Information Theoretic Auditory Grouping

Paris Smaragdis
Machine Listening Group
MIT Media Laboratory
Room E15-401C
Cambridge, MA 02139-4307 USA
paris@media.mit.edu

Abstract

Computational gestalt grouping has been so far dictated by heuristic rules. But unfortunately such descriptions result in tedious and inaccurate translations to computer programs, and they do not carry across different perceptual domains. In this paper we present a unified view of gestalt grouping based on information theory.

1.0 Introduction

Recent developments in the field of mathematical computation have enabled us to deal with raw data in more sophisticated manners than ever before. The introduction of new information theoretic computations and algorithms have provided new tools from which to examine data and have proven to be very valuable in various fields of research. For many researchers this has been a very fortunate since elegant theories linking perception with statistics and information theory, dating back to the mid-50's [Attneave 1954], are now ready to be verified or even used successfully. This paper will make an attempt to incorporate such techniques in the realm of auditory processing. Even though the focus will be mainly on substituting the gestalt heuristics with a more elegant and simple rule, the main idea to be conveyed is advocacy towards a broader integration of information theory with computational auditory perception systems. The choice to concentrate on gestalt grouping was made due to its close semantic concept with information theoretic ideas.

Gestalt psychology has found its way into a lot of work on auditory perception [Handel 1993, Bregman 1990] and early computational auditory segregation systems [Vercoe and Cumming 1988; Duda *el al*. 1990; Cooke 1991; Mellinger 1991; Ellis 1992]. Although it is fairly popular and has provided moderate success, it suffers by being dictated by vague and disconnected principles which are not easily translated to computers; and if they are, they only apply to the specific representation used. Implementations are forced to use heuristic rules that have fuzzy trade-off boundaries and are bootstraped for specific tasks. One of the goals of this paper is to attempt to shed light to a deeper principle of which the gestalt laws are parts of, or by-products.

The development of information theory by Shannin [Shannon and Weaver 1963], although designed for analysis of electrical information transmission systems, does have deeper extensions dealing with the general communications process. Perception, is a form of communication, it is the reception of sensory information and its translation to another format (mid- or high-level representation). Both of these processes are main themes of information theory and will be exploited in the following sections.

2.0 Main hypothesis

Seminal papers on sensory systems by Barlow [Barlow 1959; Barlow 1961; Barlow 1989] and Attneave [Attneave 1954] have drawn strong ties between perception and information theory. More applied work by Linsker [Linsker 1988], Redlich [Redlich 1993] and Atick and Redlich [Atick and Redlich 1990], has yielded interesting results by integrating these two domains. And recently, work on Independent Component Analysis (ICA) [Comon 1989] has provided some robust solutions to problems similar to long-standing ones in the realm of perception. This was work related to the cocktail party problem, initiated by Bell and Sejnowski [Bell and Sejnowski 1995] and Amari [Amari *et al.* 1996], as well as in the formation of perceptual preprocessors by Bell [Bell 1996; Bell 1996a], Deco and Obradovic [Deco and Obradovic 1995]. The common theme

between all of this work was the idea of redundancy reduction or, entropy minimization.

Sensory input has a very large bandwidth and conveys a lot of redundant information. Our sensory systems, have learned to latch on that redundancy to perform their tasks. If we are presented with a scene of non-redundant signals (e.g. white noise or single periods of tones), we cannot make any kind of perceptual analysis; in fact in the imaginary 'noise-land', where all stimuli are white noise, it would be impossible for perception and cognition to develop. It is the high degree of structure that we take advantage of, to perceive the world. That observation has lead some researchers to conclude that the perceptual system is a sophisticated redundancy reduction machine (or more bluntly, a data compression engine!).

Gestalt grouping is seen as a function of the perceptual system and as hinted above we can examine it from a redundancy reduction perspective. Upon closer examination of the gestalt principles it is clear that they are all describing statistical dependencies of various types. These dependencies contribute to the redundancy of a signal and they are what we use to perform grouping. So we could assume that if we make a system to outline these dependencies we would in fact do gestalt grouping using only one rule.

