ELSEVIER

Speech Communication 27 (1999) 351–366

SPEECH COMMUNICATION

A blackboard architecture for computational auditory scene analysis

Darryl Godsmark, Guy J. Brown *

Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK

Received 3 January 1998; received in revised form 1 October 1998

Abstract

A challenging problem for research in computational auditory scene analysis is the integration of evidence derived from multiple grouping principles. We describe a computational model which addresses this issue through the use of a 'blackboard' architecture. The model integrates evidence from multiple grouping principles at several levels of abstraction, and manages competition between principles in a manner that is consistent with psychophysical findings. In addition, the blackboard architecture allows heuristic knowledge to influence the organisation of an auditory scene. We demonstrate that the model can replicate listeners' perception of interleaved melodies, and is also able to segregate melodic lines from polyphonic, multi-timbral audio recordings. The applicability of the blackboard architecture to speech processing tasks is also discussed.

Keywords: Auditory scene analysis; Computer model; Blackboard

1. Introduction

According to Bregman's (1990) influential account, auditory organisation may be regarded as a two-stage process. In the first stage, early auditory processes decompose the mixture of sounds reaching the ears into a collection of 'sensory elements'. Subsequently, elements which are likely to have arisen from the same environmental source are grouped together, forming a mental representation of the sound source termed a 'stream'. The grouping process may involve the application of knowledge about sound sources such as speech and music (so-called 'schema-driven' grouping), or may involve mechanisms that operate independently of the characteristics of the sound source (so-called 'data-driven' or primitive' grouping).

It has long been thought that principles similar to those proposed by the Gestalt psychologists underlie primitive auditory grouping (Miller and Heise, 1950). Such principles include temporal and frequency proximity, common onset and offset, harmonicity, coherent amplitude and frequency modulation, and similarity of spatial location and timbre (see (Bregman, 1990) for a review). Undoubtedly, the ability of the auditory system to use such a diverse collection of heuristics contributes importantly to its remarkable robustness. However, little is known about the organisational framework within which these principles are applied. For example, several grouping principles may suggest mutually exclusive organisations; how does the auditory system resolve such a conflict? This question is critically important to the

* Corresponding author. Tel: +44-114-222-1821; fax: 44-114-222-1810; e-mail:g.brown@dcs.shef.ac.uk

0167-6393/99/$ – see front matter
PII: S0167-6393(98)00082-X

development of computational auditory scene analysis (CASA) systems – indeed, it has been suggested that the integration of evidence from multiple grouping principles is one of the hardest problems facing CASA (Bregman, 1998).

2. A framework for auditory organisation

The organisational framework proposed here is motivated by the observation that auditory organisation is both *context-sensitive* and *retroactive*. For example, consider the stimulus shown schematically in Fig. 1, which is due to Bregman and Tougas (1989). Listeners were presented with a repeating cycle consisting of a tone A followed by a pair of tones B and C. In some conditions a tone D was also included, otherwise a silent gap was left in each cycle which was the same duration as D. Subjects were asked to judge how clearly A and B could be heard as a repeating pair. They reported that the AB grouping was more salient when tone D was present; apparently, C and D tended to form a group so that the fusion of B and C was weakened, and hence it was easier for A to capture B into a sequential stream. Clearly, then, the organisation of tones B and C is dependent upon the context in which they are presented. Furthermore, the organisation of B and C cannot be determined until the presence of tone D has been confirmed or denied, suggesting that auditory organisations may be formed retroactively.

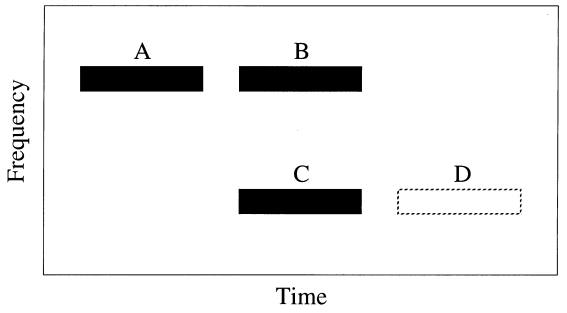

Fig. 1. Schematic illustration of one cycle of the stimulus used by Bregman and Tougas (1989). The tendency of tone A and tone B to be heard in the same perceptual stream is dependent on the presence or absence of tone D, illustrating that auditory perceptual organisation is both retroactive and context-sensitive.

Further evidence for retroactive auditory organisation has come from the study of perceptual restoration, in which a sound is perceived as continuous even though a part of it has been replaced with a noise burst (e.g. Warren, 1984). Perceptual restoration will only occur when there is sufficient evidence to suggest that the sound was occluded; in particular, there must be no evidence that the sound stopped before the noise burst and restarted after it. Hence, the mechanism of perceptual restoration appears to operate retroactively.

Taken together, the context sensitivity and retroactivity of auditory organisation suggest a means by which conflicts between grouping principles may be resolved. When faced with contradictory interpretations of the acoustic evidence, the auditory system might simply postpone its decision, with the expectation that disambiguating evidence will emerge in the near future. Evidence from studies of perceptual restoration suggest a limit on this temporal delay; the longest noise bursts that can induce perceptual restoration have a duration of 250–300 ms (Kluender and Jenison, 1992; Warren, 1984).

