Proceedings of the Institute of Acoustics

THE PERCEPTION OF PROSODY: A MULTI-SCALE AUDITORY MODEL

Neil P. McAngus Todd[1] and Guy J. Brown[2]

[1]Department of Psychology, University of Manchester, Oxford Road, Manchester M13 9PL, U.K. and [2]Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, U.K.

1. INTRODUCTION

Understanding speech is not just a matter of recognising words, but also of recognising the intended meaning of those words. Indeed, recent experience with large vocabulary speech recognisers has suggested that a wide range of linguistic information must be incorporated into speech understanding systems in order to allow natural interaction between speaker and machine [15].

In particular, speakers use intonational cues to indicate the intended interpretation of their utterances. At the word level, stress can assist in identifying words from phoneme strings. At higher levels, prosodic information can resolve syntactic and semantic ambiguities and assist in the segmentation of utterances into sentences and phrases. Additionally, prosody can assist in the management of dialogues between speakers and machines. For example, Rowles *et al.* [16] have described a spoken dialogue understanding architecture in which prosodic information is used to manage turn-taking during a telephone enquiry.

Clearly, if prosody is to be utilised in speech understanding machines then computational techniques are required that can extract prosodic information from spoken language. One approach to the development of such a computational technique is to seek inspiration from the known physiology and psychophysics of human auditory function; specifically, we might ask "*what are the processes and representations that underlie the auditory extraction of prosodic features from the speech signal?*". This paper presents one answer to this question in the form of a computational model of auditory rhythmic grouping. It should be noted that the model is proposed as a general theory of rhythm perception, and is not intended specifically for speech processing applications. Indeed, the model has previously been applied to the rhythmic analysis of music with some success [19].

2. SPEECH RHYTHM AND PROSODY

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 duration, loudness and pitch [3]. However, perceived syllable stress is context dependent, being relative to the prominence of other speech sounds. Hence, as with musical rhythm, perceived *speech rhythm* - a regular ordering of stressed and unstressed speech sounds - is a joint function of stress assignments at many different levels, from the segment to the sentence [8]. Consequently, the notion of speech rhythm is intimately connected with prosody, which refers to the 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 rhythm and stress.

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

3. A COMPUTATIONAL MODEL

In general, theories of linguistic rhythm (e.g., [9],[17]) agree that the perceived stressing of an utterance reflects the combined influence of two components. Firstly, a *grouping component* indicates the hierarchical organisation of phonological units, from phonemes at the lowest level to syllables, feet and phrases at higher levels. Selkirk [17] refers to this hierarchy as the 'prosodic constituent structure'. Since the grouping component is hierarchical, it is conventionally represented as a tree. Secondly, a *metrical component* or 'metrical grid' describes the temporal pattern of relatively stressed and relatively unstressed syllables.

In the following sections, we describe a model of auditory rhythm analysis which incorporates both of the two components described above. We refer to the representation derived from the model as the *rhythmogram*. The hierarchical grouping of auditory events is indicated by the tree-like structure of the rhythmogram. Similarly, information about the 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 auditory periphery model are given in [1], and a detailed account of the rhythmogram theory can be found in [19].

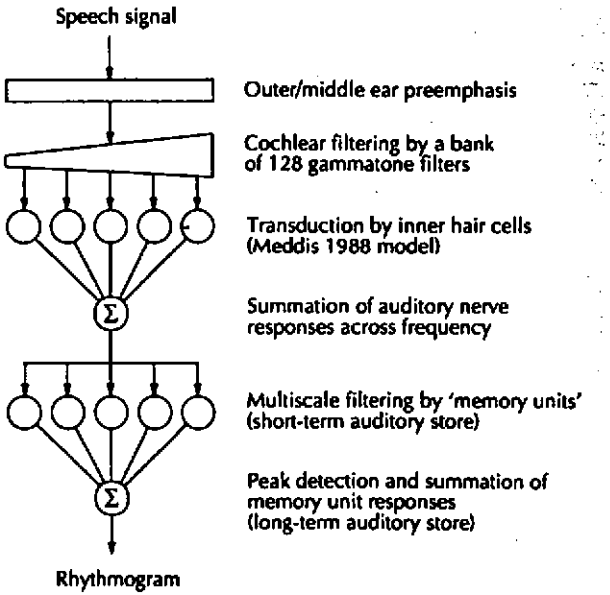

Figure 1: Block diagram of the auditory model. The model consists of three main stages; auditory periphery, multiscale filtering and rhythmogram formation.