Throughout this paper the terms of statistical dependence, structure, entropy and redundancy will be used to better describe a gestalt situation. These measures are dependent on each other and are only different facets of the same thing. The increase of statistical dependencies will result in more redundancy, more structure and less entropy. Since of these measures, only entropy is relatively easy to compute we will be using it as a cost function.

3.0 Experimental data

3.1 Method

As advocated in preceding section, we will do whatever we can to reduce redundancy. In order to be able to perform grouping we will assume that we have a set of sound atoms or objects that comprise an auditory scene. The decision on whether we need to group a pair/set of these objects will be based on their correlations, or on how much their partitioned sum reduces redundancy. Although the definition of an object is better left as an abstract concept, for the purposes of this paper we will adopt the sinusoid because of its frequent appearance in the psychoacoustic literature, and because it is a convenient vehicle to illustrate gestalt grouping. The type of object that we choose is not important for our purposes since the idea presented is abstract and devoid of a fixed front-end. In the last section of the paper we will elaborate on what would be more appropriate objects which would complement the overall philosophy in this paper.

In order to verify our suspicion that gestalt-like grouping corresponds to entropy minimization we will be measuring the entropy of different configurations of sinusoids and see whether our observations correspond to a gestalt-dictated grouping. In all of the short experiments we will have a parameterized, two object, auditory scene over a variable *n*. For only one value of *n* the scene will provide a situation that according to the gestalt groups the objects should fuse. We will expect that the entropy of the resulting sound of that scene will be minimized for that value of *n*.

Although in some cases it is algebraically possible to derive the exact value of entropy we will be using a numerical method for estimating it. The algebraic computation is extremely tedious and, with the exception of trivial cases, only practical using symbolic computation programs. The way entropy will be computed is standard and straightforward. We estimate the PDF of the function to examine from its histogram and then apply the entropy computation formula:

$$h = -\sum_{x} p(x) \cdot \ln p(x) \qquad (1)$$

Before the computation, the histogram of the function is inspected so that it is a fair approximation to the expected PDF (i.e. has no gaps) and then normalized to integrate to unity. Because of the special nature of the PDFs of the objects that we decided to use (sinusoids), it would not be accurate to use more sophisticated algorithms for estimating entropy, such as Parzen histogram estimation or common maximum likelihood approaches†.

The following section presents the experimental measurements that we made. Note that the figures presented are not all in the same scale. The variance of entropy through different experiments was high and for the sake of better visualization the plots were zoomed accordingly.

3.2 Common Fate

Frequency and amplitude modulation

Common fate in the auditory domain is usually linked to two parameters, frequency and amplitude, and it is described as common modulation of these. If a set of sinusoids feature common modulation in either amplitude or frequency (or both), then they are fused together as one sound. Redundancy in this case is created from the statistical dependencies that a common frequency or amplitude track will generate.

---

†. The reader is forewarned though that this method of estimating entropy is not optimal. If the bin widths of the histograms are too wide or too narrow we will have poor estimates of entropy. A more practical approach would be to measure higher order cross cumulants instead, but this approach would be distracting us from the pure idea presented.

