Proceedings of the Institute of Acoustics

COMPUTATIONAL AUDITORY SCENE ANALYSIS: MODELLING THE COMPETITION BETWEEN ALTERNATIVE PERCEPTUAL ORGANISATIONS

D J Godsmark & G J Brown

University of Sheffield, Department of Computer Science, Sheffield, UK

1. INTRODUCTION

Bregman (1990) contends that the mixture of sounds reaching the ears is subject to an *auditory scene analysis*, in which acoustic elements likely to have arisen from a common environmental source are grouped in a perceptual representation termed a *stream*. Furthermore, Bregman contends that this grouping process is performed in accordance with the perceptual principles described by the Gestalt psychologists (e.g. Koffka, 1936). Such principles include grouping on the basis of temporal and frequency proximity (Bregman & Campbell, 1971; Van Noorden, 1977; Bregman, 1978), common onset (Dannenbring & Bregman, 1978; Darwin, 1981; Ciocca & Darwin, 1993), harmonicity (Darwin, 1981; Darwin & Gardner, 1988) and similarity of timbre (Wessel, 1979; Singh, 1987).

Given the number of organisational principles which appear to be employed by the auditory system, it is reasonable to expect that situations exist where such principles will suggest conflicting organisations. For example, temporal proximity may suggest that two tones should be grouped, whilst frequency proximity suggests that they be segregated. This implies the existence of a mechanism capable of combining the evidence from a variety of organisational principles in order to select which of the potential organisations will be perceived. This paper presents a consideration of the properties of this mechanism, along with a computational model of auditory organisation which demonstrates many of these properties.

2. PROPERTIES OF AN EVIDENCE COMBINATION MECHANISM

Whilst direct experimental evidence concerning the exact nature of an evidence combination mechanism is limited, the psychoacoustic literature contains many investigations which indicate the basic properties such a mechanism would need to possess. This literature is briefly reviewed below.

2.1 Grouping Processes

Bregman (1990) argues that primitive (innate) auditory organisation is the result of two basic processes of organisation; sequential (horizontal) and simultaneous (vertical). Simultaneous processes organise auditory elements which occur simultaneously, but in different regions of the spectrum, whereas sequential processes organise elements which succeed each other in time. Again, it is to be expected that there will be situations where sequential and simultaneous grouping processes suggest alternative organisations. Such conflicts have been documented by several workers (e.g. Dannenbring & Bregman, 1978; Steiger & Bregman, 1981, 1982). This implies that any organisational mechanism cannot treat simultaneous and sequential grouping individually; they must interact cotemporaneously in order to conflict.

2.2 Streaming Phenomena

Auditory streaming is the phenomenon whereby a sequence of strictly consecutive tones of differing frequencies perceptually separates into two or more streams. The impression of streaming grows stronger if the sequence is presented rapidly, or contains large

frequency separations. One important observation concerning streaming is that the strength of the streaming percept accumulates as the sequence is presented (Bregman, 1978). Typically, the sequence is initially perceived as a coherent whole, with the percept of two or more streams emerging gradually.

2.3 Effect of Context

There have been a number of experiments which demonstrate that auditory organisation is highly context dependant. For example, Bregman & Rudnicky (1975) found that identifying the order of two tones (A and B) became difficult if they were surrounded by two flanking tones F (i.e. FABF). However, if the flanking tones were embedded within a sequence of similar frequency captor tones C (i.e. CCCFABFCCC) it becomes easier to identify the order of tones A and B.

Bregman & Rudnicky argue that this is due to the captor tones capturing the flanking tones in a separate stream, on the basis of frequency proximity. Thus tones A and B are isolated in a second stream, making their identification easier. This implies that the decision to group or segregate the flanking tones and the target tones is dependant not just on the relationship *between* the target and flanking tones, but also on the *context* in which they appear. Bregman (1978b) demonstrated similar results.

The auditory system also appears to employ a "wait and see" attitude to organisation; the organisation imposed upon elements at a particular moment in time can be modified by elements arriving at a future time. A demonstration of this was presented by Bregman & Tougas (1989). They found that when presented with the stimulus shown in figure 1a, subjects reported that tones B and C tended to fuse.

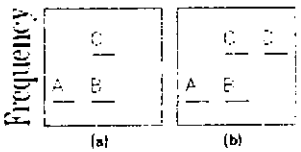


FIGURE 1: Stimulus patterns used by Bregman & Tougas (1989) to demonstrate constraint propagation.