Accordingly, the computational framework described here introduces the concept of an *organisation hypothesis region* (OHR). The OHR is a temporal window of width 300 ms, which slides over the auditory scene. Within the window, grouping principles interact to suggest alternative organisations, and the grouping of acoustic elements remains mutable. However, once elements pass beyond the limit of the temporal window, a fixed organisation is imposed upon them. Additionally, organisations within the OHR are debated at many levels of abstraction. At the lowest levels, grouping principles operate directly upon an early auditory representation of the acoustic evidence in a source-independent manner (primitive grouping). At higher levels, auditory organisation is influenced by knowledge about specific sound sources (schema-driven grouping).

In the remainder of this article, we describe the computer model and evaluate its performance on several musical sound separation tasks. The study extends our previous work (Godsmark and Brown, 1996a,b, 1998) by integrating emergent properties (pitch and timbre) and schema-driven

grouping principles (predictions of meter and melody) into a coherent computational framework. Where possible, the model is informed by psychophysical findings, although our approach should be regarded as functional – the model organises the acoustic input in a manner that is consistent with the behaviour of human listeners, but we do not make strong claims that the mechanisms of the model have a direct biological counterpart.

3. The computer model

The computer model consists of two major processing stages, corresponding to the two conceptual stages of auditory scene analysis. The input to the model is a single acoustic signal, which represents the superimposed activity of several sound sources. The acoustic waveform is sampled at a frequency of 25 kHz with 16 bit resolution. Initially, the sampled signal is processed by a model of the auditory periphery, yielding a time-frequency representation called *synchrony strands* (Cooke, 1993). This representation provides the substrate for the organisational phase of the model, which is based upon the blackboard metaphor of problem solving (Erman et al., 1980; Engelmore and Morgan, 1988). A blackboard system can be thought of as a group of independent knowledge sources (hereafter referred to as *experts*) that are able to communicate only by manipulating information on a globally accessible data structure (the blackboard). Given the state of the blackboard, an expert may indicate that it wishes to perform an action (it *fires*). Coordination of experts is achieved by a *scheduler*, which determines the sequence in which actions are executed. Several properties of the auditory scene analysis problem make it particularly suitable for a blackboard-based approach; it involves a large solution space, noisy and unreliable data, a need to integrate diverse types of information, and many semi-independent sources of knowledge are required to form a solution (see also (Nii, 1988)).

The organisational phase of the model can be further divided into four stages: primary, primitive, emergent and schema-driven grouping. Each stage of organisation is associated with one or more layers of abstraction on the blackboard (see Fig. 2). The primary stage concerns the formation of synchrony strands (level 1). Primitive stages involve the formation of featured strands and note hypotheses (levels 2 and 3). The emergent stage relates to emergent properties (level 5) and hypothesised melodic lines (level 6). At the highest level of the blackboard (level 8), meter and motif predictions relate to schema-driven organisation. Additionally, the blackboard has levels for evaluated primitive and emergent hypotheses (levels 4 and 7). These levels (strictly, they are meta-levels) do not correspond to a particular organisation process; rather, they are present for reasons of computational convenience.

Following the organisation phase of the model, the synchrony strands on the blackboard are arranged into groups, such that the strands in each group are likely to have arisen from the same sound source. This is the result of the CASA process. Currently, we have quantified the performance of the model using a music transcription task. A musical score is derived from the groups of synchrony strands on the blackboard, and this is matched against the score of the corresponding acoustic input. Metrics have been devised to quantify the success of segregation, i.e. whether the system is able to segregate a polyphonic, multitimbral musical performance into its constituent melodic lines.

Inevitably, much of the following discussion is biased towards the segregation of musical sounds. However, it should be stressed that the architecture proposed here is quite general, and could equally be applied to the segregation of speech from interfering sounds. Similarly, other evaluation techniques could be employed, such as resynthesis from groups of synchrony strands (Cooke, 1993) or the comparison of signal-to-noise ratio before and after processing by the model (Brown and Cooke, 1994a).

3.1. Auditory periphery

Cochlear frequency selectivity is simulated by passing the sampled acoustic signal through a bank of bandpass filters with overlapping