3.1. Peripheral Auditory Processing

In the first stage of the model, the speech signal is processed by a simulation of the auditory periphery, consisting of outer and 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, and the frequency selective properties of the basilar membrane are modelled by a bank of bandpass 'gammatone' filters [14]. The filters are distributed across frequency according to the ERB-rate scale of Glasberg and Moore [4]; 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 [12] model of inner hair cell transduction. The output of the hair cell model is a probabilistic representation of spiking activity in the auditory nerve. The Meddis model is configured according to the parameters given in [13], which simulate an auditory nerve fibre with a high spontaneous firing rate.

3.2. Multiscale Filtering

In the second stage of the model, auditory nerve firings are summed across all centre frequencies to give a 'pooled' representation of auditory nerve activity. The pooled response is then filtered by a multiscale mechanism, which can be seen as an auditory analogue of the multiscale analysis that forms the first stage of Marr's [11] computational theory of vision. The multiscale mechanism can be interpreted as a form of auditory sensory memory. As such, it is consistent with psychophysical evidence for two forms of auditory 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 its time constant. 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. This accumulation process is activated by the onset of an auditory event, indicated by an abrupt increase in the pooled auditory nerve activity.

The multiscale 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 [19] and [20].

3.3. Rhythmogram Formation

In the last stage of the model, peak responses accumulated in the long-term echoic store 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 multiscale analysis performs a temporal integration, which 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. Hence, the form of the rhythmogram is influenced by the three main acoustic determinants of prosody; intensity, duration and fundamental frequency.

4. EXAMPLES

On the following pages, rhythmograms of spoken language are considered 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. The utterances employed were spoken by a male native English speaker with a RP accent, and sampled at a frequency of 16 kHz with 16 bit resolution.

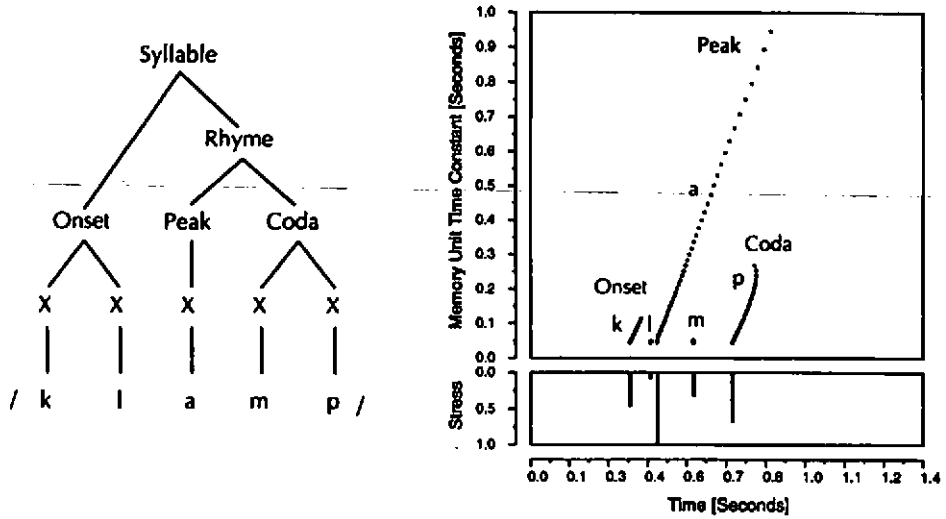

Figure 2: Phonological structure (left, redrawn from [3]) and rhythmogram (right) for the monosyllabic word 'clamp'.

