

Predicting the Grouping of Rhythmic Sequences using Local Estimators of Information Content *

Steven M. Boker

Department of Psychology
The University of Notre Dame
Notre Dame, Indiana 46556

Michael Kubovy

Department of Psychology
The University of Virginia
Charlottesville, Virginia 22903

July 20, 2005

Abstract

The hypothesis is proposed that auditory events are perceived to be partitioned according to boundaries constructed at times of maximum surprise. One way of quantifying surprise is via information theoretic predictions. The results of two experiments are presented that test the plausibility of this hypothesis using simple repeating auditory rhythmic sequences. Local estimators of information content within an auditory sequence are used to construct predictors of perceived segmentation. These predictors are fit to results of the experiment by using a structural equation model and compared Garner's Run-Gap model (Garner, 1974). The information theoretic model is found to be a significantly better predictor of the experimental results than the Run-Gap model.

1 Introduction

The segmentation and ordering of an continuous sensory stream into a series of recognizable events presents one of the fundamental problems in perception (Lashley, 1952). The auditory system must partition the incoming stream in a meaningful way; one that preserves relationships within the stream, but also breaks the stream into a sequence and thus allows the recognition of words, phrases or sentences.

In music, these sequential units are formalized and regularized in such a way that much of the possible ambiguity in segmentation is removed. This process of disambiguation is achieved in a variety of ways, primarily by stress or accent (Handel, 1989). Why is there a need to disambiguate the structure of such regular musical patterns? The rhythmic structure of spoken language carries content that helps the listener organize the sounds of speech into grammatical sense during the process of comprehending the meaning of the sentences (Thomas, Hill, Carrol, & Garcia, 1970). Why is there a need for prosodic elements in speech?

By studying the nature of the ambiguity in rhythmic sequences we can understand and predict the organization which will be perceived to be inherent in the timing of the sequence. This will lead

*Presented to the International Joint Conference on Artificial Intelligence Workshop on Computational Auditory Scene Analysis, August 19–20, 1995

to a more precise understanding of the interrelationships between stress and timing which create the unambiguous perception of segmentation of auditory streams.

A simple repeating rhythmic pattern which contains no stressed elements may be perceived as having a variety of starting points. Figure 1 shows a repeating sequence that could be perceived as having one of three potential starting points. Some starting points have a higher probability of being perceived than others, but each of these probabilities is greater than zero.

```
1 0 1 1 0 1 1 0 1 1 0 1 ...
0 1 1 0 1 1 0 1 1 0 1 1 ...
1 1 0 1 1 0 1 1 0 1 1 0 ...
```

Figure 1: A repeating sequence of length 3 has 3 potential starting points.

Garner and his colleagues (Royer & Garner, 1966; Garner & Gottwald, 1968; Garner, 1974) studied these types of rhythmic patterns and devised heuristics, which they named the *run principle* and *gap principle*, by which predictions could be made regarding the organization that would be perceived by individual subjects. The work presented here replaces Garner’s heuristics with a more formal information theoretic (Shannon & Weaver, 1949) estimation of the probability of perceiving any starting point as a segmentation boundary. This relationship between local information content in the perceptual stream and the perception of temporal segmentation is likely to generalize to the other sensory modalities.

2 Methods

2.1 Experiment 1

2.1.1 Subjects

Eleven subjects participated in Experiment 1, 8 males and 3 females. Age of the subjects ranged from 18 to 41. Six subjects reported having received more than 4 years of training in playing a musical instrument, whereas the remaining five subjects reported no formal training in playing a musical instrument. All but two subjects reported being right-handed.

2.1.2 Experimental Procedure

Each subject was asked to complete the experimental procedure once on each of five occasions, where the occasions were separated by as little as 24 hours and by as much as three weeks. Each session required approximately 45 minutes to complete.

