

A COMPUTATIONAL MODEL OF PROSODY PERCEPTION

Neil P. McAngus Todd¹ and Guy J. Brown²

¹Department of Music and ²Department of Computer Science, University of Sheffield
Sheffield S10 2TN, United Kingdom
INTERNET: N.Todd@shef.ac.uk, G.Brown@dcs.shef.ac.uk

ABSTRACT

This paper describes a computational model of auditory rhythm perception, and demonstrates its application to the extraction of prosodic information from spoken language. The model consists of three stages. In the first stage, the speech waveform is processed by a simulation of the auditory periphery. Secondly, the output of the auditory periphery is processed by a multiscale filtering mechanism, analogous to a short-term auditory memory. Finally, peaks in the response of the multiscale mechanism are accumulated in a long-term auditory store, and plotted to give a representation referred to as a *rhythmogram*. It is demonstrated that there is a close relationship between the rhythmogram of an utterance and its corresponding stress hierarchy derived by phonological analysis.

I. INTRODUCTION

Recently, the use of prosody has attracted attention as a means of improving the performance of systems for automatic understanding of spoken language. Prosodic information can assist in the segmentation of utterances into sentences and phrases, and in solving syntactic or semantic ambiguities. Additionally, it can assist in the management of dialogues between speakers and machines. For example, a spoken dialogue understanding system could use prosodic information to manage turn-taking during a telephone enquiry [7].

To realise this goal, computational techniques are required which are able to extract prosodic information from spoken language. This paper presents such a technique, in the form of a multiscale model of auditory rhythmic grouping. The model is proposed as a general theory of rhythm perception, and is not intended to be speech specific. Indeed, it has previously been applied to the rhythmic analysis of music with some success [8].

II. PROSODY AND SPEECH RHYTHM

The term *speech rhythm* refers to the perception of a regular ordering of stressed and unstressed speech sounds; this is equivalent to the notion of *metre* in music, although there are some differences in the way that these are defined [4]. In English, stress is correlated with the acoustical properties of speech syllables. Stressed syllables are produced

with a stronger burst of initiatory energy, resulting in an increase of perceived loudness, duration and pitch [4]. However, perceived syllable stress is context dependent, being relative to the variation in prominence of other speech sounds. Hence, as with musical rhythm, perceived speech rhythm is a joint function of stress assignments at many different levels, from the segment to the sentence.

Consequently, the notion of speech rhythm is intimately connected with *prosody*. This term refers to properties of the speech signal which span more than one segment. Prosody includes patterns of pitch, duration, loudness and other factors that affect the perception of stress and rhythm.

In the remainder of this paper, a model of auditory rhythm perception is presented, and it is argued that the model provides a mechanism for extracting prosodic information from spoken language. Indeed, evidence is presented that the rhythmic analysis performed by the model is sensitive to the main acoustic determinants of prosody.

III. THE MODEL

Theories of linguistic rhythm [7,11] generally agree that the perceived stressing of an utterance reflects the combined influence of two components:

(i) A *grouping component* which indicates the hierarchical organisation of phonological units, from phonemes at the lowest level to syllables, feet and phrases at the highest level. This hierarchy is referred to by Selkirk [11] as the *prosodic constituent structure*. Because the grouping components is hierarchical, it is usually represented as a tree.

(ii) A *metrical component* or 'metrical grid' which describes the temporal pattern of stressed and unstressed syllables.

The output of the auditory model described here is a hierarchical representation referred to as a *rhythmogram*. This representation incorporates the two components described above; the tree-like structure of the rhythmogram indicates the rhythmic grouping of auditory events, and information about the perceived loudness of each event in the rhythmogram can be used to construct a metrical grid.

The model consists of three stages, described briefly below and summarised in Figure 1. Further details of the peripheral auditory model can be found in [1], and a full description of the rhythmogram theory is given in [12].

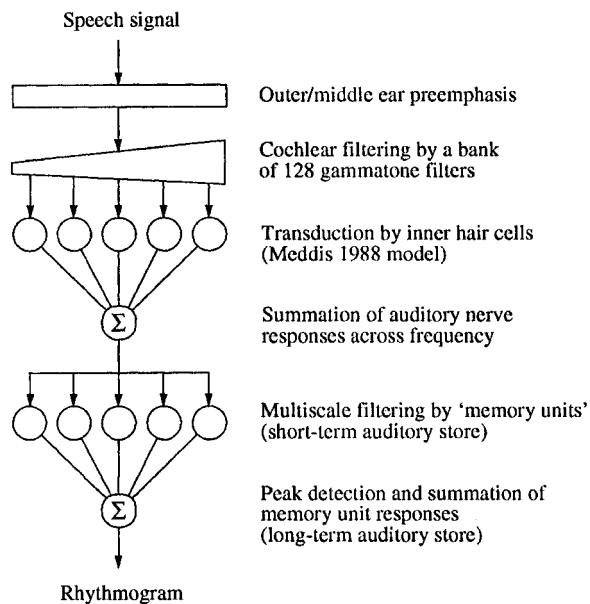


Figure 1. Schematic diagram of the auditory model.

