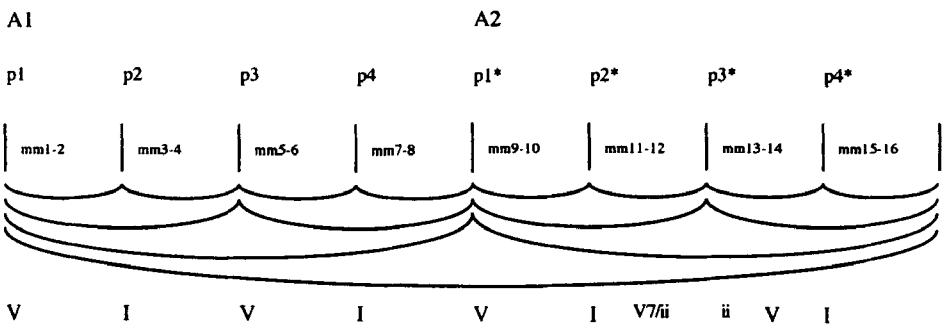


*Fig. 10.* A conventional musical analysis of the Chopin Prelude Op. 28 No. 7.

Figure 11 shows the rhythmogram structure obtained from a sample of a performance by a concert pianist. It can be seen that this structure neatly picks out the phrasal grouping structure. Further, the rhythmogram reflects the hierarchical organisation of the phrases in the Prelude and gives a measure of the relative importance of the phrases within the structure. According to this analysis the climax of the whole piece is the F-sharp dominant seventh chord (V7 of ii) in section $A_2$ phrase $p_2$*.

## IV. RHYTHM AND METRE

Although it is beyond the scope of this paper to properly cover the issue of metre, for the purpose of being able to see the previous work in relation to the complete theory of rhythm, we discuss the basic outline of the theory.

### IV.1. *An Auditory-Motor Theory*

Elsewhere an auditory-motor model (Todd and Lee 1994) has been proposed which has the following architecture (see Figure 12): (1) An *Auditory System*, which consists of (a) peripheral processing and (b) central processing in the form of band-pass and low-pass AM processes (including the multi-scale model

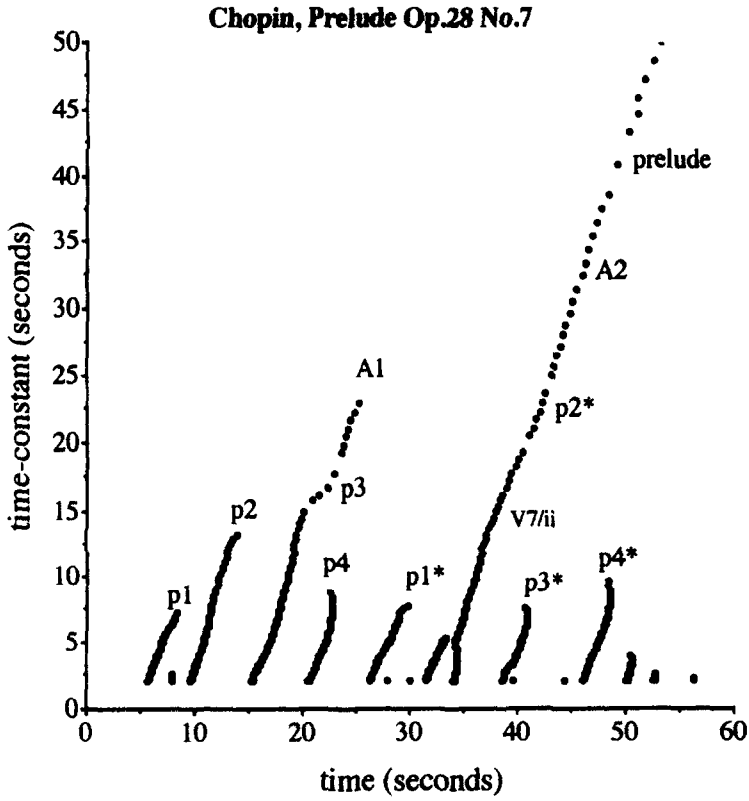


*Fig. 11.* Rhythmogram showing the hierarchical phrase structure from a performance of the Chopin Prelude Op. 28 No. 7.

above); (2) A *Motor System*, which consists of (a) sensory-motor filters representing the global dynamics of the motor system, (b) a central-pattern generator which is centrally programmed and (c) an output system; (3) An *Interpretation and Control System*, which both combines information from the various auditory processes and programs the central pattern generator.

