Proceedings of the Institute of Acoustics

STREAMER: MAPPING THE AUDITORY SCENE

S M WILLIAMS(1,2), R I NICOLSON (2) & P D GREEN(1)

(1) Department of Computer Science
(2) Department of Psychology
University of Sheffield, SHEFFIELD, ENGLAND, S10 2TN.

ABSTRACT

Bregman (e.g. [6]) has argued that the ecological function of audition is to perform an 'auditory scene analysis' (ASA), that is, to decompose the complex and noisy acoustic environment into 'streams' corresponding to separate acoustic events — sources of sound. In this position paper we apply Marr's [18] metatheoretical framework to audition, suggesting that Bregman has identified correctly the 'computational problem' of audition as ASA but that the data available are not sufficient to discriminate between different algorithms for solving the computational problem. Our proposed research has two inter-related themes. First, the collection of a large corpus of psychophysical data, using the MITSYN [14] sound synthesis system to reconstruct ASA demonstrations which will then be extended by varying the parameters systematically. The data thus obtained should then inform the development of computational models of the processes involved. It is envisaged that a computational implementation of these processes will be based around the STREAMER prototype [24], using the CLOS object-oriented programming environment [15].

1. AUDITORY SCENE ANALYSIS

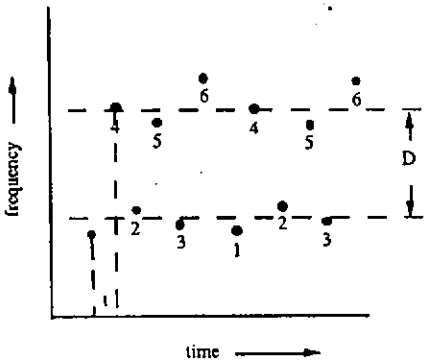
Bregman [6] has argued persuasively that the function of an organism's auditory system is to identify the various sound sources that make up the complex and noisy acoustic environment so that the organism may attend selectively to those sources which are of particular significance to it. Given the evolutionary advantages of such a system, it seems inevitable that an animal's auditory system would have developed to achieve this function.

He terms the process involved 'auditory scene analysis' [5, 6] and uses the term 'stream' to denote the representation of the time course of each of the acoustic events which together constitute the acoustic environment. The phrase 'acoustic event' is preferred to 'sound' in this context to denote a single experienced event which may extend over time and consist of many separate sounds, such as a series of footsteps. The function, then, of audition is to decompose the incoming sound train into its constituent streams, from which the hearer can select significant streams to "derive a useful representation of reality" (Bregman, [6], p3.) from them.

STREAMER: MAPPING THE AUDITORY SCENE

Bregman cites three classic demonstrations of the streaming process;

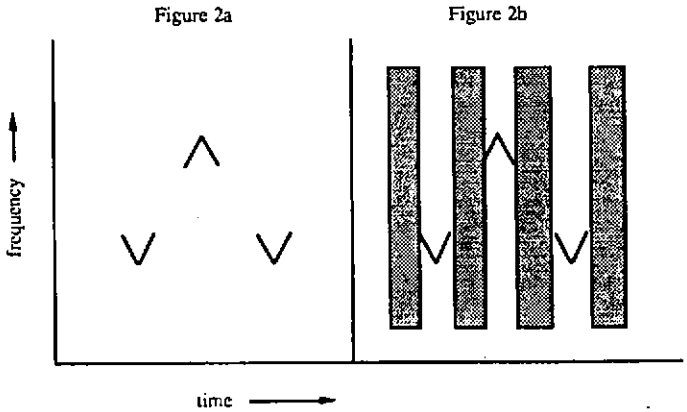
Figure 1. The Auditory Streaming Effect



First the 'auditory streaming effect' [1]. If a sequence of 6 tones (3 high, 3 low) is played repeatedly (see Fig. 1), in due course they appear perceptually as 2 streams - a low frequency stream (123123123) and a high frequency stream (456456456) even though there is no physical basis on which to separate them. The streaming effect increases with the separation between the two frequencies (D) and increases as the temporal interval (t) between the tones decreases.