However, when presented with the stimulus shown in figure 1b, subjects reported a greatly reduced tendency for tones B and C to fuse. Thus the presence of tone D has a retroactive effect on the fusion of tones B and C. Bregman & Tougas explain this in terms of a *constraint propagation* mechanism. The grouping of tones C and D on the basis of frequency proximity reduces the fusion of tones B and C, which strengthens the grouping of tones B and A and so on. Obviously, this propagation must occur over a limited time period; otherwise constraints would be propagated back indefinitely and a stable percept would never emerge.

It has also been demonstrated that a harmonic can be "heard out" if it terminates asynchronously with the remaining harmonics of a complex tone (Dannenbring & Bregman, 1978). The decision to segregate the harmonic must be made retroactively, following the harmonics termination. A similar argument extends to the *perceptual restoration* of a signal obscured by a loud masker (Bregman, 1990). Such restoration is dependant upon the sensory evidence present both before and *after* the masker, implying that the restoration is retrospective.

Thus it would appear that both perceptual restoration and constraint propagation rely upon the operation of a retroactive mechanism. Assuming that they rely on the *same*

mechanism, the maximum duration of noise through which perceptual restoration occurs should indicate the time period across which constraints are propagated. Perceptual restoration has been observed through noise bursts ranging from 50 to 350 msec (e.g. Ciocca & Bregman, 1987; Kluender & Jenison, 1992).

2.4 Properties of the Evidence Combination Mechanism
In summary, there are a number of important factors concerning auditory organisation and evidence combination:

- Primitive auditory organisation is performed by two main grouping processes - simultaneous and sequential - which can suggest conflicting organisations.
- Each of these processes employs a number of grouping principles, each of which can also suggest conflicting organisations.
- The mechanism is context-sensitive and accumulates evidence over time.
- The organisation of auditory elements can be modified by elements arriving at a future time.

3. OVERVIEW OF THE MODEL

3.1 Previous Models of Auditory Organisation
A number of computational models of auditory organisation have been described, many of which were intended to either segregate speech from interfering noise (e.g. Scheffers, 1983; Brown, 1992; Cooke, 1993) or to segregate melodic lines from polyphonic music (Mellinger, 1991; Kashino & Tanaka, 1992; Brown & Cooke, 1994). However, few of these models have addressed the problem of integrating the evidence presented by contradictory grouping principles. Regardless of their ability to segregate sounds, these models would be incapable of emulating the organisation of simple pure tone sequences.

One exception is the model by Kashino & Tanaka, which attempts to directly integrate evidence from psychoacoustic experiments. The model supports the integration of evidence from two grouping principles, harmonicity and onset asynchrony. However, the model does not consider the context in which tones appear, and does not accumulate evidence over time.

Similarly, there have been a number of models aimed at reproducing the results of auditory experiments, such as Williams (1989) STREAMER - which takes a high-level symbolic approach to simulating the organisation of pure tone stimuli - and the peripheral channelling model presented by Beauvois & Meddis (1991). Again, these models do not consider the context in which an element is presented, nor do they exhibit any retroactive behaviour.

3.2 Modelling Constraint Propagation
The primary objective behind the development of this model was to demonstrate a mechanism capable of combining a number of grouping principles within a constraint propagation framework. As argued previously, the constraint propagation mechanism must operate within a finite time period. This is modelled using a temporal window, which feeds information to a propagation mechanism.

3.2.1 The Temporal Window. The temporal window is a sliding rectangular window of finite width. The organisation of auditory events can only be modified while they are contained within the temporal window. Once an event has departed from the temporal window, a permanent organisation is imposed.

In the current implementation, there is a single temporal window of fixed (350 msec) width. However, certain grouping principles - in particular those concerned with emergent properties such as pitch and timbre (e.g. Krumhansl & Iverson, 1992) - operate over a significantly wider time scale. This suggests that either the width of the temporal window is variable, or there a number of temporal windows with different widths. A computational scheme which adopts the latter approach is discussed in detail in Godsmark (1994).

3.2.2 The Propagation Mechanism. Currently, the model employs a symbolic representation of sequences of tones, consisting of start time, duration and frequency information. As tones enter the temporal window, every possible interpretation of these tones is created and added to a structure called the *scene interpretation tree* (SIT). Consider, for example, a consecutive sequence of three tones, labelled A,B and C. When tone A enters the temporal window, there is only one possible interpretation. When tone B enters the temporal window, it can be grouped with tone A or segregated. The arrival of tone C suggests a number of possible interpretations. Figure 2 shows the SIT after the three tones have arrived.