4.1. Phonological structure of single syllables

Before considering the prosodic structure of speech, it is instructive to consider the response of the rhythmogram to a single syllable. According to the sonority theory, it is held that the phonological representation of a syllable can be structured hierarchically into tiers (see [3] for a review). For example, consider the phonological structure for the word 'clamp' 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; for example, each X-position in the syllable onset represents a single consonant phoneme. At higher levels of the hierarchy, X-positions are grouped into an onset, peak and coda; in turn, the peak and coda are grouped into a rhyme. The peak of the syllable is associated with the segment that is more sonorous than both of its neighbours.

It is clear from the above description that the phonological structure of a single syllable contains a hierarchical and a metrical component. This is reflected in the rhythmogram for the word 'clamp', shown in the right panel of Figure 2. The upper panel of the rhythmogram shows the grouping structure, and the lower panel indicates the estimated stress of each auditory event. Memory units with a short time constant respond to structure at the segment level, giving rise to an event in the rhythmogram for each X-position of the syllable. Memory units with longer time constants respond to supra-segmental structure, generating three large events that correspond to the onset, peak and coda of the syllable. Additionally, the event occurring at the syllable peak has the highest stress value and occupies a dominant position in the grouping hierarchy.

4.2. Phonological structure of phrases

At the phrase level, there are some intriguing similarities between the rhythmogram and stress hierarchies of the form employed in the metrical phonology literature (e.g., [9], [17]). A stress hierarchy for the phrase 'tennessee air' is shown in the left panel of Figure 3. The 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). 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.

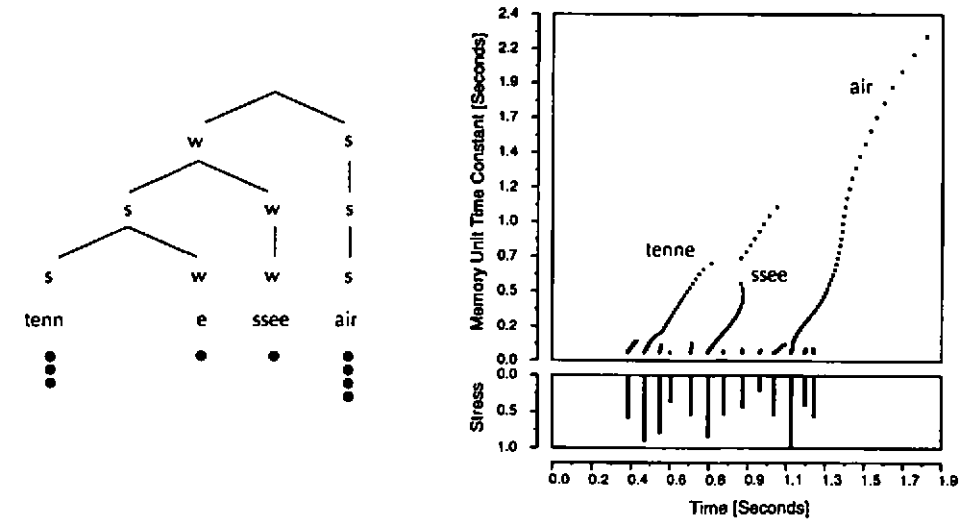

Figure 3: Stress hierarchy with metrical grid (left, redrawn from [5]) and rhythmogram (right) for the utterance 'tennessee air'.

The rhythmogram for the utterance 'tennessee air' is shown in the right panel of Figure 3. The second word in the phrase receives the highest stress and the first word has initial syllable stress ('TENNessee

AIR'). This is reflected in the grouping component of the rhythmogram, in which the second word has a dominant position, and also in the estimated stress of each auditory event. Furthermore, the rhythmogram exhibits a binary branching structure that closely resembles the corresponding stress hierarchy.

A final example is shown in Figure 4. Here, the second word in the phrase receives the strong branch and has initial syllable stress ('mississippi LEGislature'). Again, this is reflected in the grouping structure of the rhythmogram, in which the event corresponding to the fifth syllable occupies a dominant position. Additionally, the overall structure of the grouping component is similar to the branching pattern of the stress hierarchy.