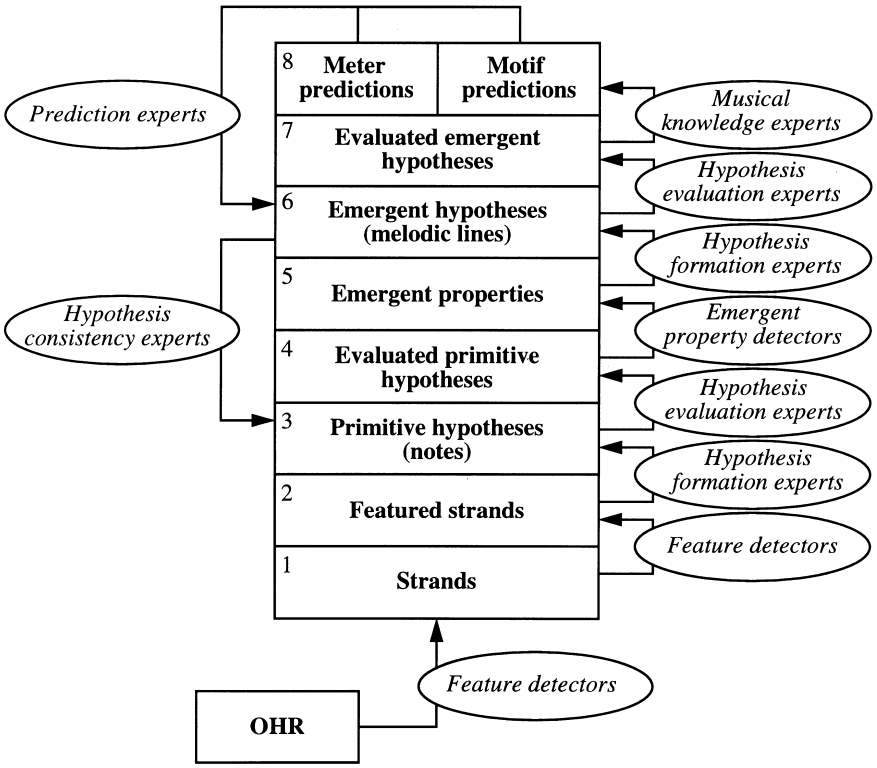

Fig. 2. Schematic overview of the model.

pass-bands. More specifically, we use a bank of 'gammatone' filters (Patterson et al., 1988) which have the (complex) impulse response

$$g(t) = t^{n-1} \exp(-bt) \exp(\mathrm{i}\omega t). \quad (1)$$

Here, ω is the radian centre frequency of the filter, n is the filter order and b is related to bandwidth. We use fourth-order filters (i.e., $n = 4$) with centre frequencies linearly distributed between 25 Hz and 11 kHz on the ERB-rate scale of Glasberg and Moore (1990). Additionally, the gains of the gammatone filters are adjusted to reflect the transfer functions of the outer and middle ears, using data from the ISO standard for equal-loudness contours (ISO, 1988).

Cooke (1993) has demonstrated that the instantaneous response frequency of an auditory filter can be modelled in terms of a relatively slowly changing component which corresponds to the dominant frequency, and a periodic component which is correlated with the acoustic signal. The slowly varying component (instantaneous frequency), can be conveniently obtained from a digital implementation of the gammatone filter as the following quantity:

$$v(t) = \frac{1}{2\pi}\left[\omega + \frac{\Im(t)\frac{\mathrm{d}}{\mathrm{d}t}\Re(t) - \Re(t)\frac{\mathrm{d}}{\mathrm{d}t}\Im(t)}{\Im^2(t) + \Re^2(t)}\right]. \quad (2)$$

Here, $v(t)$ is the instantaneous frequency and $\Re(t)$ and $\Im(t)$ represent the outputs of the real and imaginary parts of the gammatone filter. We compute median-smoothed estimates of $v(t)$ using a window length of 10 ms. Since $v(t)$ is relatively slowly varying, a degree of data reduction is introduced by calculating the median every 0.5 ms.

Similarly, the instantaneous amplitude $e(t)$ at the output of the gammatone filter is available as the following quantity:

$$e(t) = \sqrt{\Re^2(t) + \Im^2(t)}. \quad (3)$$

Again, $e(t)$ is smoothed and downsampled by computing the median within a 10 ms window at intervals of 0.5 ms.

3.2. Calculating place-groups

Since adjacent filters in the gammatone filterbank have overlapping pass-bands, a highly synchronised response is observed in sections of the filterbank that are excited by the same spectral component (i.e., contiguous sections of the filterbank exhibit a nearly identical instantaneous frequency). This redundancy can be exploited by noting that filters with centre frequencies above a dominant component respond with a frequency below their centre frequency, and vice versa (Cooke, 1993). More specifically, we compute the function

$$E(t) = \upsilon(t, f) - f, \tag{4}$$

where $\upsilon(t, f)$ is the instantaneous frequency of the filter with centre frequency f at time t. Filter channels which lie between successive maxima and minima of $E(t)$ are combined to form place-groups, each of which represents a dominant spectral component. Finally, a frequency is computed for each place-group by forming a mean of $\upsilon(t, f)$ over the channels that comprise the group, weighted by their instantaneous amplitudes. A typical place-group representation is shown in Fig. 3.