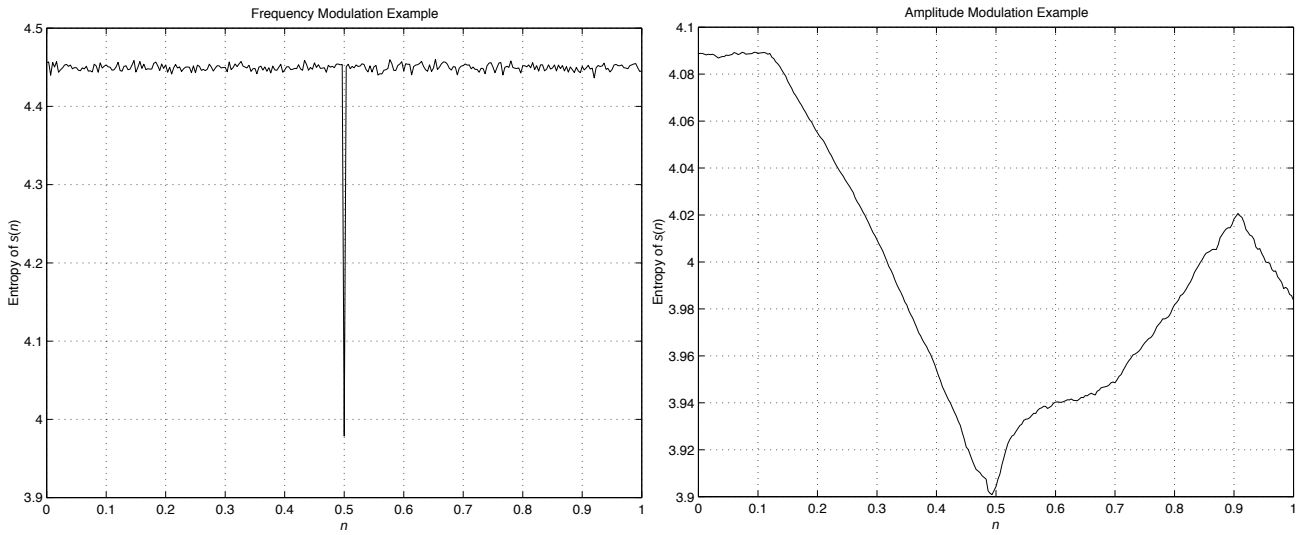

Figure 1. Entropy measurements for common fate experiments. Note the different scale across plots.

According to our principle, we would expect the entropy measure to drop when we have co-modulated sinusoids. The following two short experiments prove to behave as the gestalt rules predict for common modulation cases.

In the first case we use the parameterized sum of two frequency modulated sinusoids, modulated by two uncorrelated and arbitrary functions f_1 and f_2 as

$$s(n) = \cos\left(f \cdot \frac{(f_1 + f_2)}{2} \cdot t\right) + \cos(2f \cdot (n \cdot f_1 + (1-n) \cdot f_2) \cdot t) \quad (2)$$

where $n = 0 \ldots 1$ and $f = 1300Hz$. Obviously if n takes the value of $\frac{1}{2}$ then the two sinusoids develop the same frequency modulation therefore fusing as one sound. If we were to measure the entropy of *s(n)*, with respect to *n*, we would expect a dramatic drop at that point because of the fusion.

Similarly we set up the same experiment with amplitude modulation where:

$$s(n) = \frac{(f_1 + f_2)}{2} \cdot \cos(1300 \cdot t) + (n \cdot f_1 + (1-n) \cdot f_2) \cdot \cos(2600 \cdot t) \quad (3)$$

Our measurements validate the hypothesis as shown in Figure 1.

Note that in the amplitude modulation case the entropy difference is much more subtle, something that hints that amplitude modulation is not a very strong grouping cue, compared to frequency modulation (in other words if these modulations were to compete at the levels presented, frequency modulation would dominate).

Common onsets/offsets

Common onset and offset of partials is another example of common fate and also a key clue in fusion. If two sinuoids coincide in time, then they exhibit a correlation which results in redundancy. Just like in the previous section, this redundancy is maximized when the on and off boundaries are exactly the same for both sounds. When that happens, we would expect a drop of entropy.

The following function is set up:

$$s(n) = f(t) + f(2t + n) \quad (4)$$

where:

$$f(t) = \begin{cases} t_1 \le t \le t_2 & f(t) = \cos(t) \\ otherwise & f(t) = 0 \end{cases} \quad (5)$$

The function *f* is essentially a time-bounded sinusoid from t_1 to t_2. By varying the value of n we produce two sinusoids that are misaligned in time for $n \neq 0$, and perfectly aligned otherwise. We would expect to see a drop in the entropy of *s(n)* at the point where the two sinusoids align. The results are shown in Figure 2. Similar results are obtained when we change only the onset or the offset of the second sinusoid rather than both.