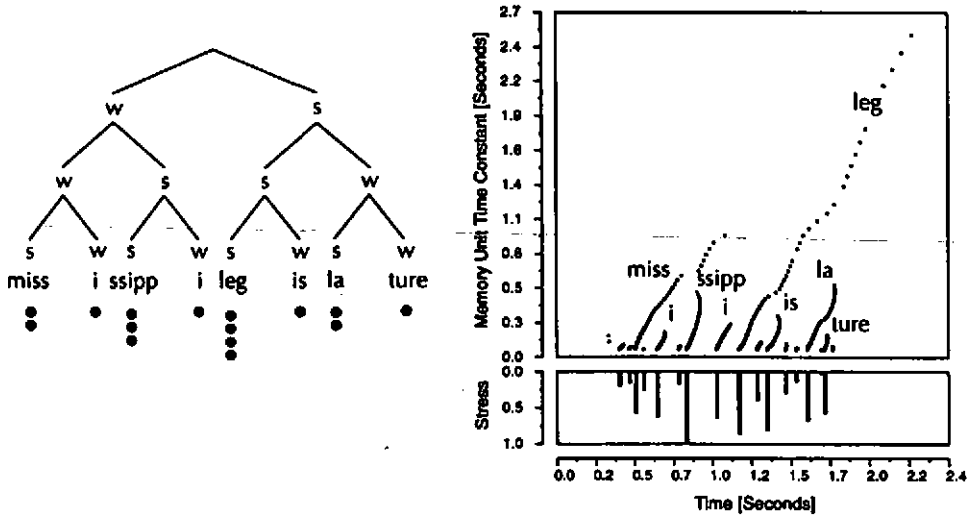

Figure 4: Stress hierarchy with metrical grid (left, adapted from [5]) and rhythmogram (right) for the utterance 'mississippi legislature'.

5. DISCUSSION AND CONCLUSIONS

The rhythmogram appears to be a powerful tool for phonological analysis. Although the evaluation presented here has been qualitative and based on a small data set, we believe that the rhythmogram is a promising method for extracting prosodic information from spoken language.

A previously published analysis of the properties of the rhythmogram has demonstrated that it is sensitive to loudness and duration, two of the acoustic correlates of prosody [19]. In addition, the rhythmogram is likely to be sensitive to variations in pitch, because these lead to corresponding fluctuations in the pooled auditory nerve response. However, the pitch sensitivity of the rhythmogram has yet to be fully investigated, and it is possible that the model could benefit from a more ex-

plicit representation of voice pitch. One possibility would be to extend the model to extract periodicity information both at pitch frequencies (50-500 Hz) and rhythmic frequencies (0.05-20 Hz). Interestingly, physiological studies have suggested that such a unified analysis of pitch and rhythm is performed in the auditory cortex [7].

Although this paper has emphasised the similarity between the rhythmogram and stress hierarchies of the form used in the metrical phonology literature, there is no reason to expect an exact correspondence. Stress hierarchies assume a binary strong-weak relation that seldom occurs in natural speech. Similarly, although English is widely regarded as a 'stress-timed' language, it is clear that stresses are only approximately isochronous. Rather, speakers use intonation to convey emotional and semantic messages, and this influences the pattern of strong and weak stresses in the hierarchy. Indeed, expressive deviations from isochrony are an important means by which a speaker can convey the intended structure of an utterance. In conclusion, then, our experience with the rhythmogram suggests that linguistic rhythm is considerably more complex than suggested by much of the phonology literature.

6. SUMMARY AND FUTURE WORK

A multiscale model of auditory rhythmic grouping has been described, and its application to the extraction of prosodic features from speech has been investigated. The results on a small data set suggest that the model is a very promising tool for phonological analysis.