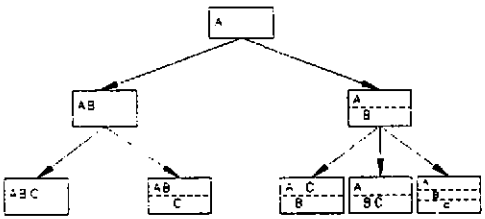


FIGURE 2: An interpretation tree for three tones A,B and C.

Each interpretation in the SIT is evaluated by a range of grouping "experts". This evaluation is entirely local as only the relationship between the most recently added tones is considered. In the case of the example given above, the root of the SIT receives a score of 0, as there is nothing to evaluate. The second level of the tree will be scored based on the relationship between tones A and B. Finally, the lowest level of the tree will be evaluated on the basis of the relationship between tones B and C.

The experts score the interpretations on the basis of how well they abide by a particular principle. For example, if tones B and C are close in frequency, a frequency proximity expert will give a high score to any interpretation with tones B and C grouped, and a low score to any interpretation with tones B and C isolated in different streams. The local score for any node is the combination of the scores of all the grouping experts. Presently, these scores are combined through simple addition.

In addition to this local score, each interpretation also has a global score. This global score is its own local score combined with the local scores of *all* the interpretations derived from it. Thus when a new tone is added, all possible interpretations of it are added to the SIT. The local scores of these new interpretations are propagated up the tree and combined with the global score of the interpretation from which they were grown. This global score is in turn propagated back up the tree to the previous level, and so on. Eventually, this propagation process will modify the global scores of the nodes at the root of the tree. Thus the score of the root nodes will be dependant upon every tone within the temporal window.

As tones exit the temporal window, the root node with the highest *global* score is selected and the tones are organised according to that interpretation. All alternative interpretations

are then pruned from the SIT. As this global score is highly context dependant, the chosen organisation is a direct result of the context of tones occurring at a future date; thus the mechanism embodies both context sensitive grouping, and a retroactive component.

3.3 Emergent Properties

An interesting emergent property of this approach is that, for a simple tonal sequence ABAB presented rapidly enough to invoke streaming, the global score of the root node continually increases over time. If the magnitude of the global score is taken as a measure of the strength of the percept, then this simulates the percept of streaming continually increasing over time.

A further property of this approach is that it inherently incorporates competition between individual grouping principles, and between grouping processes. As the SIT contains all possible interpretations, every possible conflict will be considered, and automatically resolved by the selection of the root node according to its global score. Additionally, this mechanism is capable of incorporating as many grouping principles as required. As competition between principles is handled implicitly by the propagation mechanism, all that is required to add a new principle is an evaluation metric to produce the local score.

4. RESULTS

4.1 Evaluating the Model

The model has been evaluated on the basis of direct comparison with psychophysical findings. In particular, psychophysical experiments explicitly demonstrating either competition or constraint propagation were selected. Half of these experiments were used to "train" the three grouping experts currently implemented (frequency proximity, temporal proximity and onset asynchrony), and the remainder to evaluate the models performance.

4.2 Example: Bregman & Tougas

One of the most important experiments used to evaluate the model was that by Bregman and Tougas (1989). As previously discussed (section 2.3), they presented their subjects with a three or four tone repeating sequence (see figure 1). If tone D is absent, tones B and C tend to fuse; if tone D is present tones B and C tend to segregate. Figure 3 shows the SIT after the first occurrence of tones A,B and C have been presented to the model.