In each trial, subjects were presented with one repeating rhythmic auditory pattern and were asked to respond by striking a key on a synthesizer keyboard synchronous with the perceived starting point of the pattern. Subjects were asked to continue to strike the key at the beginning of each repetition until confident that they had perceived the starting point. Once subjects were confident of their response, they were asked to press a mouse button ending the trial.

The PsyLog software (Boker & McArdle, 1992), running on a NeXT workstation, was modified to present the stimuli and gather the responses. Subjects were presented with a short set of practice

trials and then a set of 115 experimental trials. Subjects were informed that they could rest at any time that they became fatigued and that the software would wait for them.

A single rhythmic stimulus was composed of a fixed number of *beats*, equal intervals of time which could either be empty or be filled with a percussive sound at the beginning of the interval. Thus, a measure could be represented by a binary number where each binary digit represents a beat: a zero representing an empty interval and a one representing an interval with a percussive sound at its beginning. The set of stimuli for Experiment 1 consisted of all of the unique rhythmic patterns of length 8 or less, 115 patterns in all.

In each trial, the stimulus initially began with a short beat length of 10 ms, which quickly slowed to a steady beat length of 250 ms. The beat on which the stimulus was initiated was chosen at random for each presentation of the stimulus. The combination of these two methods minimized the subjects' ability to associate the beginning of the presentation of the first beat of the stimulus with the perceived beginning of a measure within the repeating pattern.

The percussive sound used in this experiment was a synthesized musical cowbell produced by a Roland MT-32 MIDI wavetable synthesis module and delivered to the subjects binaurally via Sennheiser HD-414-SL headphones. The user responded by striking a key on a Kawai K5 Digital Synthesizer keyboard. Several variables were measured for each keypress: the time of response in milliseconds relative to the beginning of the presentation of the stimulus; the time of response in milliseconds relative to the beginning of the pattern as represented internally by the computer software and the velocity of the response as an integer between 1 and 127.

2.2 Experiment 2

2.2.1 Subjects

Twenty eight subjects participated in Experiment 2, 17 males and 11 females. Age of the subjects ranged from 18 to 21. Thirteen subjects reported more than 4 years of training in playing a musical instrument. Sixteen subjects reported being right-handed.

2.2.2 Experimental Procedure

The experimental procedure was identical to that of Experiment 1 with the following two exceptions. The subjects in Experiment 2 were only tested on one occasion. The set of rhythmic stimuli in Experiment 2 consisted of a random sample of 115 stimuli drawn half from the unique patterns of length 8 or less and half from the unique patterns of length 12.

3 Models

3.1 Run-Gap Predictions

A latent variable structural equation model was employed to test the goodness of fit of Garner's "run-gap" heuristic predictions to the data gathered from the two experiments. Figure 2 shows a path model of Garner's run-gap heuristics. The predictor variables are *Run*, the run principle, and *Gap*, the gap principle. The latent variable is *S*, the perceived structure of the rhythmic pattern. The measured outcome variables are *RB*, the response within the beat; *A*, the accuracy of the response; and *V*, the velocity of the response.