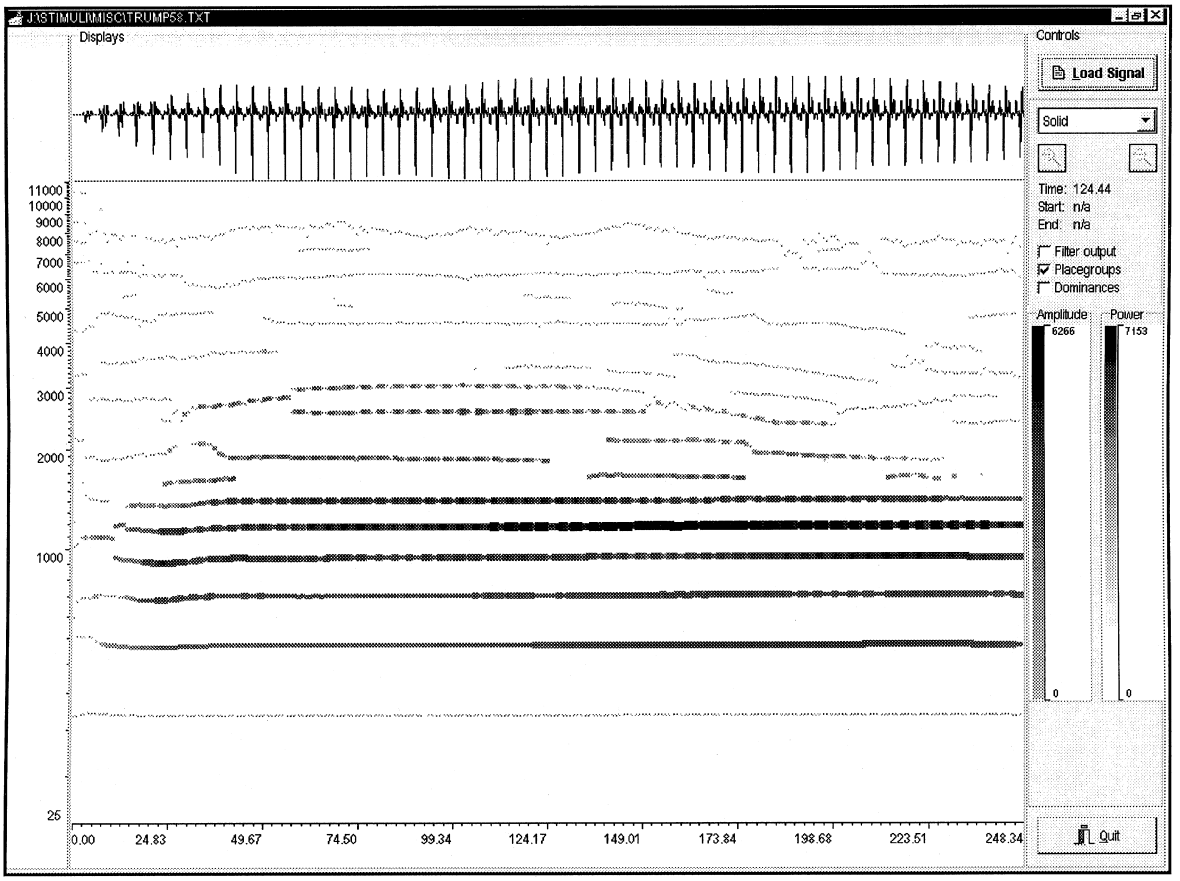

Fig. 3. Acoustic waveform (upper panel) and place-group representation (lower panel) for a trumpet tone. The thickness of a place-group is related to its amplitude.

4. Implementation of the grouping strategy

The place-group representation of the acoustic stimulus resides at the lowest level of the blackboard. During a primary organisation stage, place-groups are aggregated to form synchrony strands (Cooke, 1993), each of which represents a dominant spectral component that extends through time and frequency. As a representational substrate for CASA, synchrony strands have an advantage over frame-based spectral representations because temporal continuity is made explicit (see also (Brown and Cooke, 1994a)). They are also sufficiently few in number to allow grouping strategies which attempt to explain every synchrony strand in an auditory scene.

Subsequently, a primitive organisation stage fuses *synchrony strands* that are likely to have originated from the same environmental source. For example, a group of strands might correspond to the components of a musical note or a vowel. Additional properties emerge as a consequence of this grouping, including fundamental frequency and timbre. An emergent organisation phase combines groups of strands on the basis of these emergent properties, forming structures that correspond to Bregman's (1990) notion of a stream (for example, a melodic line or a sequence of speech sounds uttered by the same speaker).

Finally, properties of streams are identified through the application of source-specific knowledge. In our current implementation, this level of the blackboard identifies properties such as the meter of a melody, and the presence of recurrent musical phrases. In turn, this information is used to generate predictions about future auditory events, which are exploited at lower organisational levels of the blackboard.

The organisational stages of the blackboard architecture are driven by the OHR. At each time frame, the OHR is stepped forward and any new place-groups that are generated are added to the blackboard. This may cause a number of experts to fire, leading to further changes in the state of the blackboard that may activate other experts. In some cases, movement of the OHR only leads to changes at the lower levels of the blackboard; for example, a single place-group might be appended to an existing synchrony strand. Changes at the higher levels of the blackboard occur when a synchrony strand is terminated: new hypotheses might be generated based on the offset time of the strand, or on its proximity in frequency to other strands. The OHR is not advanced to the next time frame until the blackboard has reached a state of equilibrium; that is, there are no experts which are able to fire. Hence, the set of hypotheses on the blackboard is refreshed every time the OHR is progressed.

4.1. Primary organisation

As place-groups appear in the OHR, synchrony strands are formed on the basis of three principles; temporal continuity, frequency proximity and amplitude coherence. More specifically, in order to extend an existing strand a new place-group must be temporally contiguous with the final place-group in the strand, and must also be within 3% of its frequency. Place-groups that are unable to extend an existing synchrony strand become the first component of a new strand, and existing strands that cannot recruit further place groups are terminated.

The extension of synchrony strands by temporal continuity and frequency proximity allows slowly-varying frequency components to be successfully tracked. However, it is possible (particularly for sounds such as polyphonic music) that harmonics belonging to different acoustic events will be coincident. If there is no temporal gap between such harmonics, a single synchrony strand may be inappropriately formed from them.

This problem can be resolved by noting that the appearance of a new acoustic event will be accompanied by an increase in intensity. Hence, the model also exploits a principle of amplitude coherence; we allow a place-group to start a new strand if it would otherwise have caused an abrupt increase in the amplitude of an existing strand. Amplitude increases are detected by convolving strands with a bipolar kernel, which simultaneously smooths and differentiates the input (Mellinger, 1991). In practice, it is preferable to search for amplitude increases on a number of time-scales; accordingly, we use four kernels with

widths between 5 ms and 20 ms, and start a new strand if an onset is detected by any kernel.