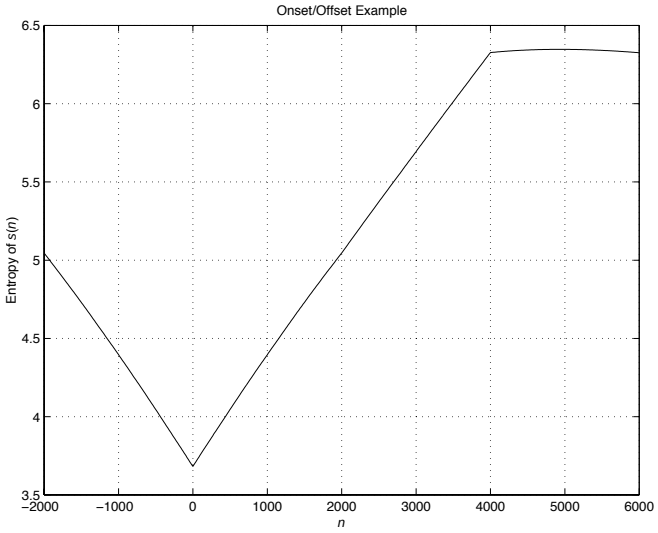

Figure 2. Entropy measurements for common time onset/offset

3.3 Harmonicity - Proximity

Harmonicity is one of the major gestalt principles in auditory grouping. Gestalt auditory theory states that two sinusoids, where one of which has a frequency which is an integer multiple of the other's (harmonic relation), are fused together. Harmonic relations exhibit more redundancy than non-harmonic relations (a short proof is that a harmonic pair of partials only requires a period of the lowest frequency to be encoded in a wavetable, whereas a non-harmonic pair would always require more). The following experiment attempts to perform grouping based on harmonicity by reducing entropy.

In this example we have a parameterized sum of two sinusoids as:

$$s(n) = \cos(f \cdot t) + \cos(n \cdot \dot{f} \cdot t) \qquad (6)$$

where $n = 1 \ldots 15$ and $f = 1000Hz$. We would expect the entropy to be minimum when n is an integer, therefore the two sinusoids would be harmonic. The results are show in Figure 3. Our prediction was right and we also make two more observations. First, we appear to have lesser entropy dips where $n = k + \frac{1}{2}, k = \text{integer}$. At these points we have the following effect. The two sinusoids act as a first and second harmonic to a non-existing fundamental at $\frac{f}{2}$, this is a lesser degree of harmonicity, but valid nevertheless. We also note that when $n = \frac{1}{2}$ we get stronger fusion than when $n = 3$, something which we can (subjectively) verify by ear. The note sensation is indeed $\frac{f}{2}$ for the first case and f for the second one. This also leads us to the second observation to be made; proximity is also tracked in this procedure. Note that the further away the second sinusoid goes the less entropy increases.

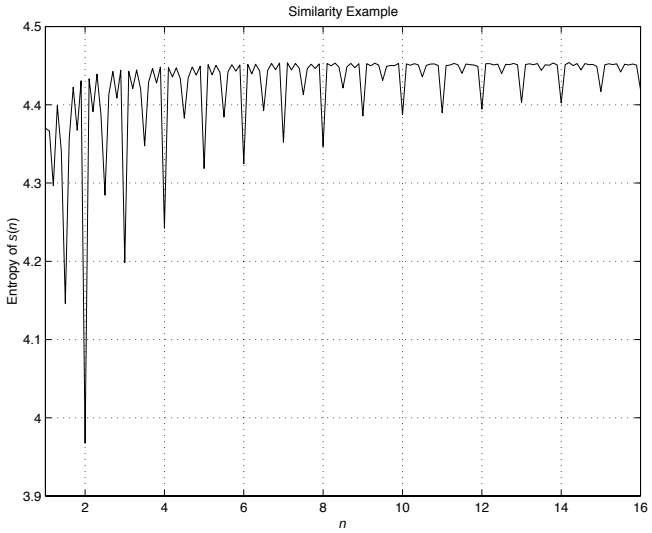

Figure 3. Entropy measurements for similarity experiment

In order to verify some of our suspicions on proximity we can insert another sinusoid in Eq. 6 and transform it to:

$$s(n) = \cos(f \cdot t) + \cos(2f \cdot t) + \cos(n \cdot \dot{f} \cdot t) \qquad (7)$$

We would expect this to make the entropy dips in higher values of n deeper. It should also move the deeper dip to the second harmonic, since that would be the closest harmonic to both existing sinusoids. The results are shown in Figure 4.