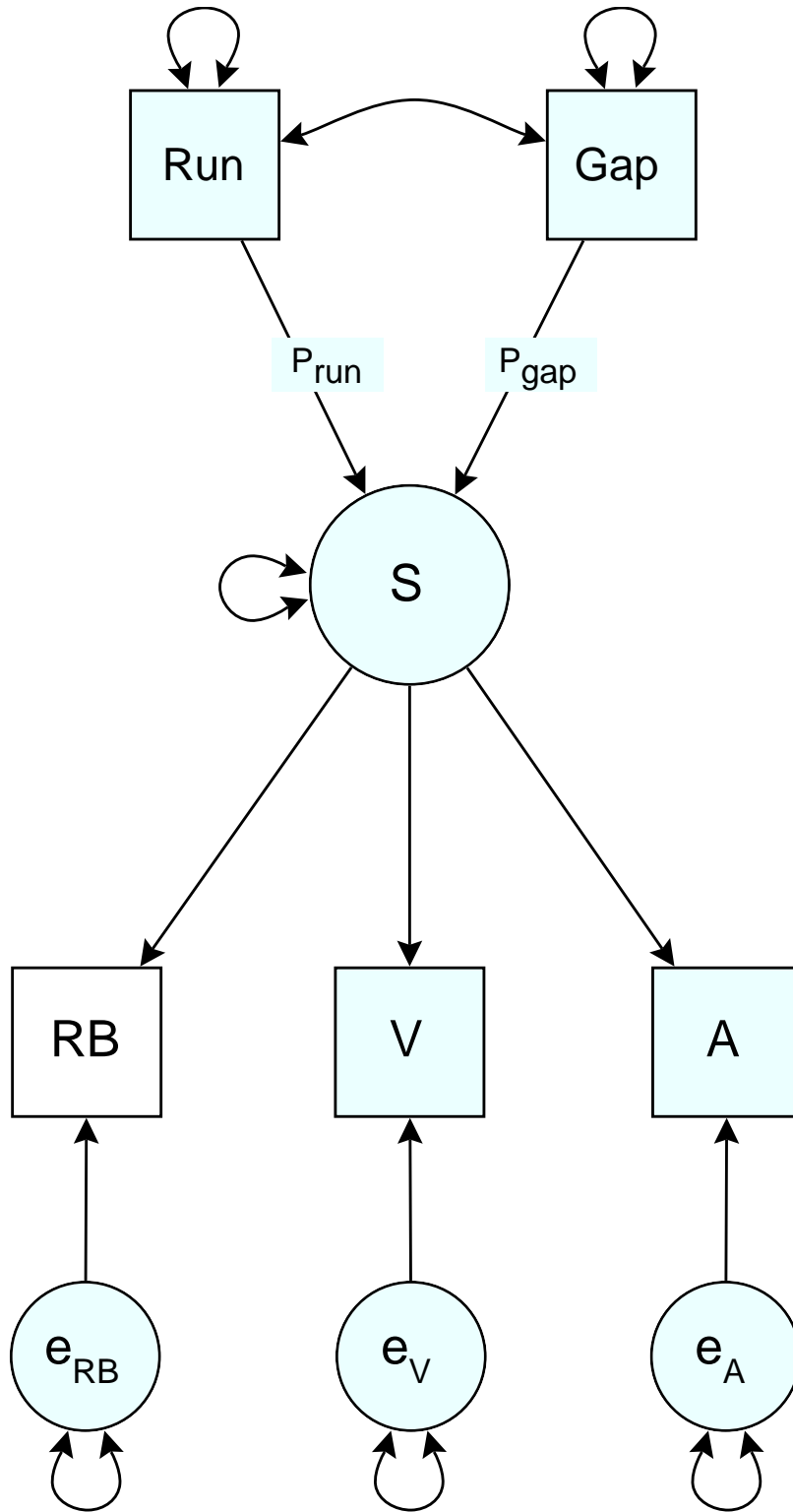


Figure 2: Path diagram showing a structural model of Garner's basic run-gap theory.

The *Run* variable is constructed as follows: if the current stimulus beat value is 1 and the previous stimulus beat value is 0, then $Run = 1 +$ the number of stimulus beats following the current stimulus beat before another stimulus beat with a value of 0 is encountered. The *Gap* variable is constructed similarly: if the current stimulus beat value is 1 and the previous stimulus beat value is 0, then $Gap = 1 +$ the number of stimulus beats preceding the current stimulus beat before a stimulus beat with a value of 1 is encountered.

The outcome variables were coded as follows. If a keypress occurred within ± 0.5 of the onset of the current stimulus beat, then RB was coded as 1, otherwise as 0. If RB was 1, the MIDI velocity of the keypress (range 1–127) was coded as V . If RB was 1, A was coded as

$$A = 1 - \frac{2(t_k - t_b)}{b}$$

where t_k is the elapsed time to the keypress, t_b is the elapsed time to the onset of the stimulus beat, and b is the duration of the beat. This means that the accuracy of the keypress was 1 if the keypress occurred simultaneously with the onset of the stimulus beat, and the accuracy was 0 if the keypress occurred halfway between two stimulus beats.