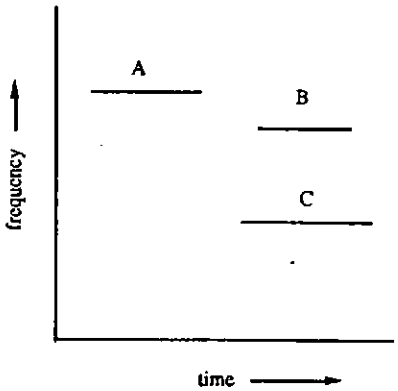
The second streaming effect [9] is known as the 'continuity illusion'. In Fig. 2a the three sound glides are heard separately, as one would expect.

Figure 2. The continuity illusion



However, when bands of masking noise are inserted into the gaps (fig. 2b) the three are heard as one continuous sound.

Figure 3. Perceptual Decomposition



The third streaming demonstration comes from the perceptual decomposition of simultaneous sounds. A repeating sound made up of three pure tones A, B and C (see figure 3) is ambiguous in that it can be heard either as a pure tone A followed by a complex tone B+C, or as a stream A+B plus a pure tone C. Bregman and Pinker [3] showed that the probability of the former was increased by increasing the onset and offset synchrony of B and C, and by increasing the frequency difference between A and B.

1.1. Gestalt Laws of Grouping -vs- Scene Analysis.

Interestingly the three streaming phenomena discussed above may also be explained by the Gestalt Laws of Grouping. Gestalt psychology, the study of the laws of organisation, was developed in the early part of this century, mainly by the German psychologists, Wertheimer, Kohler and Koffka, at Frankfurt. It attempts to explain human perceptual recognition and development of understanding by predicating that objects exist as organised "wholes" which have a definite structure or form and by hypothesising an inherent human propensity to recognise or identify this structure. Key Gestalt principles included those of 'proximity' and 'similarity', which predicts that the closer two components are and the more similar they are, the more likely they are to belong to the same unit, and that of 'good continuation', which requires that, in order for sequential components to be assigned to the same object, there should be smooth transitions between any changes of state which occur. Gestalt grouping concepts of 'similarity' [23], 'proximity' [7], 'good continuation' [3] 'habit' or 'familiarity' [19, 11], 'belongingness' [2, 4] and 'common fate' [13], first demonstrated in vision research have now also been shown in auditory perception.

The auditory streaming effect is an example of the Gestalt Proximity Principle (the perceptual grouping increases as the elements get closer together); the continuity illusion demonstrates the Gestalt effects of Closure and Belongingness, and the perceptual decomposition effects may be accounted for by the Gestalt Principle of Common Fate. Bregman notes (as did Marr, [18]) that the Gestalt grouping processes are merely a description of phenomena and in no way offer an explanation as to how or why they arise in ASA. By contrast the scene analysis perspective suggests that the Gestalt grouping processes are surface manifestations of the underlying heuristics. Thus the auditory streaming effect is caused by the 'principle of exclusive allocation' - a scene element may only be assigned to one description at a time; the continuity illusion results from a masking compensation system using 4 heuristics (which Bregman [6] (p.663) terms the 'no discontinuity in A rule', the 'sufficiency of evidence' rule, the 'A1 - A2 grouping' rule, and the 'A is not B' rule); and the perceptual decomposition effects derive from the 'synchronicity principle', the 'harmonicity principle' and the 'old plus new heuristic' ([6], p.656).

STREAMER: MAPPING THE AUDITORY SCENE

1.2. Auditory Scene Analysis: A Marrian Perspective

Surprisingly despite his comprehensive coverage of audition, the many analogies he draws with vision, and the clear parallels between his framework and that of David Marr (eg. [18]) Bregman makes no reference to Marr's immensely influential analysis of the processes involved in early vision. In brief, Marr identified three different levels of explanation, the level of the 'computational problem' which specifies the task to be performed by the system,and the knowledge that the system has available to solve the problem; the 'algorithm level', which specifies the algorithm used to solve the computational problem; and the 'implementation level' which specifies the neural hardware used to implement the algorithm. Marr [17] claimed that many problems then current in vision research could be attributed primarily to failure to distinguish clearly between these levels of explanation, and in particular to attempts to investigate hardware and algorithms in the absence of a specification of the underlying computational problem. Bregman highlights an important gap within Marr's analysis by introducing the crucial distinction between 'primitive' (innate) processes and 'schema-based' (learned) processes in sound analysis. Marr's analyses were confined to primitive processes. Bregman argues ([6] p.669) that primitive processes are responsible for the Gestalt grouping phenomena and automatically partition the evidence. By contrast schema-based processes can act under attentional control and select from the evidence without partitioning it. It seems likely that the principles underlying their operation are radically different and yet, as Bregman concedes, both processes are normally inextricably interwoven in most ASA, including speech analysis.