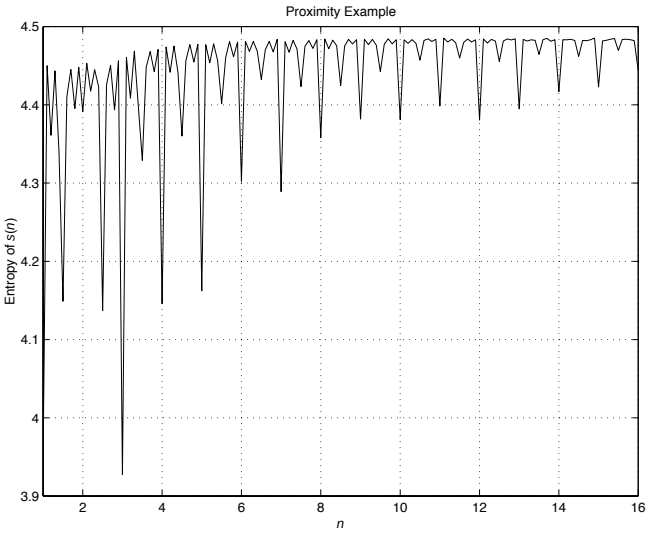

Figure 4. Entropy measurements for proximity experiment

3.4 Continuation

Continuation is another important cue in the auditory domain. Often a static† sound is temporarily interrupted either by silence or by another masking sound. Whether or not the original sound existed during that interruption or not, we still believe that the sound after the interruption is somehow linked the sound before. If the second section of the sound can be predicted or anticipated by the first one, that means that the two segments share common information, hence exhibit redundancy. In this case entropy is minimized and the sounds are thought of as one continuous sound (or at least as the same entity). If not, then the two sound segments are perceived as different sounds. We set up an experiment to see how we can track this using information theory.

We generate a sinusoid described by:

$$s(n) = \cos(f(t, n))) \tag{8}$$

where:

$$f(t, n) = \begin{cases} t < x_1 & f(t, n) = a \cdot t + b \\ t > x_2 & f(t, n) = a \cdot t + b + n \\ x_1 \le t \le x_2 & f(t, n) = 0 \end{cases} \tag{9}$$

This creates an upwards gliding sinusoid, which will interrupt for a set time period and them resume. Unless $n = 0$, it will resume at a position which we wouldn't project the original sound to be at, and the second part will be heard as another sound. If $n = 0$ however, the second part of the sound will be a possible continuation of the original sound and will be perceived as such. Figure 5 illustrates this scenario in the frequency domain.

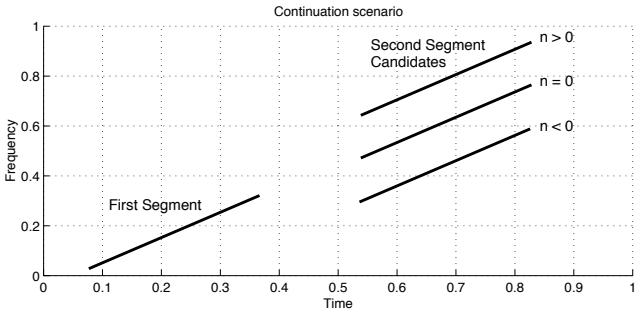

Figure 5. Continuation scenario. For values of *n* that are non-zero the alignment between first and second segments is off.

As in all preceding sections we measure the amount of entropy of $s(n))$ with respect to n. In this case we would expect that the entropy will be lowest for $n = 0$ where the two sinusoids align with each other. The results are shown in Figure 6.

In terms of computational methods that can detect this type of grouping we need very complicated algorithms that have to incorporate additional knowledge (such as possible frequency tracks etc.). The information-theoretic approach is bypassing the need for a model or a knowledge base by just observing the correlations of the two segments. Analogous results are obtained if during the interrupted period we place white noise rather than silence.