Note that strongly amplitude-modulated sounds may cause spurious onsets to be detected, forming a sequence of short synchrony strands. However, these fragments are likely to be regrouped by later organisational processes.

4.2. Feature detection

As synchrony strands are formed, features can be computed from them. These include the onset and offset time, initial and final frequency, and a frequency transition history (computed by taking the ratio between the frequencies of adjacent place-groups in the strand). The onset and offset times of a strand are simply taken to be the times at which the first and last place groups occur; this was found to be adequate in the current model, although a more sophisticated scheme would be needed if grouping by common onset and offset were applied at emergent levels of the blackboard (e.g., the grouping of notes by common onset).

Features are detected by opportunistic experts; in other words, an expert only fires when the blackboard contains sufficient evidence for the presence of a feature. For instance, an expert that detects the offset time of synchrony strands will not fire if all the strands on the blackboard are still evolving; it will only be activated when a strand has been terminated.

4.3. Primitive organisation

Bregman (1990) has argued that fusion is the default state of auditory organisation; for example, although white noise contains a random collection of fusion and segregation cues, it is still perceived as a perceptual whole. Accordingly, primitive organisation in our blackboard architecture begins by allocating all synchrony strands to the same stream.

As strand features become available, alternative organisations of the auditory scene can be considered. However, it is impractical to consider every possible interpretation of the acoustic evidence – even for a small number of synchrony strands, exhaustive search is computationally intractable. Hence, our model uses a number of heuristics to generate only those organisational hypotheses which are in some sense plausible. These heuristics are embedded in *hypothesis formation experts*, such that each expert represents a particular grouping principle. Note that a hypothesised organisation must account for all of the synchrony strands in the OHR; accordingly, the hypotheses generated by formation experts may be regarded as independent.

The format of a hypothesis differs according to the layer of abstraction on the blackboard (see Fig. 2). Primitive hypotheses (level 3) are a logical grouping of synchrony strands that overlap in time; in the context of musical stimuli, a primitive hypothesis corresponds to a note. At level 5 of the blackboard, emergent properties such as fundamental frequency and timbre are added to these note hypotheses. Grouping on the basis of these emergent properties allows hypotheses to be formed about sequences of acoustic events that are separated in time; these are the emergent hypotheses (melodic lines) represented at level 6 of the blackboard.

4.3.1. Hypothesis formation experts

Each hypothesis formation expert operates in a similar manner. Given a 'target' strand with some newly-derived features, the expert attempts to find another strand which is close (in some respect) to the target. For example, in the case of an expert that detects the common onset of acoustic events, the closest strand will be the one whose start time is most similar to the start time of the target strand. If the relationship between the two strands is nearly ideal (e.g., they have almost identical onset times) then the expert will only generate hypotheses in which the two strands are grouped. Conversely, if the two strands conform to the principle very weakly, the expert will only consider hypotheses in which the two strands are segregated. If the relationship lies between these two extremes, the expert generates both sets of hypotheses.

Since hypothesis formation experts operate independently of one another, it is possible that several experts will generate identical hypotheses; in such cases, the hypotheses are merged to

preserve uniqueness. In the current model, experts have been implemented for principles of onset and offset synchrony, temporal and frequency proximity, harmonicity and common frequency movement. A detailed description of the function of each expert can be found in (Godsmark, 1998).

4.3.2. Local hypothesis evaluation

When place-groups pass beyond the limit of the OHR, their organisation is fixed according to the 'best' hypothesis on the blackboard. Hence, an evaluation scheme is employed which ranks hypothetical organisations according to their quality. As before, evaluation schemes are embedded in the experts for each grouping principle. Hypotheses are evaluated under two situations; a synchrony strand that forms part of an organisation may have been moved to a different stream by a hypothesis formation expert, or a new strand feature may have been detected.

Evaluation functions relate the likelihood that two strands will be grouped to their degree of conformance with a particular grouping principle. For the majority of grouping principles, the evaluation function is two-dimensional; the only exception is the proximity principle, which combines temporal and frequency proximity into a three-dimensional function (Fig. 4). The output of each evaluation function is a score between −0.5 (indicating that the organisation has weak support) and +0.5 (indicating that the organisation has strong support). This mapping ensures that weak grouping relationships actively penalise a hypothesis.

For strands that have already been segregated into different streams, an evaluation score is derived by negating the evaluation function. For example, consider two strands that have almost identical onsets. A hypothesis that places these two strands in the same stream will receive a score close to +0.5 from the onset evaluation function, since this organisation is consistent with grouping by a