We argue, first, that although much audition research has tended to concentrate on speech analysis, it may also be valuable to isolate and analyse the primitive processes, in that understanding of the principles of operation of the primitive processes may allow a dissociation of primitive from schema-based processes in speech perception. Second, we suggest that Bregman has, in effect, produced a specification for the computational problem in auditory speech analysis, both in terms of the problem to be solved and the knowledge that may be exploited. The next stage in explanation of auditory scene analysis is therefore investigation of the algorithms used to solve the computational problems. This analysis provides the motivation for our current research programme, but before discussing our research plans it is valuable to outline how the project fits in with ongoing research in Sheffield.

2. RELATED AUDITORY RESEARCH AT SHEFFIELD

The interdisciplinary research planned complements three existing research themes within the SPLASH speech research group at Sheffield. In particular, Marr's analysis has provided a framework for a representational approach to speech processing; starting with the 'Speech Sketch' [12], from which the Auditory Speech Sketch (ASS) [8] has developed. Current work in the ASS may be seen as complementary to the planned modelling of primitive auditory grouping processes. The STREAMER model [24], developed within the ASS research programme, was an explicit attempt to model Gestalt grouping processes, and has provided much of the motivation for our current analyses of streaming. Third, Simons and Green [22, 12] have consistently advocated the use of OOP techniques for computational modelling. Streams may be represented as objects, and hence the OOP methodology appears to provide a powerful but natural modelling system. We outline each theme in turn.

STREAMER: MAPPING THE AUDITORY SCENE

2.1. Auditory Speech Sketch

Since 1986, the speech research group in the Department of Computer Science (SPLASH) has been developing an approach to the computational modelling of speech perception which draws inspiration from Marr's analysis. Following Marr, it was hypothesised that low-level speech processing may be analysed into a hierarchy of processes, each effecting a representational transformation which makes explicit the objects required at the next level, thus building a structure called the 'Speech Sketch' (SS). In sharp contrast with the 'rush to recognition' characteristic of conventional speech technology, the argument is that one should seek to interpret the *objects* in the SS [12], rather than the original parameter values.

Although early Speech Sketch implementations were based on spatial analysis of the conventional spectrogram, a *principled* approach to speech processing should appeal to the auditory system rather than the visual: there should be an 'Auditory Speech Sketch'. The ASS attempts to model grouping phenomena studied in neurophysiology and psychophysics in order to construct low-level semi-symbolic auditory representations of speech events. The ASS is reported on elsewhere in these Proceedings [8], so an outline will suffice here. The lower levels of the ASS transform the acoustic data using a model of the auditory periphery, and make explicit the perceptually important representations of synchrony, onsets and modulation of the objects in the sound train. Higher level representations may then be derived via harmonic constraints which make explicit stream-related information such as time-frequency representations of synchronous activity.

2.2. STREAMER model

A prototype computational model, STREAMER, [24], developed within the ASS research, attempts to portray the principles which lead to identification of relationships between primitive auditory objects, and, on the basis of these relationships, to assign the objects to independent streams for interpretation. These objects are assumed to be contiguous segments of simple sounds, having single frequency and amplitude values for each moment in time, such as pure tones. The model is currently capable of demonstrating the action of several grouping principles across the independent auditory objects which make up a particular 'scene'. Relationships identified during the grouping are made explicit through the creation of 'group' objects which contain information about the successful grouping method which gave rise to the grouped object, as well as listing the primitive auditory objects involved in the group. STREAMER can also indicate the hierarchical nature of the structures which develop from combining information from the group objects derived in the first stage, showing how conflicts arise during this process.