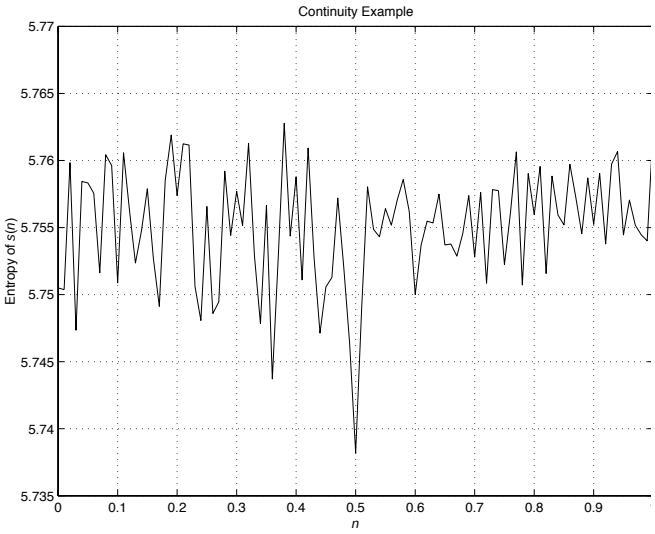

Figure 6. Entropy measurements for continuation experiment

3.5 Prägnanz, Contexts and Higher-order Gestalt principles

Due to the fact that gestalt principles were formulated to be general and applicable to all perceptual functions, there are a lot of special statistical dependencies that are not described. As an effort to include all of these, the prägnanz principle was used. The prägnanz principle states that "of several geometrically possible organizations that one will actually occur which possesses the best, simplest and most stable shape." [Koffka 1935]. Although it is arguable, descriptions such as simple, best and stable, tend to denote strong statistical dependencies. A square for example which would fall under the best/simple/stable description exhibits a lot of structure and redundancy, when compared to the complicated and unstable random polygon. This is of course a direct statement of this paper's argument. Simplicity, stability, predictability and order are features of low entropy systems.

Context is also deemed an important factor in grouping, and that too is an expression of dependence.

†. By static we refer to sounds with no dynamically changing parameters. Under this definition we can include sounds that have a linearly growing frequency, or a periodic amplitude modulation, as long as there is no higher order change in the parameters.

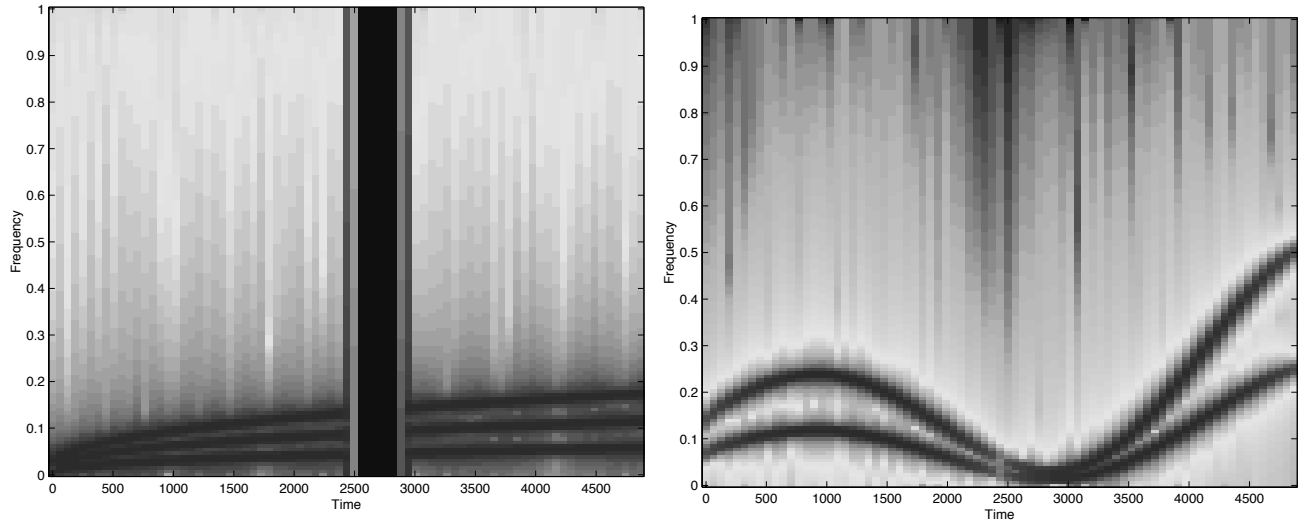

Figure 7. The two sets of sinusoids. The first set features a silence in the middle, the second is amplitude and frequency modulated by one period of a sine wave.