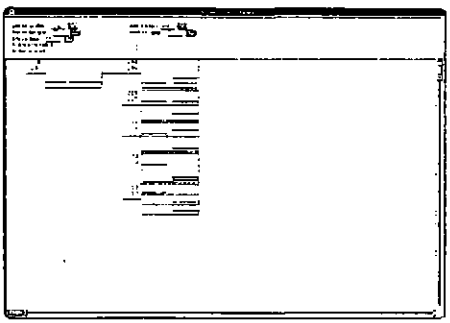
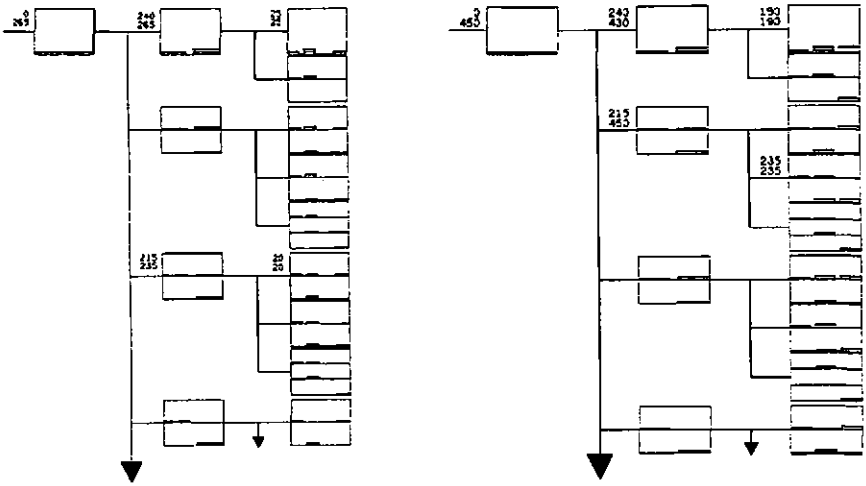


Figure 3: The SIT when tones A,B and C are within the temporal window. Each interpretation is displayed as a miniature sketch.The numbers to the left of each interpretation represent the local (upper) and global (lower) score. A horizontal bar across the centre of the interpretation represents a stream.



(a) (b)

FIGURE 4: SIT after (a) the next repetition of tone A has entered the temporal window if tone is absent and (b) tone D is present and has entered the temporal window.

It can be seen that the interpretation with B and C fused is currently regarded as superior (a score of 240) than the interpretation with tones B and C segregated (215). Assuming tone D is absent, the next tone to enter the temporal window will be the next repetition of tone A, resulting in the SIT shown in figure 4a. In this case, the fusion of B and C is still clearly favoured (a score of 265) over segregation (a score of 235).

In contrast, if tone D is present, it will be the next tone to enter the temporal window, resulting in the SIT shown in figure 4b. Now the segregation of tones B and C (a score of 450) is quite clearly favoured over fusion (a score of 430). The arrival of tone A further strengthens the segregation interpretation.

In both cases, because the sequence is repetitive the local scores will continually cycle through a limited set of values, increasing the global score of both interpretations by a constant amount every repetition. However, in each case the interpretation initially favoured is being strengthened the most. For example, in figure 4a, the segregation condition is being strengthened by 20, while the fusion condition is being strengthened by 25. Thus the favoured interpretation will steadily grow much stronger, demonstrating the cumulative aspects of streaming.

5. CONCLUSIONS

The model described above is capable of explaining a range of organisational phenomena. In particular, the model is capable of simulating both context-sensitive organisation in a retroactive manner, and the gradual cumulation of the streaming percept. It is also inherently capable of combining as many grouping principles as are

required without modifying the central mechanism. It thus presents an extremely useful experimental tool for evaluating the contribution made by the various grouping principles, and the relative importance of context.

However, in its current form, the model is unsuitable for use as a practical application. The symbolic representation employed by the model is capable of describing only tonal sequences consisting of stable pure tones, and the SIT exhibits a high degree of combinatorial explosion. Even with simple sequences consisting of tones with just four or five harmonics, the SIT typically contains several thousand nodes.

Development has recently commenced on a more efficient model capable of accepting much more complex stimuli. The new model features a computational model of the auditory periphery and thus accepts digitised sound in place of the symbolic representation employed here. Also, the backward propagation mechanism of the current model has been replaced with a functionally equivalent forward propagation mechanism. With this new mechanism, the SIT *never* grows beyond a single level. It is expected that the new mechanism will typically consider a maximum of several dozen interpretations, as opposed to several thousand in the current implementation. This model is described in detail in Godsmark (1994).

In summary, it is felt that the mechanism of propagation suggested here presents a natural framework for the inclusion of top-down (*schema-driven*) grouping principles. Many top-down principles can be integrated by strengthening interpretations in the SIT which are favoured by these top-down principles. Once again, competition between schema-driven and primitive principles is implicitly modelled by the propagation mechanism, and all that is needed is an evaluation metric for the local score.

6. ACKNOWLEDGEMENTS

DJG is supported by a grant from the BBSRC. GJB is upported by SERC grant GR/H53174 and the Nuffield Foundation.