2.3. Object Oriented Programming

Object-oriented programming systems support the design of programs which focus on the abstract properties of objects, by making it easy to represent relationships among classes of objects and providing a flexible means of sharing, or inheriting, structure and data [15]. A class is a description of a set of mutually similar objects which defines the structure of the objects and their shared values and behaviours. Classes are organised into inheritance hierarchies so that classes lower in the hierarchy inherit from their superclasses, all the classes on the ascending pathway to the top (or root) of the network. A class may inherit both variables and methods (the algorithms which implement the behaviours of the objects which are its instances) from its superclasses [21].

We argue that OOP techniques should be particularly powerful for streaming analyses (and for other speech analyses) owing to the support they provide for object manipulation [22]. Nonetheless, modelling of auditory streaming using OOP raises a number of intriguing problems, which will also be addressed by the research programme.

3. RESEARCH PLANNED

As discussed earlier, the twin objectives of the research programme are to investigate which algorithms are capable of solving the auditory streaming computational problem and to attempt to build computational models which implement plausible algorithms. Unfortunately, while there is an abundance of reported research to demonstrate that auditory streaming effects exist ([6], p.48-639!), few systematic attempts have been made to map their interactions across auditory space at the level of detail required for a computational model. Further, unlike the three dimensions of physical space onto which visual perception must be mapped, the three dimensions along which an acoustic signal may vary — time, frequency and amplitude (or energy level) — are each measured against a different scale, making comparisons between intervals along these dimensions less straightforward.

Consequently an early objective of the research is to replicate in detail several of the classic streaming experiments, using a data collection technique in which the parameters known to affect the phenomena are varied systematically in small steps across the critical regions, thus providing a rich data lattice which should support detailed curve fitting techniques for comparative evaluation of different computational models. In this set of experiments only those streaming effects thought to be 'primitive' will be investigated, in the expectation that the data will therefore be relatively uncontaminated by the effects of learning.

3.1. Psychophysical Studies

Experimental stimuli will be produced using the MITSYN sound synthesis system [14], and we are indebted to Al Bregman and his team for their support and advice on implementing the system and their streaming methodology in Sheffield. It is intended to replicate the three streaming effects described earlier, but it may be useful to provide a more detailed plan for the third effect, the perceptual decomposition of simultaneous sounds.

3.1.1. Synchrony Mapping. This represents a systematic extension of the technique introduced by Dannenbring and Bregman [10] who devised a series of experiments to explore the relative importance of the onset and offset asynchronies of single components in 3-tone harmonic complexes. They selected 3 alternative onset lead or offset lag times which were 0, 35 or 69 msecs. relative to 137 msec tones (a lead or lag of 0 meant that the onset or offset was concurrent with that of the other two tones in the complex) and compared the effects of these by using a captor tone alternating with the 3-tone complexes. The tones were selected at octave intervals on a 500Hz fundamental. For the purposes of our model we need to determine whether the effects of asynchrony of onset and offset times are related to, or independent of, either the length of concurrent tones or the length of the intervals between presentations of the tones. We also need to know to what extent the fusion relies on the harmonicity of the concurrent components. This package aims to define the relationship between strength of fusion and degree of asynchrony in either absolute or relative terms, as appropriate. We anticipate that this will also be the basis for a further investigation into synchrony effects in speech data where small asynchronies in formant onsets are significant for interpretation purposes without apparently decomposing the speech stream.

3.1.2. Frequency Proximity. Here we will extend the work of Bregman and Pinker [3] introduced above. We will attempt to ascertain the independent contributions of time-synchrony and frequency-proximity by developing a set of stimuli in which the synchrony of the tone-pair is held as a constant but many more frequency values are tested for the captor tone. Combining this

range of captor values with two-tone pairs covering a range of frequency disparities (including some simple harmonics) should enable us to study the frequency proximity effect in isolation, providing sufficient data to deduce approximate mathematical functions which will represent this fundamental grouping effect in our model.