The correlations between history and background to a current situation are factors influencing our analysis of a scene. Like all correlations they would be encapsulated in entropy measurements.

3.6 Putting it all together

A fortunate feature of trying to perform grouping in such a manner is that we do not need to perform different measurements for each grouping criterion. Instead, with a simple measurement of entropy we could perform the same task more efficiently. A short example of how we could do this is presented in the following experiment.

We assume that we have collected a set of sinusoids by analyzing an auditory scene. In this case we will look at a set of five sinusoids comprising two sounds. The spectrograms of the two original sounds are shown in Figure 7.

All five sinusoids were submitted, untagged and isolated, to an algorithm which tried to find their grouping that would result in the least amount of entropy. Out of the 52 possible partitions that these sinusoids can form, the one with the lowest entropy was the one that formed the two original sounds, and the one picked by the algorithm. As the algorithm ran the only measurements were entropy values of different partitions. There were no pitch track, amplitude or on/offset estimates, something that reduces the complexity of the system and increases parallelism. The estimation of the optimal grouping could be done using either by an exhaustive search (which is fine up to about 9 sinusoids, after which the problem size explodes), or a combinatorial optimization algorithm such as simulated annealing. Simple versions of these approaches were used to obtain the results.

4.0 Conclusions

It was shown in this paper that gestalt grouping and entropy minimization are two closely related functions. The development of this approach was driven by three ideas.

First, the need to unify as many perceptual functions under one simple principle which has a clear and mathematical definition. The search for a compact answer to explain multiple phenomena is central to scientific research (not to mention, a form of entropy minimization!). Such abstract reasoning also has the added advantage of supporting multimodality in a very elegant manner. For example, integration of simultaneous visual and auditory gestalt would require the same principles presented in this paper, only generalized to higher dimensions.

Second, was the need to work with low level data so as to avoid preprocessing. Raw data contains a lot of information which, by definition is all original. Any form of preprocessing will produce artifacts and dependencies which were not originally present and could bias or complicate estimations. Rather than imposing a representation linked to an algorithm it is found more elegant to deal with measures such as information, which works on raw data and because of that is invariant throughout other perceptual domains. In fact the proposed model of dealing with a front-end providing sinusoids was only used in this paper to provide a link to the existing literature. By constricting our auditory atoms to be sinusoids we are biasing our approach and we deviate from the ideal set throughout this paper. Ideally we should expect do derive the form of the auditory atoms from statistical analysis similar in form to the entropy minimization we are performing. A possible atom set would be basis functions derived from Principal Component Analysis, Higher Order

Analysis or Independent Component Analysis as described by Bell [Bell 1996a]. Such an approach would be extremely desirable since the system would be build around one computational kernel, invariant of the domain of application.

Finally it is the author's belief that we need the foresight to look past digital computers. The programming principles that are imposed to us by the current generation of computers push us towards discrete and serial processing. Our body, on the other hand, is a massively parallel and continuous system. The methods that are proposed and used in this paper are very similar to natural processes, notably thermodynamic systems. It is our hope to derive first principles which are not bootstraped to today's computers but rather, more natural algorithms, which in the future might be easily and efficiently implemented in upcoming generations of computing machinery. By ensuring this link to natural processes we achieve not only more biological plausibility but also the prospect of much more elegant and efficient structures.

Although the main subject of this paper was an attempt to unify gestalt principles, the main idea which is hoped to be conveyed is the utility of statistical and information theoretic measures for perceptual computing. Similar work has been also done on pitch tracking and waveform identification by using the same principles, and has yielded favorable results. Although the work presented has been more in the form of combinatorial verification procedures, it does not preclude the use of entropy in more sophisticated ways. By imposing models it seems possible to construct systems to predict future events, perform segregation etc. in a very elegant manner. It is the author's hope that this paper might inspire such work.