### IV.2. *Auditory Processes*

It has now been realised that the multi-scale process described above can be seen as a component of a more general theory. The basic idea of the more general theory is that following transduction a number of populations of cells responsive to amplitude modulation (AM) at different levels of the auditory system extract information from temporal patterns in the auditory nerve response. A similar scheme has been proposed for auditory segregation (Yost and Sheft 1993). Neurophysiology suggests perhaps three or four levels, including the cochlear nucleus, inferior colliculus, thalamus and cortex – the cut-off frequency becoming lower the higher one ascends the system (Popper and Fay 1992). A priori, the uncertainty principle dictates that two types of AM processes must

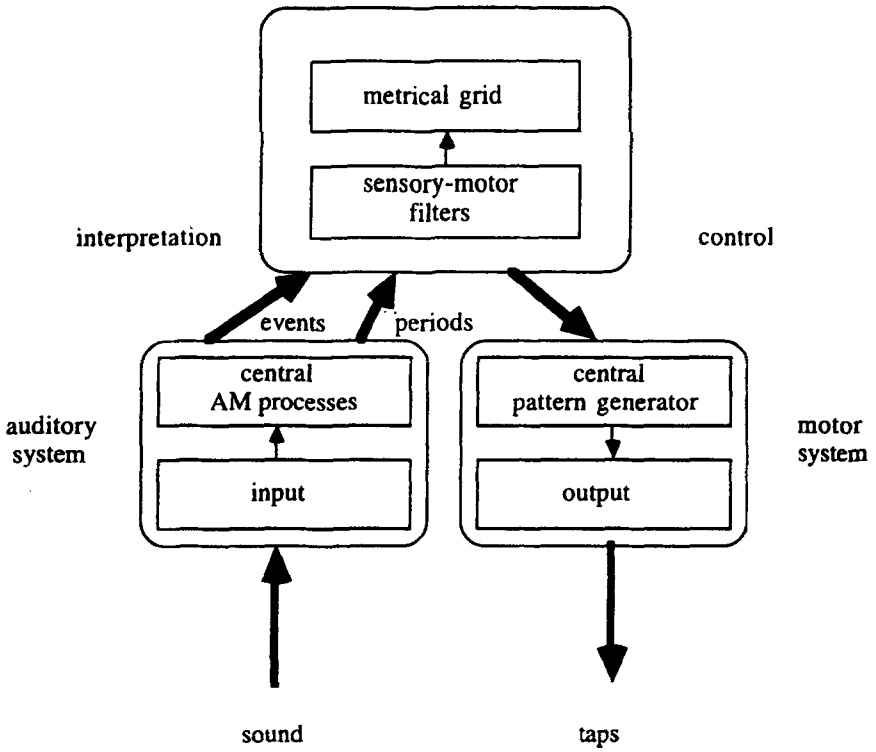


*Fig. 12.* An auditory-motor model of rhythm and metre.

exist, i.e. low-pass and band-pass, since it is not possible to simultaneously localise an event in time and obtain information on event periodicity. Given such a division, it is possible to classify the various psychophysical processes accordingly. Thus, on the one hand, phenomena such as temporal grouping (Todd 1994a) of multiple events or temporal integration (Todd 1994b) for single events can be viewed as the product of a low-pass process, whereas, on the other hand, phenomena such as periodicity pitch or tempo sensitivity can be seen as the product of a band-pass type process. Despite the apparent diversity of these phenomena, the general computational theory describing the behaviour of the underlying AM process may be seen in three stages for each case: (1) the nerve response is decomposed by an array of low/band-pass filters; (2) the decomposition is coded in terms of zero-crossings and has three dimensions (firing rate, AM frequency, time) so that three sorts of information are available – (a) phase (AM frequency, time), (b) amplitude (firing rate, time) and (c) spectrum (firing rate, AM frequency); (3) the zero-crossings are accumulated in a sensory memory which has a decay time-constant.