RM was coded as 0 if no keypress occurred within ± 0.5 measure of the current stimulus beat. Thus RM began the trial as a string of 0's and became a string of 1's when the subject began to respond to the stimulus.

3.2 Local Information Predictions

A similar latent variable model was constructed to test the fit of the local information content predictions to the data from Experiments 1 and 2 (see Figure 3). The outcome variables were coded in exactly the same manner as for the Run–Gap Model above. The predictor variables were calculated in terms of redundancy for features of a particular size.

Redundancy can be stated in terms of a ratio of entropies (Barlow, 1961), and was explored by Redlich (Redlich, 1993) as an active mechanism in visual perception. In the case of segmentation, the quantity which needs to be calculated is a measure of local information content rather than redundancy. The measure of local information content which will be used here is simply $1 - R$, where R is redundancy.

Consider a repeating sequence of beats. If r is the number of contiguous repetitions with which a feature x of size s has previously occurred, then one possible measure of local surprise upon the reoccurrence of x is

$$\begin{aligned} L_x &= \frac{H_x}{H_{\not{x}}} \\ &= \frac{-r/(r+1) \log_2(r/(r+1))}{-1/(r+1) \log_2(1/(r+1))} \\ &= \frac{r \log_2(r/(r+1))}{\log_2(1/(r+1))} \end{aligned}$$

whereas if x does not re-occur then

$$L_x = \frac{H_{\not{x}}}{H_x}$$

$$= \frac{\log_2(1/(r+1))}{r \log_2(r/(r+1))}.$$

For each beat in the rhythmic sequence, three local measures of information content were calculated L_1 , L_2 , and L_3 . These three predictor variables have a value for each beat of each rhythmic sequence.

4 Results

The two models were fit to the data from Experiment 1 and Experiment 2 using the structural equation modeling procedure in SAS (PROC CALIS). The results of fitting the Run-Gap Model and the Local Information Model are presented in Table 1 and Table 2. The estimated parameters and χ^2 statistics are presented side by side for the two models.

Table 1: Comparison of prediction model parameters and χ^2 for models fit to the data from Experiment 1.

	Run-Gap	Entropy	
Run→S	0.043	0.295	$L_1 \rightarrow S$
Gap→S	0.131	-0.056	$L_2 \rightarrow S$
		-0.132	$L_3 \rightarrow S$
S→RB	=1	=1	S→RB
S→V	0.539	0.543	S→V
S→A	0.618	0.614	S→A
eRB→RB	-0.025	0.021	eRB→RB
eV→V	0.032	0.032	eV→V
eA→A	0.107	0.108	eA→A
Run↔Gap	0.603	0.009	$L_1 \leftrightarrow L_2$
		0.001	$L_1 \leftrightarrow L_3$
		0.000	$L_2 \leftrightarrow L_3$
Var(Run)	0.990	0.091	Var(L_1)
Var(Gap)	0.965	0.029	Var(L_2)
		0.015	Var(L_3)
Var(S)	0.105	0.124	Var(S)
χ^2	2623	79	χ^2
DF	4	6	DF
N	174781	174145	N

In Table 1 the Run-Gap model has a χ^2 fit statistic of 2623 with 4 degrees of freedom. Although this might seem large, recall that the effective sample size is 174,781 separate stimulus response pairs which contribute to a null model χ^2 of 923,649 with 10 degrees of freedom. The Run-Gap model certainly fits much better than the null model. However, notice that the Local Information (Entropy) Model has a χ^2 of only 79 with 6 degrees of freedom. Although these two models are