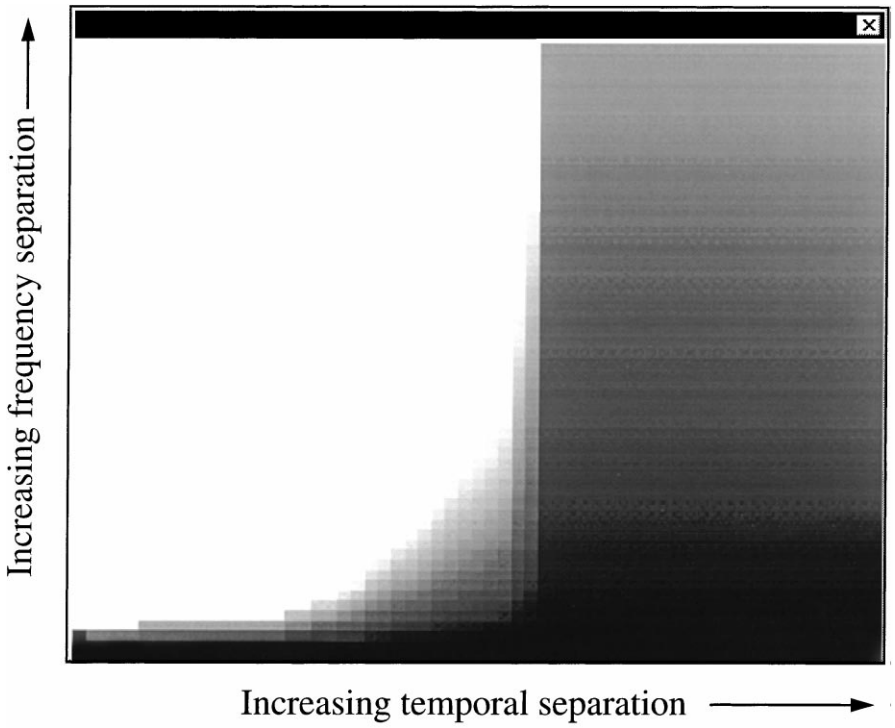

Fig. 4. The evaluation function for proximity. Black represents a high score for hypotheses in which two strands are grouped, and white represents a low score.

principle of common onset. Conversely, a hypothesis that places the two strands in different streams will be penalised with a score close to −0.5.

When a strand is moved to a new organisation, or when a new strand feature becomes available, the relationship between this 'target' strand and every other strand in the OHR is evaluated by each hypothesis evaluation expert. A *local evaluation score* is then formed for the target strand, which is simply the sum of the scores derived from each evaluation expert.

4.3.3. Global hypothesis evaluation

In addition to the local evaluation score, each hypothesis is also given a *global evaluation score*. The global score is a cumulative sum of the local scores which evolve for a hypothesis over time. When a hypothesis is initially generated, the local and global scores are identical. However, as new place-groups appear in the OHR, the hypothesis is modified and a set of local evaluation scores are generated for the new organisations. These local evaluations are then incorporated into the global evaluation score.

The global evaluation score is used to determine the 'best' organisation of a collection of acoustic events. When the OHR passes beyond a place-group, the current set of hypotheses is searched and the one with the highest global score is selected, effectively committing the place-group to a particular stream. It should be noted, however, that although place-groups eventually leave a hypothesis, the hypothesis for that organisation (which may contain other strands) continues to evolve (and hence the global evaluation score continues to accumulate). Consequently, the grouping process is highly context-sensitive; the global score reflects an overall evaluation of the relationship between a target strand and other strands not just within the OHR, but across the entire history of the acoustic signal.

A complication arises with this scheme when an expert changes the organisation of a strand that has been evaluated as part of another hypothesis. For example, consider a case in which two strands are harmonically related and have similar onset times. These strands will be grouped on the basis of both onset synchrony and harmonicity, and a global evaluation score will evolve for this organisation. However, the strands may be segregated at a later time if they become inharmonic; hence the evaluation based on onset synchrony is no longer valid, as the strands are no longer allocated to the same stream. Consequently, when a strand is moved to a new stream, it is necessary to remove any contribution that it made to the global evaluation score of its previous organisational hypothesis.

4.3.4. Parameter estimation

The parameters associated with the hypothesis formation experts and hypothesis evaluation experts have been set to ensure that they are consistent with known psychophysical data. Initially, the evaluation function for the proximity principle was estimated using alternating-tone sequences of the form employed in van Noorden's early streaming experiments (van Noorden, 1975) (see Fig. 4). A hill-climbing optimisation was used, in which the evaluation function of the proximity expert was adjusted until the output of the model matched van Noorden's perceptual data. Further grouping principles were then added incrementally and their parameters were estimated in a similar fashion, using increasingly complex stimuli that incorporated cues such as harmonicity and onset asynchrony. For example, the sequence of tones used by Bregman and Tougas (Fig. 1) was one of several stimuli that were used to estimate parameters for the onset grouping expert.

By deriving parameters for the model from a large range of psychophysical findings, we avoid some of the limitations of previous approaches. For instance, Kashino and Tanaka (1992) derived an evaluation function for an onset grouping principle directly from a single psychophysical experiment. As such, their approach describes the behaviour of a grouping principle only in relation to a specific acoustic stimulus, and does not acknowledge the context-sensitivity of auditory organisation. Furthermore, it is difficult to design experimental paradigms which genuinely isolate the behaviour of a single grouping principle; even for simple stimuli, auditory organisation may reflect the operation of multiple grouping mechanisms.

4.4. Emergent organisation

The fusion of a group of synchrony strands allows emergent properties to be derived – in our current implementation, we compute estimates of fundamental frequency and timbre. The emergent organisation stage of the model attempts to further organise groups of strands on the basis of these properties. For example, for a stimulus consisting of musical sounds, the primitive organisation stage determines which strands are part of the same musical note, and emergent organisation determines which notes are part of the same melodic line.