The multi-scale mechanism proposed earlier is just one example of a central low-pass AM process. This process provides information on onsets, loudness and grouping. However, it turns out that in order assign a metrical description to a rhythmic pattern it is necessary to compute the periodicity content of the

rhythm by means of a population of cells, assumed to be at the level of the auditory cortex, which behave as a bank of band-pass filters of AM. Similarly to the low-pass process the pooled nerve response is processed by a bank of band-pass filters of AM spaced at 24 per octave with a range 0.5 to 20 Hz. The impulse response of the filters are those of damped simple harmonic motion, i.e. exponentially decaying cosines, with a Q of about 3 so that the decay is quite rapid. The output of each filter is represented by only peak responses, i.e. it is assumed that at the neurophysiological level the response of a band-pass unit is in the form of a series of spike trains. The peak responses across the filter-bank are stored in a buffer which simulates a "sensory memory". Since the Q of the band-pass units is quite small the sensory memory also decays quite rapidly. Figure 13 shows the overall responses for a single time interval of 600 ms.

*Time Perception*

In addition to the neurophysiological evidence further support is given to this component of the theory since it able to account for tempo sensitivity (Drake and Botte 1993). Figure 14 shows the predicted sensitivity for single time intervals. According to this model the auditory system is maximally sensitive to intervals of between 300 and 600 ms because the fundamental period and its harmonics are maximally represented in a population of cortical units which

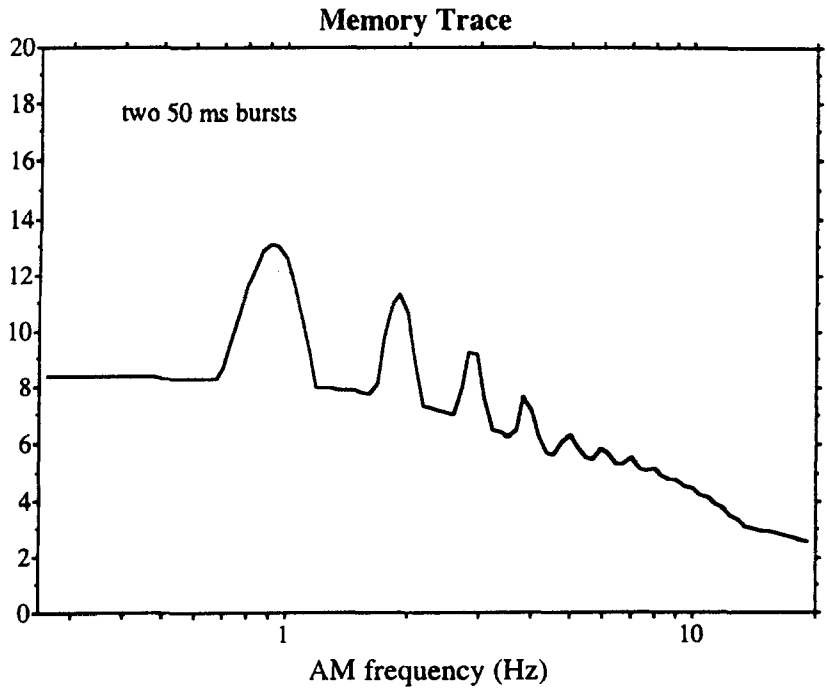


*Fig. 13.* The spectrum induced in the band-pass of AM process by a pair of 50 ms tone bursts with IOI 600 ms. Thus, a single (empty) time interval is represented in the form of a harmonic series.