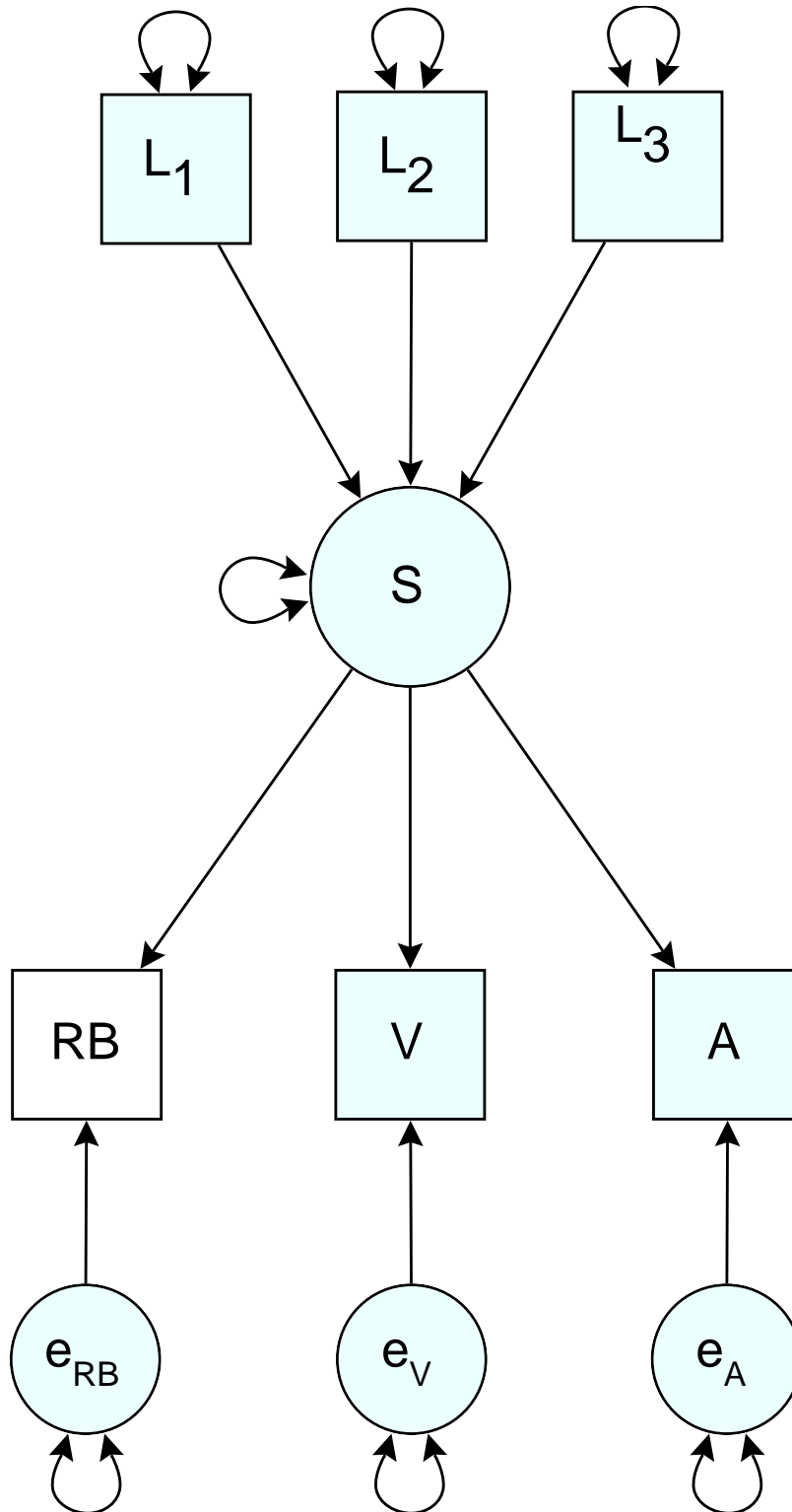


Figure 3: Path diagram of entropy prediction model where the predictors are: L_1 , entropy of features of length 1; L_2 , entropy of features of length 2; and L_3 , entropy of features of length 3.

not nested and so cannot be compared precisely in terms of χ^2 goodness of fit, the difference in χ^2 is so great that it overwhelms the possible loss of accuracy due to the non-nested nature of the comparison.