3.1 Peripheral auditory processing

In the first stage of the model, the speech signal is processed by a simulation of the auditory periphery, comprising outer/middle ear preemphasis, cochlear filtering and transduction by inner hair cells.

The transfer function of the outer and middle ears is approximated by a simple high-pass filter of the form

$$y(t) = x(t) - 0.95x(t-1) \quad [1]$$

where $x(t)$ is the input signal at time t and $y(t)$ is the filtered output. The frequency selective properties of the basilar membrane are modelled by a bank of gammatone filters [5]. The impulse response of a gammatone filter with order n and centre frequency f_0 Hz is given by

$$g(t) = t^{n-1} \exp(-2\pi b t) \cos(2\pi f_0 t + \phi) \quad [2]$$

where ϕ is phase and b is related to bandwidth. Here, fourth order filters are used. The filters are distributed across frequency according to the ERB-rate scale of Glasberg and Moore [2]. Specifically, 128 overlapping filters were spaced equally in ERB-rate between centre frequencies of 100Hz and 5 kHz.

After cochlear filtering, each channel is processed by the Meddis model of inner hair cell transduction [4]. This yields a probabilistic representation of firing activity in the auditory nerve.

3.2 Multiscale filtering

In the second stage of the model, auditory nerve firings are summed across all centre frequencies. This 'pooled' representation of auditory nerve activity is then filtered by a multiscale mechanism, which can be interpreted as a form of auditory sensory memory. The multiscale mechanism is motivated by psychophysical evidence for two forms of auditory sensory memory, namely *short-term echoic store* which lasts for 200-300 ms and *long-term echoic store* lasting for several seconds or more [2].

Pooled auditory nerve firings are passed through a bank

of low-pass filters ('memory units') which correspond to a short-term echoic store. Each memory unit is implemented as a Gaussian approximation filter, which has a finite delay ('memory') proportional to the time constant of the filter. Subsequently, peaks in the response of the memory units are identified, and a sum of the peak responses is accumulated in a simplified model of the long-term echoic store. The accumulation process is activated by the onset of an auditory event, indicated by an abrupt increase in auditory nerve firing activity.

This mechanism is able to account for a number of important auditory phenomena, including temporal integration, persistence and masking. A detailed analysis of the properties of the model is given in [7]. Additionally, there is some physiological justification for the model, since the multiscale mechanism is consistent with the form of neural 'maps' found throughout the higher auditory system (see [1] for a review).

3.3 Rhythmogram formation

In the final stage of the model, peaks in the output of the multiscale filterbank are plotted on a time-constant/time graph. The resulting tree-like pattern indicates the rhythmic grouping of auditory events, and is referred to as a *rhythmogram*.

The form of the rhythmogram for a speech signal is determined by a number of factors. The temporal integration performed by the multiscale analysis ensures that acoustic events of a long duration or high intensity occupy relatively more important positions in the grouping hierarchy than events of shorter duration or lower intensity. Similarly, the rhythmogram is sensitive to changes in spectral energy density that accompany changes in fundamental frequency. Accordingly, the form of the rhythmogram for a speech signal is influenced by the three main acoustic determinants of prosody, namely duration, intensity and fundamental frequency.

IV. EXAMPLES

Rhythmograms of spoken language are considered here at two levels of phonological analysis. First, the rhythmogram of a monosyllabic word is related to its subsyllabic structure. Second, the metrical phonology of words and phrases is considered in terms of their stress hierarchies. It is demonstrated that there is a close relationship between the rhythmogram of an utterance and its corresponding stress hierarchy.

The utterances analysed in this section were spoken by a male native English speaker with a RP accent. The utterances were sampled at a frequency of 16 kHz with 16 bit resolution.

4.1 The phonological structure of a monosyllabic word

According to the sonority theory of the syllable (see [4] for a review), it is held that the phonological representation of a syllable can be structured hierarchically into tiers. For example, the phonological structure of the monosyllabic word 'clamp' is shown in Figure 2. Above the segment tier are X-positions, which indicate the number of epochs that a given segment occupies in the syllable. Further up in the

hierarchy, X-positions are grouped into an onset, peak and coda; in turn, the peak and coda are grouped into a rhyme. The syllable peak is associated with the segment that is more sonorous than both of its neighbours.

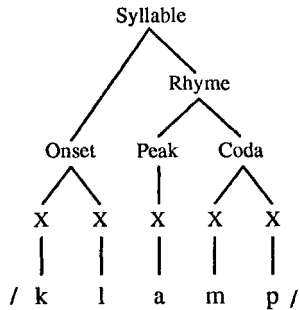


Figure 2. Phonological structure of the monosyllabic word 'clamp'. Redrawn from [4].