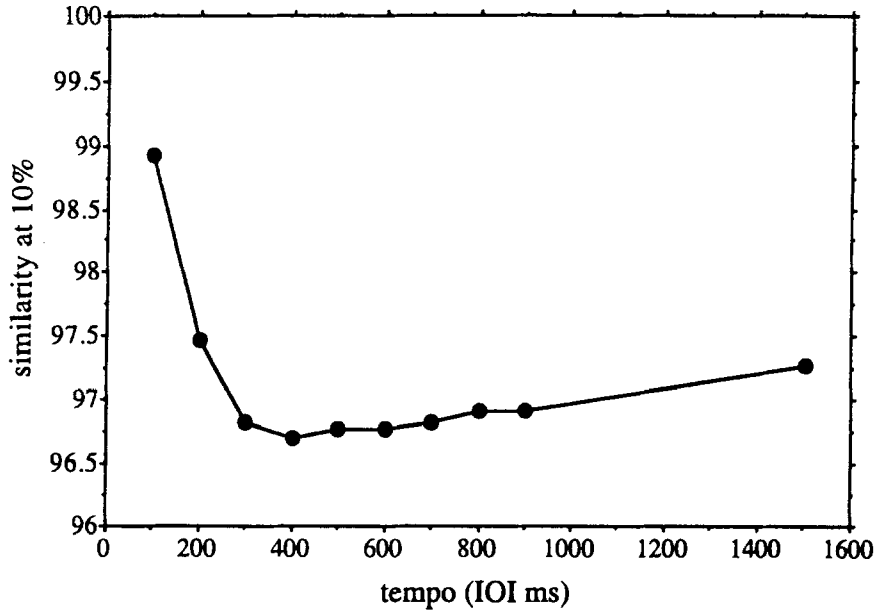


*Fig. 14.* The predicted tempo sensitivity curve for single intervals by the band-pass of AM process. Similarity is computed from the correlation coefficient between spectra induced by intervals which differ by 10%, e.g. 100 ms IOI and 110 ms IOI.

functionally behave as a bank of band-pass filters of AM whose range is approximately 0.5 to 20 Hz. For fundamental periods shorter than about 300 ms the harmonics become attenuated and for fundamental periods longer than about 1000 ms the fundamental becomes attenuated. Sensitivity increases with increasing interval number because the cortical units take time to build a resonance. However, sensitivity as a function of interval number reaches an asymptote quite quickly because the cortical units have a finite and small Q.

IV.3. *Motor Processes: Phase-locking a "Foot-tapper"*

After a short time for any fixed rhythm the auditory system will have produced both phasic (i.e. low-pass) and periodicity (i.e. band-pass) AM images. This enables an interpretation and control system to construct a motor programme for a foot-tapper which may be synchronised with stressed elements in the image. The basic procedure then, for metrical induction in the model, is the following: (1) set the tactus period equal to the strongest resonance in the band-pass sensory memory which is closest to the natural frequency of motor output; (2) lock the tactus phase to the most stressed event in the low-pass sensory memory within the time-span of the tactus period.

The natural frequency of the motor output system is represented in the form of a bio-mechanical filter which models the dynamics of the system during locomotion or foot tapping. It turns out that the sensitivity of the auditory system to tempo is coincident with the natural frequency of the motor system – perhaps a product of the coevolution of the auditory and motor systems.

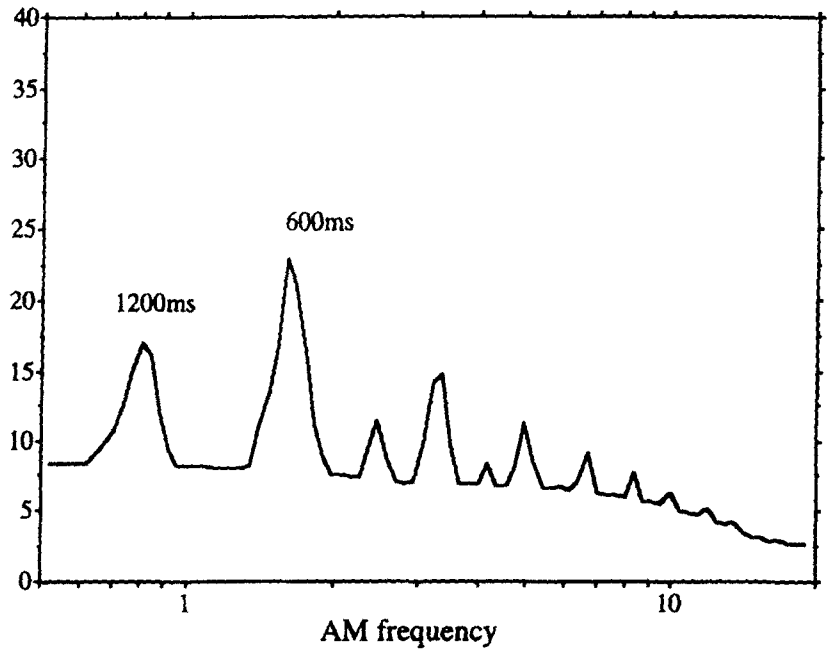


*Fig. 15.* The spectrum induced in the band-pass of AM process by an isochronous sequence of 50 ms tone bursts with a binary stressed/unstressed pattern.