Future work will quantify the performance of the rhythmogram on a larger corpus of prosodically annotated speech. Additionally, the role of the rhythmogram in speech synthesis will be investigated. For example, speech could be resynthesised from the rhythmogram, in much the same way that other auditory representations have been inverted (e.g., [18]). Additionally, it is widely acknowledged that rhythm is a crucial factor in obtaining natural sounding synthetic speech [10] and the rhythmogram may provide a novel means of coding prosodic information for speech synthesis.

Finally, there is some evidence that rhythm plays a role in auditory stream segregation [6]. Hence, the rhythmogram could form one component of a computational system for the segregation of concurrent sounds.

ACKNOWLEDGMENTS

Neil Todd is supported by MRC grant G9018013. Guy Brown is supported by SERC Image Interpretation Initiative grant GR/H53174, EPSRC Standard Research Grant GR/K18962 and the Nuffield Foundation.

REFERENCES

[1] GJ Brown & MP Cooke, 'Computational auditory scene analysis', *Speech Communication*, in press (1994)
[2] N Cowen, 'On short and long term auditory stores' *Psychological Bulletin*, **96**, pp. 341-370 (1984)
[3] HJ Giegerich, *'English phonology: An introduction'*, Cambridge University Press (1992)
[4] B Glasberg & BCJ Moore, 'Derivation of auditory filter shapes from notched-noise data', *Hearing Research*, **47**, pp. 103-138 (1990)
[5] S Handel, *'Listening'*, MIT Press (1993)
[6] S Handel, MS Weaver & G Lawson, 'Effect of rhythmic grouping on stream segregation', *Journal of Experimental Psychology: Human Perception and Performance*, **9**, pp.637-651 (1983)
[7] B Hose, G Langner & H Scheich, 'Topographic representation of periodicities in the forebrain of the mynah bird: one map for pitch and rhythm?', *Brain Research*, **422**, 367-373 (1987)
[8] F Lerdahl & R Jackendoff, *'A generative theory of tonal music'*, MIT Press (1983)
[9] M Liberman & A Prince, 'On stress and linguistic rhythm', *Linguistic Inquiry*, **8**, pp. 249-336 (1977)
[10] J Local, 'On the phonetic interpretation of rhythm in non-segmental speech synthesis', *Proceedings of the Institute of Acoustics*, 14, 473-480 (1992)
[11] D Marr, *'Vision'*, WH Freeman & Co (1982)
[12] R Meddis, 'Simulation of mechanical to neural transduction in the auditory receptor', *Journal of the Acoustical Society of America*, **79**, pp. 702-711 (1986)
[13] R Meddis, 'Simulation of auditory-neural transduction: further studies', *Journal of the Acoustical Society of America*, **83**, pp. 1056-1063 (1988)
[14] R Patterson, I Nimmo-Smith, J Holdsworth, & P Rice, *'SVOS final report: the gammatone filter bank'*, MRC Applied Psychology Unit Report 2341 (1988)
[15] C Rowles, X Huang & G Aumann, 'Natural language understanding and speech recognition: Exploring the connections', *Proceedings of the Third Australian Conference on Speech Science and Technology*, pp. 374-379 (1990)
[16] C Rowles, X Huang, M de Beler, J Vonwiller, R King, C Matthiesson, P. Sefton & M O'Donnell, 'Using prosody to assist in the understanding of spoken English', *Proceedings of the Fourth Australian International Conference on Speech Science and Technology*, pp. 248-253 (1992)
[17] EO Selkirk, 'The role of prosodic categories in English word stress', *Linguistic Inquiry*, **11**, pp. 563-605 (1980)
[18] M Slaney, D Naar & RF Lyon, 'Auditory model inversion for sound separation', *Proceedings of the IEEE International Conference on Acoustics, Speech & Signal Processing*, **2**, pp. 77-80 (1994)
[19] NP Todd, 'The auditory primal sketch: A multiscale model of rhythmic grouping', *Journal of New Music Research*, **23**, pp. 25-70 (1994)
[20] NP Todd & GJ Brown, 'Visualisation of rhythmic structure', *Artificial Intelligence Review*, **8**, in press (1995)