The rhythmogram for the word 'clamp' is shown in Figure 3. The upper panel indicates the grouping structure, and the lower panel indicates the estimated stress of each auditory event. Memory units with the shortest time constants respond to structure at the segment level; there is an event in the rhythmogram for each X-position of the syllable. Units with longer time constants respond to supra-segmental structure, giving three distinct events that correspond to the onset, peak and coda. Note that the event corresponding to the syllable peak occupies the most dominant position in the grouping hierarchy, and that this event also has the highest stress value. It is significant, however, that the stress values for the word do not exhibit a single peak, as would be predicted by the sonority theory of the syllable.

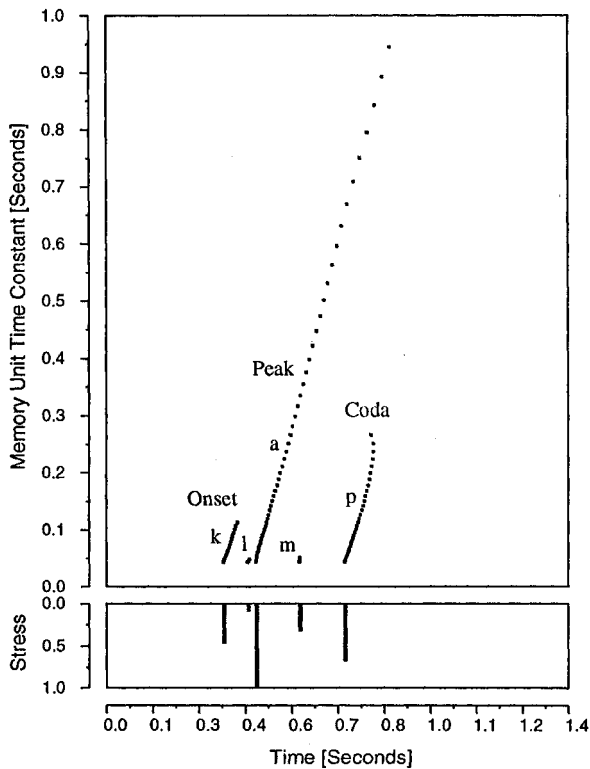


Figure 3. Rhythmogram of the monosyllabic word 'clamp'.

4.2 The rhythm of words and phrases

We now consider the relationship between the rhythmogram and stress hierarchies of the form employed in the metrical phonology literature (e.g., [7,11]). A stress hierarchy is a binary tree, in which one branch leads to a relatively stronger node (S) and the other leads to a relatively weaker node (W). The stress hierarchy for the word 'reconciliation' is shown in Figure 4.

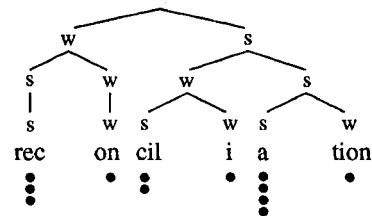


Figure 4. Stress hierarchy and metrical grid for the utterance 'reconciliation'. Redrawn from [5].

Below each node in the lowest level of the stress hierarchy, a vertical line of dots indicates the relative stress of each syllable. This so-called 'metrical grid' indicates the temporal pattern of strong and weak beats in the rhythm of the utterance. The metrical grid should be compared with the black bars in the lower panel of the rhythmogram, which indicate the estimated stress of each auditory event.

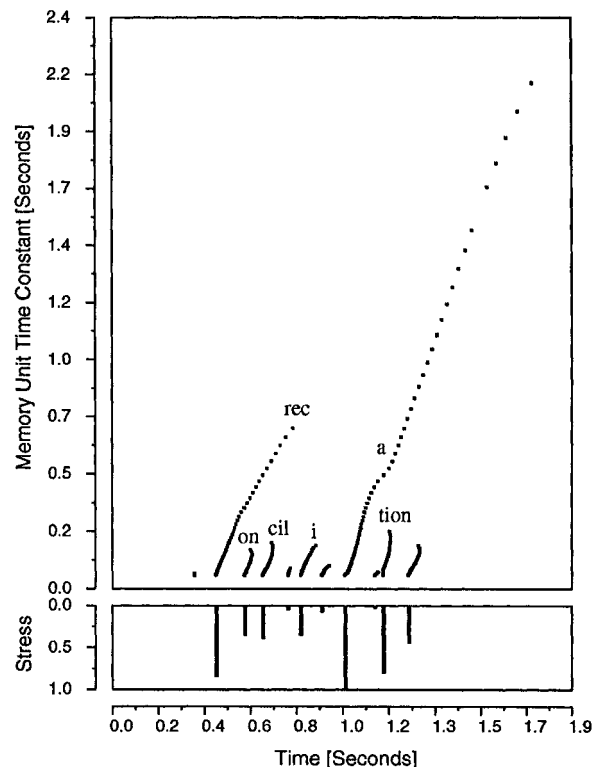


Figure 5. Rhythmogram of the utterance 'reconciliation'.

The rhythmogram for the utterance 'reconciliation' is shown in Figure 5. There is an alternation of stressed and unstressed syllables in this utterance, a fact that is reflected both in the metrical grid of Figure 4 and in the stress estimates derived from the rhythmogram. Additionally, the rhythmogram correctly indicates that the fifth syllable is the