As an example, the response of the model to a binary stressed-unstressed pattern is shown in Figures 15 and 16. Research is currently being undertaken to investigate the behaviour of the model to metrical verse recital.

## V. CONCLUSION

We have demonstrated, albeit for a small set of data, that the multi-scale mechanism provides a method for the visualisation of rhythm in the form of a "rhythmogram" at all levels of rhythmic structure, from individual phonemes to the structure of a complete poem. A number of important issues have been raised.

(1) Rhythmogram structures are not as simple as BTs. Are these differences purely to do with the individual variation of the speakers, or do they suggest a more fundamental difference, between an acoustically based representation on the one hand, and a more abstract schema based representation on the other?

(2) Whilst there are often striking differences with the BTs, the rhythmograms often show equally striking similarities with TSRs. This seems to suggest that Lerdahl and Jackendoff's approach is closer to the perceptual reality than the idealised structures of conventional metrical phonology. This view is strengthened by the fundamental similarity of large-scale musical and stanzaic grouping structure.

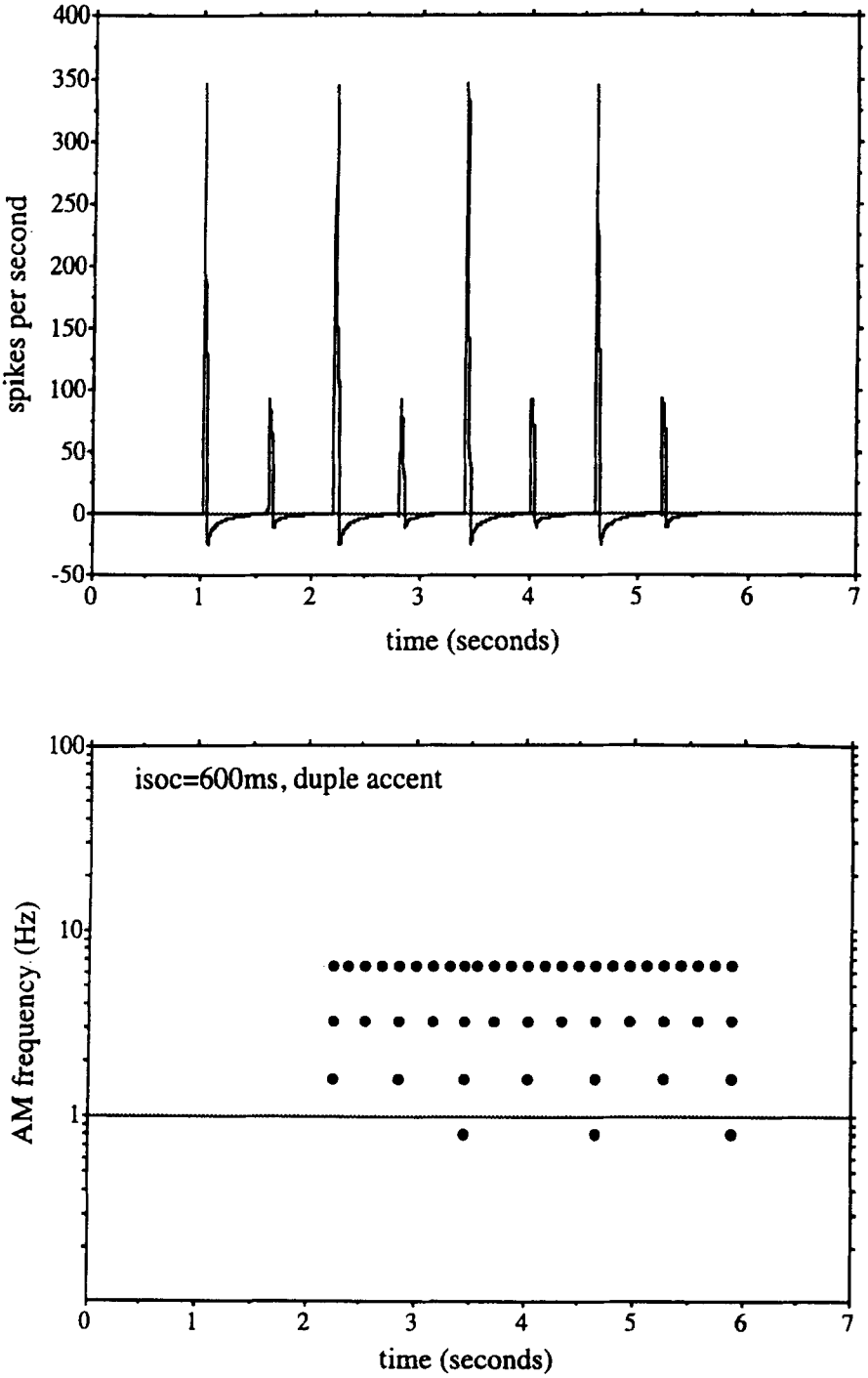


*Fig. 16.* Hair-cell input (top) vs the metrical-grid output (bottom) of the model for a binary stressed/unstressed pattern.

(3) A fundamental difference though with either BTs or TSRs is that the edges which make up the rhythmogram have a right slope – a consequence of the causality of the Gaussian filters – unlike the TSR which may have left and right sloping.
(4) There are many occasions when onset or coda consonant phonemes may have greater prominence in structure than an unstressed vowel, which is counter to the sonority principle. Does this suggest that the sonority theory may have to be revised? Or again, does this suggest that the acoustic level is an inappropriate level for the comparison with phonological principles.