5.0 Acknowledgments

The author would like to express his gratitude to the members of the machine listening group at the MIT Media Lab as well as to Tony Bell at Interval Research for stimulating this work through his discussions.

6.0 References

[Amari *et al.* 1996] Amari, S-I., A. Cichocki, and H.H. Yang. 1996. A new learning algorithm for blind signal separation. In *Advances in Neural Information Processing Systems* ***8***. MIT Press, Cambridge, MA.

[Atick and Redlich 1990] Atick, J.J. and A.N. Redlich. 1990. Towards a theory of early visual processing. *In Neural Computation* ***2***. pp. 308-320. MIT Press, Cambridge, MA.

[Attneave 1954] Attneave, F. 1954. Informational aspects of visual perception. *Psychological Review* ***61***, pp. 183-193.

[Barlow 1959] Barlow, H.B. 1959. Sensory mechanisms, the reduction of redundancy, and intelligence. In *National Physical Laboratory Symposium No. 10*, The Mechanization of Thought Processes.

[Barlow 1961] Barlow, H.B. 1961. Possible principles underlying the transformation of sensory messages, In *Sensory Communication*, W. Rosenblith, ed., pp. 217-234. MIT Press, Cambridge, MA.

[Barlow 1989] Barlow, H.B. 1989. Unsupervised learning. In *Neural Computation* ***1***, pp. 295-311. MIT Press, Cambridge, MA.

[Bell and Sejnowski 1995] Bell, A.J. and T.J. Sejnowski. 1995. An information maximization approach to blind separation and blind deconvolution. In *Neural Computation* ***7***. pp. 1129-1159. MIT Press, Cambridge, MA.

[Bell and Sejnowski 1996] Bell, A.J. and T.J. Sejnowski. 1996. The independent components of natural scenes. *Vision Research*. To appear.

[Bell and Sejnowski 1996a] Bell A.J. and Sejnowski T.J. 1996. Learning the higher-order structure of a natural sound, *Network: Computation in Neural Systems,* ***7***.

[Bregman 1990] Bregman, A.S. 1990. Auditory Scene Analysis, MIT Press, Cambridge, MA.

[Comon 1989] Comon, P. 1989. Independent component analysis - a new concept? In *Signal Processing* ***36***, pp. 287-314.

[Cooke 1991] Cooke, M.P. 1991. Modeling auditory processing and organization. Ph.D. thesis, University of Sheffield, Dept. of computer science.

[Deco and Obradovic 1995] Deco, G. and D. Obradovic. 1995. An Information Theoretic Approach to Neural Computing. Springer-Verlag.

[Duda *et al.* 1990] Duda, R.O., R.F. Lyon, and M. Slaney. 1990. Correlograms and the separation of sounds. In *Proceedings Asilomar Conference on Signals, Systems and Computers 1990.*

[Ellis 1992] Ellis, D.P.W. 1992. A perceptual representation of sound. Masters thesis, MIT EECS Department.

[Handel 1993] Handel, S. 1993. Listening, An Introduction to the Perception of Auditory Events. MIT Press, Cambridge MA.

[Koffka 1935] Koffka, K. 1935. Principles of Gestalt Psychology. Hartcourt, Brace, New York.

[Linsker 1988] Linsker, R. 1988. Self-Organization in a perceptual network. In *Computer* ***21*** (March), pp. 105-117.

[Mellinger 1991] Mellinger, D.K. 1991. Event formation and separation in musical sound. Ph.D. thesis, CCRMA, Stanford University.

[Redlich 1993] Redlich, A.N. 1993. Redundancy reduction as a strategy for unsupervised learning. *Neural Computation* ***5***. pp. 289-304. MIT Press, Cambridge, MA.

[Shannon and Weaver 1963] Shannon, C. and W. Weaver. 1963. "The Mathematical Theory of Communication" University of Illinois Press.

[Vercoe and Cumming 1988] Vercoe, B. and D. Cumming. 1988. Connection machine tracking of polyphonic audio. In *Proceedings of International Computer Music Conference 1988.* pp. 211-218.