3.1.3. Relative Amplitude Mapping. Here the aim is to map out the relationships between amplitude and frequency which lead to perceptual fusion or fission of concurrent pure tones in both harmonic and inharmonic groupings. The experiments by Dannenbring and Bregman [10] described above demonstrated that an overriding factor in the onset/offset asynchrony effects was that of maintaining the relationship between amplitude and frequency of the component tones to that which would occur if the components had been generated naturally as harmonics of a sound generated from a single source (i.e. with amplitude decreasing for the higher frequency components). However, despite the speech signal always being generated from a single source, its complexity results in the perceptual prominence of different components varying from that predicted by the Dannenbring results, again without apparent decomposition of the speech stream. We need to devise variations on these experiments to determine the conditions which permit higher frequency components (higher formants) to have higher amplitude without loss of cohesion with the rest of the speech stream.

3.2. Computational Modelling

The objective of the computational modelling stages is to explore different methods and algorithms for solving the scene analysis problems. We expect that the object-oriented programming techniques used in the ASS and in STREAMER should prove particularly valuable in the modelling process. However it is important to test a range of algorithms and approaches. For instance, we intend to explore the use of relaxation techniques used for analogous problems in vision research. The restriction of OOPS themselves may be imposing constraints on how we analyse the scenes and an investigation of this aspect should form part of the computational modelling component of the research.

Modelling development will initially proceed in two main directions. Much work is needed (a) to establish the most suitable level of representation for primitive auditory objects within the scene and (b) the higher levels of the grouping hierarchy need to be defined.

3.2.1. Representation of Auditory Objects. The principle of disjoint, or exclusive, allocation allows a component to belong to only one group or 'stream' for any interpretation. However, sound is 'transparent' [6]. It does not hide other sounds so that a single component of an acoustic signal may result from either a single source generating a particular frequency at a particular amplitude for a given period of time or from more than one source contributing lower amounts of energy to that frequency and/or for different periods of the total time, which have combined in the environment to produce the total amplitude at that frequency received by the listener. If we recover the 'primitive' auditory object from the signal before grouping, it may fail to relate to other objects in the scene due to the contours of its 'components' being hidden. On the other hand, if we try to consider all possible combinations of sources which could have contributed to each auditory object, we lose the forms or structures which we are trying to compare. A possible solution is to assume that a single auditory object is present unless it will not group with any other objects and then re-evaluate it in the light of what form of components the other objects in the same scene would preferably group with in that region, subtracting the hypothesised component, creating a new component or components from the remaining energy in that frequency region and testing whether this leads to a coherent interpretation of the scene.

3.2.2. Development of grouping hierarchy. In implementing conflict resolution to complete the analysis of a scene into its component streams, both Gestalt and Scene Analysis principles can be considered.

The Gestalt principle of Pragnanz states that a figure always becomes as regular, symmetrical, simple and stable as prevailing conditions permit. Similarity, proximity, good continuation, habit or familiarity, belongingness, common fate and closure all promote pragnanz and have been identified in the context of perception, mostly for the purposes of vision research and explanation of visual illusions [16, 20]. Many of these principles have also been demonstrated in auditory perception. Stability, also known as Set, predicts that having achieved one interpretation, that interpretation will remain fixed throughout slowly changing parameters until the original interpretation is no longer appropriate. The better the pragnanz or "goodness" of the object, the greater the stability which it displays under subsequent changes [20]. We have started to develop experimental stimuli to test whether pragnanz and set appear to apply in conflict resolution.

The Scene Analysis principle of disjoint assignment applies in the development of groups up to the level of streams. This means that where an auditory object, or a group of auditory objects are perceived as components of one stream, they will not simultaneously belong to another stream in normal circumstances. However, the same set of objects can belong to a single stream in one context but be divided between streams in a different context [2], demonstrating that not only the relationships between the set of components considered but also competing relationships with other components within the same scene are significant in allocation to streams.

We need to develop methods of quantifying the "goodness" of group objects in order to investigate how pragnanz might mediate in the process of disjoint assignment. Work begun in the STREAMER prototype to assign grouping 'strengths' on the basis of closeness to an 'ideal' match and experimental evidence as to the relative weightings of particular grouping concepts in competitive environments, will be developed.

3.3. Schedule of Research

The computational research will occur, as far as possible, in parallel with the psychoacoustic experiments but of course any design decisions will be predicated upon the detailed results of the experimentation.