It is too early yet to be able to say if such visualisations will be of benefit for the assistance of comprehension. However, given their high degree of similarity to conventional analyses the use of rhythmograms is surely an area worthy of further research. Finally, given that the starting point was the analogy to edge detection in vision, the success of the rhythmogram algorithm lends further support to the view that the approach advocated by Marr is entirely appropriate for perceptual computation in general.

ACKNOWLEDGEMENT

Research supported by MRC grant G9018013.

REFERENCES

Brown, G. J. (1992). *Computational Auditory Scene Analysis: A Representational Approach.* Ph.D. Thesis, University of Sheffield.

Brown, G. J. & Cooke, M. (1994). Perceptual Grouping of Musical Sounds: A Computational Model. *J. New Music Research.*

Clarke, E. (1988). Generative Principles in Musical Performance. In Sloboda, J. (ed.) *Generative Processes in Music: The Psychology of Performance, Improvisation and Composition.* Oxford: Claredon Press.

Cutler, A. & Ladd, D. R. (1983). *Prosody: Models and Measurement.* Springer-Verlag: Berlin.

Desain, P. (1992). A (De) Composable Theory of Rhythm Perception. *Music Perception* **9**: 439–454.

Dishal, M. (1959). Gaussian Response Filter Design. *Electrical Communications* **36**(1): 3–26.

Drake, C. & Botte, M. (1993). Tempo Sensitivity in Auditory Sequences: Evidence for a Multiple-Look Model. *Perception and Psychophysics* **54**(3): 277–286.

Geigerich, H. J. (1985). *Metrical Phonology and Phonological Structure.* Cambridge University Press: Cambridge.

Glasberg, B. & Moore, B. (1990). Derivation of Auditory Filter Shapes from Notched-Noise Data. *Hearing Research* **47**: 103–138.

Hardy, T. (1923). *Collected Poems of Thomas Hardy.* Vol. I. London: Macmillan.

Kingston, J. & Beckman, M. E. (1990). *Papers in Laboratory Phonology: Between the Grammar and Physics of Speech.* CUP: Cambridge.

Large, E. (1994). The Resonant Dynamics of Beat Tracking and Meter Perception. Proceedings of *The International Computer Music Conference.* Denmark: Aarhaus.

Lee, C. S. (1991). Perception of Metrical Structure: Experimental Evidence and a Model. In Howell, P., West, R. & Cross, I. (eds.) *Representing Musical Structure*, 59–127. London: Academic Press.

Leman, M. (1994). Introduction to Auditory Models in Music Research. *J. New Music Research* **23**(1): 5–9.

Lerdahl, F. & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. MIT Press: Cambridge, MA.