The emergent organisation phase proceeds in a similar manner to primitive organisation; a temporal window is employed, and organisation within the window remains mutable. Hypotheses are created by hypothesis formation experts, which attempt to identify groups of strands that are similar in respect of an emergent property; the expert then determines whether they should definitely group, definitely segregate, or whether the organisation is ambiguous. Finally, local and global evaluation metrics are computed. Additionally, a close correspondence is maintained between the emergent and primitive levels of the blackboard. Since an event at the emergent level is also represented as a group of strands at the primitive level, any modifications to the emergent organisation must be reflected at lower levels of the blackboard. For instance, if a musical note is moved to a different stream at the emergent level, the group of strands that constitute that note must also be moved to the new stream at the primitive level.

The main difference between primitive and emergent organisation lies in the complexity of the underlying representations and the length of the organisational window. A single musical note could be several seconds in duration, and hence the window for emergent organisation must be significantly wider than the OHR. In our current implementation, we use a window width of 5 s; wider windows have little effect on the performance of the model for the test stimuli used here.

4.4.1. Grouping by pitch proximity

The pitch proximity principle works in a similar manner to the primitive proximity mechanism, except that grouping decisions are made on the basis of the pitch of a group of strands, rather than the frequency of a single strand. In fact, our model currently uses a very simplistic indicator of pitch; the median frequency of the lowest-frequency strand in a group (which is presumed to be the fundamental frequency). For the stimuli used in our current evaluation of the model, this approximation is adequate; more principled schemes, such as the time-frequency harmonic sieve proposed by Cooke (1993), could be integrated into the model in a straightforward manner.

4.4.2. Grouping by timbral similarity

Modelling the perception of timbre presents a challenging problem, not least because the acoustical correlates of timbre are far from certain. However, it is clear that timbre is multidimensional; it is not correlated with a single acoustic property. Currently, it is believed that properties such as attack and decay transients, inharmonic noise and changes in the distribution of spectral energy contribute to the perception of timbre (Iverson and Krumhansl, 1993; Handel, 1995). Accordingly, our model employs a novel representation of timbre that captures dynamic changes in amplitude and spectral shape, termed a *timbre track*.

A timbre track is a representation of changes in spectral centroid (which can be related to 'brightness') plotted against changes in amplitude. The centroid frequency of a group of strands is calculated by summing the product of the frequency and amplitude of each place-group, and dividing this by the sum of the amplitudes (see also Iverson and Krumhansl, 1993; Brown and Cooke, 1994b). The centroid is computed within a 20 ms window at 10 ms intervals, yielding a series of values that may be thought of as a 'brightness envelope'. A similar approach is used to compute an amplitude envelope for a group of strands; within the same 20 ms window, the average of all place-group amplitudes is calculated. The amplitude and brightness envelopes are then smoothed by polynomial fitting, and differentiated by convolution with the first derivative of a gaussian. Hence, a three-dimensional representation of timbre is obtained in which each point (b, a, t) represents the brightness b and amplitude a at time t.

In fact, it is convenient to project each point in the timbre space onto two-dimensions by simply taking the pair (b, a). This stage is motivated by the observation that many musical instrument sounds have a sustained portion in which a repeating pattern of timbral change occurs. By projecting onto two dimensions, these repetitive changes become visible as loops in the timbre track. Multiple instances of these loops are removed, since they introduce a dependency on duration that might confound the subsequent matching of timbre tracks. Typical timbre track representations are shown in Fig. 5. Elsewhere, we have demonstrated that timbre tracks for a given sound source (such as a musical instrument) are relatively invariant over both fundamental frequency and intensity (Godsmark and Brown, 1996b).

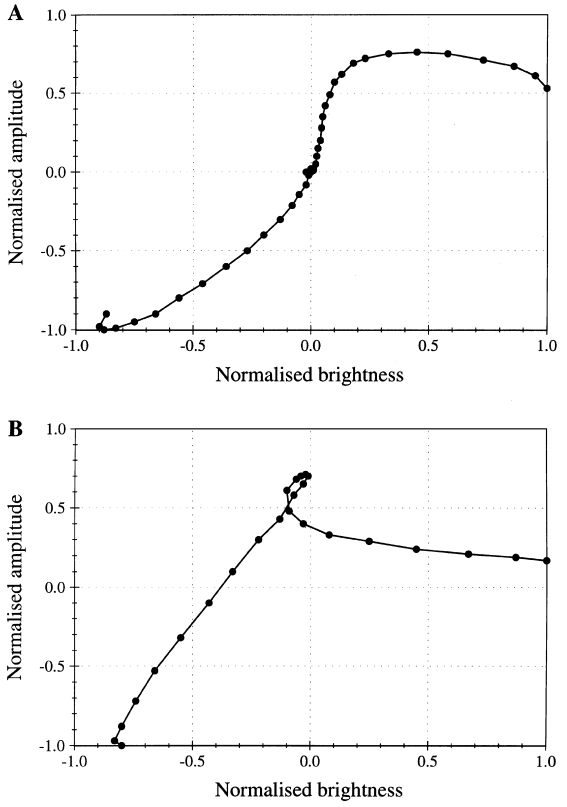

Fig. 5. Timbre tracks for a violin tone (A) and piano tone (B).

By treating each point in the timbre track as a control point on a b-spline curve, it is possible to ensure that each track contains a specific number of points (Foley et al., 1996). Hence, the similarity of two timbre tracks can be quantified by computing the sum of the Euclidean distances between equivalent points on each track. Matching timbre tracks in this way allows an evaluation function to be specified, which relates the similarity of two timbre tracks to the likelihood that the corresponding groups of strands will be placed in the same stream.