4. ACKNOWLEDGEMENTS

Work on the STREAMER prototype was supported by SERC Image Interpretation Initiative (grant no. GR/E42754 Green and Cooke). The current computational model and supporting psychoacoustic experimentation is sponsored by the joint MRC/SERC/ESRC Cognitive Science Initiative (grant no. SPG8921799 Williams, Green and Nicolson). SMW would like to thank Al Bregman and Pierre Ahad for their help in learning MITSYN and for many ideas and suggestions for the new project and the Nuffield Foundation for the travel award for my training visit to the lab at McGill, also, Martin Cooke and Mike Stannett for their comments and suggestions for this paper.

5. REFERENCES

[1] A S BREGMAN, J CAMPBELL, 1971, 'Primary auditory stream segregation and perception of order in rapid sequences of tones', J Exp Psychology, 89(2), pp 244-249.
[2] A S BREGMAN, A RUDNICKY, 1975, 'Auditory segregation: stream or streams?', J Exp Psychology: Human Perception and Performance, 1, pp 263-267.
[3] A S BREGMAN, S PINKER, 1978, 'Auditory streaming and the building of timbre', Canadian J of Psychology, 32, pp 19-31.
[4] A S BREGMAN, H STEIGER, 1980, 'Auditory streaming and vertical localization', Perception and Psychophysics, 28(6), pp 539-546.
[5] A S BREGMAN, 1984, 'Auditory Scene Analysis', Proc. 7th Int Conf Pattern Recognition, Montreal, pp 168-175.
[6] A S BREGMAN, 1990, 'Auditory Scene Analysis', MIT Press.
[7] V CIOCCA, A S BREGMAN, 1987, 'Perceived continuity of gliding and steady-state tones through interrupting noise', Perception and Psychophysics, 42(5), pp 476-484.
[8] M P COOKE, P D GREEN, 1990, 'The Auditory Speech Sketch', ibid.
[9] G L DANNENBRING, 1976, 'Perceived auditory continuity with alternately rising and falling frequency transitions', Canadian J of Psychology, 30, pp 99-114.
[10] G L DANNENBRING, A S BREGMAN, 1978, 'Streaming vs. fusion of sinusoidal components of complex tones', Perception and Psychophysics, 24(4), pp 369-376.
[11] C J DARWIN, 1984, 'Perceiving vowels in the presence of another sound: Constraints on formant perception', JASA, 76(7), pp 1636-.
[12] P D GREEN, et al., 1990, 'Bridging the gap between signals and symbols in speech recognition', in "Advances in Speech, Hearing and Language Processing", W A Ainsworth (ed), JAI Press, pp 149-191.
[13] J W HALL, M P HAGGARD, M A FERNANDES, 1984, 'Detection in noise by spectro-temporal pattern analysis', JASA, 76(1), pp 50-56.
[14] W L HENKE, 1990, 'MITSYN languages: Synergistic family of high-level languages for time signal processing [computer program]', available from the author, 133, Bright Street, Belmont, MA 02178.
[15] S E KEENE, 1988, "Object-Oriented Programming in COMMON LISP: A Programmer's Guide to CLOS", Addison-Wesley.
[16] K KOFFKA, 1936, 'Principles of Gestalt Psychology', Kegan Paul, London.
[17] D MARR, 1976,'Early processing of visual information', Philosophical Trans. of the Royal Soc., Series B, 275, pp 483-524.
[18] D MARR, 1982, "Vision", Freeman, San Francisco.
[19] S MCADAMS, 1982, 'Spectral fusion and the creation of auditory images', in "Music, Mind and Brain", Plenum Press, ch XV.
[20] G A MILLER, 1962, "Psychology, The Science of Mental Life", Penguin, England.
[21] A J H SIMONS, 1989, 'Object-Oriented Programming', AISB '89 tutorial, 25th Conf. AISB, University of Sussex, England.
[22] A J H SIMONS, 1988, ' An Object-Oriented architecture for phonetic processing', Proc SPEECH '88, 7th FASE Symposium, Edinburgh, UK, pp 361-368.
[23] H STEIGER, A S BREGMAN, 1981, 'Capturing frequency components of glided tones: Frequency separation, orientation and alignment', Perception and Psychophysics, 30(5), pp 425-435.
[24] WILLIAMS, 1989, 'STREAMER: A prototype tool for computational modelling of auditory grouping effects', Department of Computer Science Research Report No. CS-89-31, University of Sheffield.