Liberman, M. & Prince, A. (1977). On Stress and Linguistic Rhythm. *Linguistic Inquiry* **8**(2): 249–336.

Longuet-Higgins, H. C. (1976). The Perception of Melodies. *Nature* **263**: 646–653.

Longuet-Higgins, H. C. & Lee, C. S. (1982). Perception of Musical Rhythms. *Perception* **11**: 115–128.

Longuet-Higgins, H. C. & Lee, C. S. (1984). The Rhythmic Interpretation of Monophonic Music. *Music Perception* **1**(4): 424–441.

Longuet-Higgins, H. C. & Steedman, M. J. (1971, 1987). On Interpreting Bach. In Longuet-Higgins, H. C. (ed.) *Mental Processes: Studies in Cognitive Science*, 82–104. MIT Press: Cambridge, MA.

Marr, D. (1982). *Vision*. Freeman: New York.

Meddis, R. (1988). Simulation of Auditory-Neural Transduction: Further Studies. *J. Acoust. Soc. Am* **83**(3): 1056–1063.

Parncutt, R. (1994). A Model of Beat Induction Accounting for Perceptual Ambiguity by Continuously Variable Parameters. Proceedings of *The International Computer Music Conference*. Denmark: Aarhaus.

Patterson, R, D. & Holdsworth, J. (1992). A Functional Model of Neural Activity Patterns and Auditory Images. In Ainsworth, W. A. (ed.) *Advances in Speech, Hearing and Language Processing*. Vol. 3. JAI Press: London.

Popper, A. N. & Fay, R. R. (1992). *The Mammalian Anditory Pathway: Neurophysiology*. Springer-Verlag: NY.

Povel, D. J. & Essens, P. (1985). Perception of Temporal Patterns. *Music Perception* **2**(4): 411–440.

Repp, B. (1990). Patterns of Expressive Timing in Performances of a Beethoven Minuet by Nineteen Famous Painists. *Journal of the Acoustical Society of America* **88**(2): 622–641.

Repp, B. (1992). Probing the Cognitive Representation of Musical Time: Structural Constraints on the Perception of Timing Perturbations. *Cognition* **44**: 241–281.

Rosenthal, D. (1992). *Machine Rhythm: Computer Emulation of Human Rhythm Perception*. MIT Media Lab. Ph.D Thesis.

Seashore, C. (1938). *The Psychology of Music*. McGraw-Hill: New York.

Selkirk, E. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press: Cambridge, MA.

Shaffer, H. (1981). Performances of Chopin, Bach and Bartok: Studies in Motor Programming. *Cognitive Psychology* **13**: 326–376.

Sloboda, J. (1983). The Communication of Musical Meter. *Quarterly Journal Of Experimental Psychology* **35**: 377–396.

Todd, N. P. (1985). A Model of Expressive Timing in Tonal Music. *Music Perception* **3**: 33–58.

Todd, N. P. McAngus (1989). Towards a Cognitive Theory of Expression: The Performance and Perception of Rubato. *Contemporary Music Review* **4**: 405–416.

Todd, N. P. McAngus (1992). The Dynamics of Dynamics: A Model of Musical Expression. *J. Acoust. Soc. Am* **91**(6): 3540–3550.

Todd, N. P. McAngus (1994a). The Auditory "primal sketch": A Multi-Scale Model of Rhythm Grouping. *J. New Music Research* **23**(1): 25–70.

Todd, N. P. McAngus (1994b). A New Theory of Temporal Integration. *British Journal of Audiology*.

Todd, N. P. McAngus (1995). The Kinematics of Musical Expression. *J. Acoust. Soc. Am* **97**(3), 1940–1950.

Todd, N. P. McAngus & Brown, G. (1994). A Multi-Scale Auditory Model of Prosodic Perception. Proceedings of *The International Conference on Spoken Language Processing*. Yokoyama, Japan.

Todd, N. P. McAngus & Lee, C. S. (1994). An Auditory-Motor Model of Beat Induction. Proceedings of *The International Computer Music Conference*. Denmark: Aarhaus.

Yost, W. A. & Sheft, S. (1993). Auditory Perception. In Yost, W., Popper, A. & Fay R. (eds.) *Human Psychophysics*, 193–236. Springer-Verlag: NY.