7. REFERENCES

[1] M W BEAUVOIS & R MEDDIS, 'A computer model of auditory stream segregation', *Quart J Exp Psych*, 43A(3) pp517-541 (1991)

[2] A S BREGMAN, 'Auditory streaming is cumulative', *J Exp Psychology: Human Perception and Performance*, 4(3) pp380-387 (1978)

[3] A S BREGMAN, 'Auditory streaming: competition among alternative organisations', *Perception & Psychophysics*, 23(5) pp391-398 (1978b)

[4] A S BREGMAN, 'Auditory Scene Analysis', MIT Press (1990)

[5] A S BREGMAN & J CAMPBELL, 'Primary auditory stream segregation and perception of order in rapid sequences of tones', *J Exp Psychology*, 89(2) pp244-249 (1971)

[6] A S BREGMAN & A RUDNICKY, 'Auditory segregation: stream or streams?', *J Exp Psych: Human Perception and Performance*, 1 pp263-267 (1975)

[7] A S BREGMAN & Y TOUGAS, 'Propagation of constraints in auditory organisation', *Perception & Psychophysics*, 46(4) pp395-396 (1989)

[8] G J BROWN, 'Computational auditory scene analysis: A representational approach', Unpublished PhD Thesis, Sheffield University (1992)

[9] G J BROWN & M P COOKE, 'Perceptual grouping of musical sounds: A computational model', *J New Music Research*, 23 pp107-132 (1994)

[10] V CIOCCA & A S BREGMAN, 'Perceived continuity of gliding and steady-state tones through interrupting noise', *Perception and Psychophysics*, 42(5) pp476-484 (1987)

[11] V CIOCCA & C J DARWIN, 'Effects of onset asynchrony on pitch perception: Adaptation or grouping?', *JASA*, 93(5) pp2870-2878 (1993)

[12] M P COOKE, 'Modelling auditory processing and organisation', Cambridge University Press (1993)

[13] G L DANNENBRING & A S BREGMAN, 'Streaming vs. fusion of sinusoidal components of complex tones', *Perception & Psychophysics*, 24(4) pp369-376 (1978)

[14] C J DARWIN, 'Perceptual grouping of speech components differing in fundamental frequency and onset-time', *Quarterly J Exp Psychology*, 33A pp185-207 (1981)

[15] C J DARWIN & R B GARDNER, 'Perceptual segregation of speech from concurrent sounds', in M E H Schouten & M Nijhoff (eds.) 'The psychophysics of Speech Perception', pp112-124 (1988)

[16] D J GODSMARK, 'Computational listening: A cognitive model of the perceptual organisation of polyphonic music.', *Computer Science Report (in press)*, University of Sheffield (1994)

[17] K KASHINO & H TANAKA, 'A sound source separation system using spectral features integrated by the Dempster's law of combination', In *Annual Report of the Engineering Research Institute, University of Tokyo, Faculty of Engineering*, pp67-72 (1992)

[18] K R KLUENDER & R L JENISON, 'Effects of glide slope, noise intensity, and noise duration on the extrapolation of FM glides through noise', *Perception and Psychophysics*, 51(3) pp231-238 (1992)

[19] K KOFFKA, 'Principles of Gestalt psychology', New York: Harcourt, Brace and Co. (1936)

[20] D K MELLINGER, 'Event formation and separation in musical sound', Unpublished PhD Thesis, Stanford University (1991)

[21] L P A S VAN NOORDEN, 'Minimum differences of level and frequency for perceptual fission of tone sequences ABAB', *JASA*, 61(4) pp1041-1045 (1977)

[22] M T M SCHEFFERS, 'Sifting vowels: Auditory pitch analysis and sound segregation', Unpublished PhD Thesis, University of Gröningen (1983)

[23] P G SINGH, 'Perceptual organization of complex-tone sequences: A tradeoff between pitch and timbre', *JASA*, 93(6) pp3374-3389 (1987)

[24] H STEIGER & A S BREGMAN, 'Capturing frequency components of glided tones: frequency separation, orientation and alignment', *Perception & Psychophysics*, 30(5) pp425-435 (1981)

[25] H STEIGER & A S BREGMAN, 'Competition among auditory streaming, dichotic fusion and diotic fusion', *Perception & Psychophysics*, 32(2) pp153-162 (1982)

[26] D L WESSEL (1979), 'Timbre space as a musical control structure', *Computer Music Journal*, 3(2) pp45-52 (1979)

[27] S M WILLIAMS, 'STREAMER: A prototype tool for computational modelling of auditory grouping effects.', *Computer Science Report CS-89-31*, University of Sheffield (1989).