Table 2 shows the results of the analysis on Experiment 2. The Run–Gap model has a χ^2 fit statistic of 388 with 4 degrees of freedom compared to a null model χ^2 of 541,596 with 10 degrees of freedom. The Local Information (Entropy) Model has a χ^2 of 59 with 6 degrees of freedom. Again these two models aren’t nested, but the difference in χ^2 between the two models is large enough that the Local Information Model is preferred.

Table 2: Comparison of prediction model parameters and χ^2 for models fit to the data from Experiment 2.

	Run–Gap	Entropy	
Run→S	0.021	0.161	$L_1 \rightarrow S$
Gap→S	0.082	0.031	$L_2 \rightarrow S$
		-0.038	$L_3 \rightarrow S$
S→RB	=1	=1	S→RB
S→V	0.533	0.533	S→V
S→A	0.595	0.595	S→A
eRB→RB	0.039	0.040	eRB→RB
eV→V	0.031	0.031	eV→V
eA→A	0.092	0.092	eA→A
Run↔Gap	0.779	0.021	$L_1 \leftrightarrow L_2$
		0.004	$L_1 \leftrightarrow L_3$
		0.004	$L_2 \leftrightarrow L_3$
Var(Run)	1.302	0.127	Var(L_1)
Var(Gap)	1.284	0.037	Var(L_2)
		0.014	Var(L_3)
Var(S)	0.085	0.093	Var(S)
χ^2	388	59	χ^2
DF	4	6	DF
N	104512	104512	N

5 Discussion

The results of these analyses suggest that predicting segmentation boundaries in auditory streams by using local estimators of information content may result in calculated segmentations of temporal structure that mimic those perceived by human listeners. An algorithm estimating information content using ratios of entropies could be implemented in a computationally efficient manner since the algorithm only requires two \log_2 operations, one divide and a few lookups. The algorithm has the potential for parallel implementation in that the components for each feature size are calculated independently and then linearly combined.

The Local Information Model predicts that the probability of a temporal segmentation occurring at any point is directly related to the amount of information in the auditory stream at that point in time. This prediction is difficult to verify without a specific measurement model for estimating the local information in a stream on a moment by moment basis. Finding a reliable estimator for local information in complex stimuli remains an open problem, but a problem which if solved seems likely to hold rewards for research in auditory scene analysis.

6 Acknowledgments

We would like to thank Jay Friedenbergh whose help in performing this experiment was invaluable. We would also like to thank several anonymous reviewers whose comments and criticisms have substantially improved this chapter.

References

- Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.
- Boker, S. M., & McArdle, J. J. (1992). *Psylog: Software for psychometric measurement*. (Unpublished software, Department of Psychology, University of Virginia)
- Garner, W. R. (1974). *The processing of information and structure*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Garner, W. R., & Gottwald, R. L. (1968). The perception and learning of temporal patterns. *Quarterly Journal of Experimental Psychology*, *20*, 97–109.
- Handel, S. (1989). *Listening: An introduction to the perception of auditory events*. Cambridge, MA: MIT Press.
- Lashley, K. S. (1952). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley.
- Redlich, N. A. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, *5*, 289–304.
- Royer, F. L., & Garner, W. (1966). Response uncertainty and perceptual difficulty of auditory temporal patterns. *Perception and Psychophysics*, *1*(1), 41–47.
- Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: The University of Illinois Press.
- Thomas, I. B., Hill, P. B., Carrol, F. S., & Garcia, B. (1970). Temporal order in the perception of vowels. *Journal of the Acoustical Society of America*, *48*, 1010–1013.