4.5. Schema-driven organisation

At the highest level of the blackboard, organisation of auditory events is influenced by source-specific knowledge. Currently we have evaluated the model with musical stimuli, and have developed experts for identifying meter and repeated melodic phrases. The information provided by these experts is fed into a prediction mechanism which is able to strengthen hypotheses at the emergent level. For example, if a repeating musical phrase is identified, then the next musical note can be anticipated; any organisational hypotheses at the emergent level which support this prediction will be strengthened. Similarly, by deriving the meter of a musical piece, the model can predict when the next note will arrive, and strengthen hypotheses which support the presence of an acoustic event at that temporal location.

In our current implementation, the experts for meter and phrase identification work in a rather simplistic manner. Meter prediction is based upon the scheme proposed by Rosenthal (1992), in which the temporal location of future events is predicted from the inter-onset-intervals of previous events. However, an additional complication arises because the acoustic evidence on the blackboard may represent the activity of more than one sound source. Accordingly, meter predictions are made separately for each stream (recall that groups of strands are partitioned into streams at the emergent level of the blackboard).

A similar mechanism underlies the expert for melody identification, except that predictions are made about the fundamental frequency of events

as well as their timing. Predictions are based upon the recognition of sequences of *relative* changes in fundamental, and hence the expert is able to recognise transposed motifs. This approach is consistent with the work of Deutsch (1980), who has presented evidence that listeners memorise tonal sequences using relative, rather than absolute, changes in frequency. For monophonic input, melody prediction is based on the fundamental frequency of notes. For polyphonic music, melody prediction is rather simplistic because the model does not currently exploit knowledge about musical chords; the notes of a chord are grouped (because they have a similar timbre), and the fundamental frequency of this group as a whole is used for melody prediction.

5. Evaluation

The computational model has been evaluated by investigating its ability to reproduce data from a wide range of psychophysical experiments. Elsewhere, we have shown that the model is consistent with phenomena such as the build-up of streaming over time, context-sensitive and retroactive organisation, and competition between sequential and simultaneous grouping (Godsmark and Brown, 1996a,b). Here, we focus on two rather challenging evaluation tasks; reproducing listeners' perception of interleaved melodies, and segregating polyphonic music into its constituent melodic lines.

In the first experiment, we investigate whether a model whose parameters are derived from psychophysical studies using simple tonal stimuli can also predict listeners' performance in a more complex melody identification task. The second experiment further investigates the scalability of our model; can the principles and parameters derived from psychophysical studies also be successfully applied to the segregation of complex musical sounds?

In the evaluations that follow, the same set of model parameters were used for every condition; furthermore, the stimuli used for evaluation were not presented to the model during the parameter estimation stage.

5.1. Identification of interleaved melodies

Hartmann and Johnson (1991) have investigated the ability of listeners to identify interleaved melodies (i.e., a stimulus of the form $x_1y_1x_2y_2 \ldots x_ny_n$, where the x_n are notes of one melody and the y_n are notes of a second melody). The data for five of their experimental conditions are shown in Fig. 6. In the null condition, the two melodies were played in the same pitch range, with each note represented by a pure tone of equal duration. No stream segregation was apparent in this condition, and hence identification of the interleaved melodies was difficult. However, identification performance could be substantially improved by transposing one of the melodies up by an octave (condition 1). An improvement in performance was also obtained by introducing a difference in timbre between the two melodies; this was achieved by sounding the notes of one melody with a harmonic complex rather than a pure tone (condition 3). Similarly, identification was improved when the notes of one melody had a long attack and short decay, and the notes of the second melody had a short attack and long decay (condition 5). In the final condition, a difference in duration was introduced so that the notes of one melody were twice as long as the notes of the other.

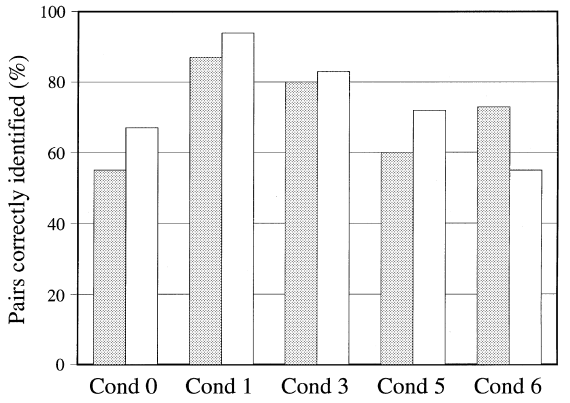

Fig. 6. The performance of listeners (grey bars) and the computer model (white bars) in identifying two interleaved melodies. The details of each condition are explained in the text. The data for listeners is from Hartmann and Johnson (1991).

Fig. 6 also shows the performance of our model for Hartmann and Johnson's stimuli. To enable this comparison, a simple template-matching mechanism for melody identification was added to the model. The model reproduces the overall pattern of listeners responses, although it outperforms human listeners in the first four conditions. It is likely that this performance difference arises because the model is able to identify melodies with greater reliability than listeners; it cannot be due to a difference in segregation ability, since the model outperforms listeners in the null condition (in which no cues for segregating the melodies are available).

It should be noted that the model performs poorly in condition 6; its performance in this condition is below that for the null condition, a result which is inconsistent with Hartmann and

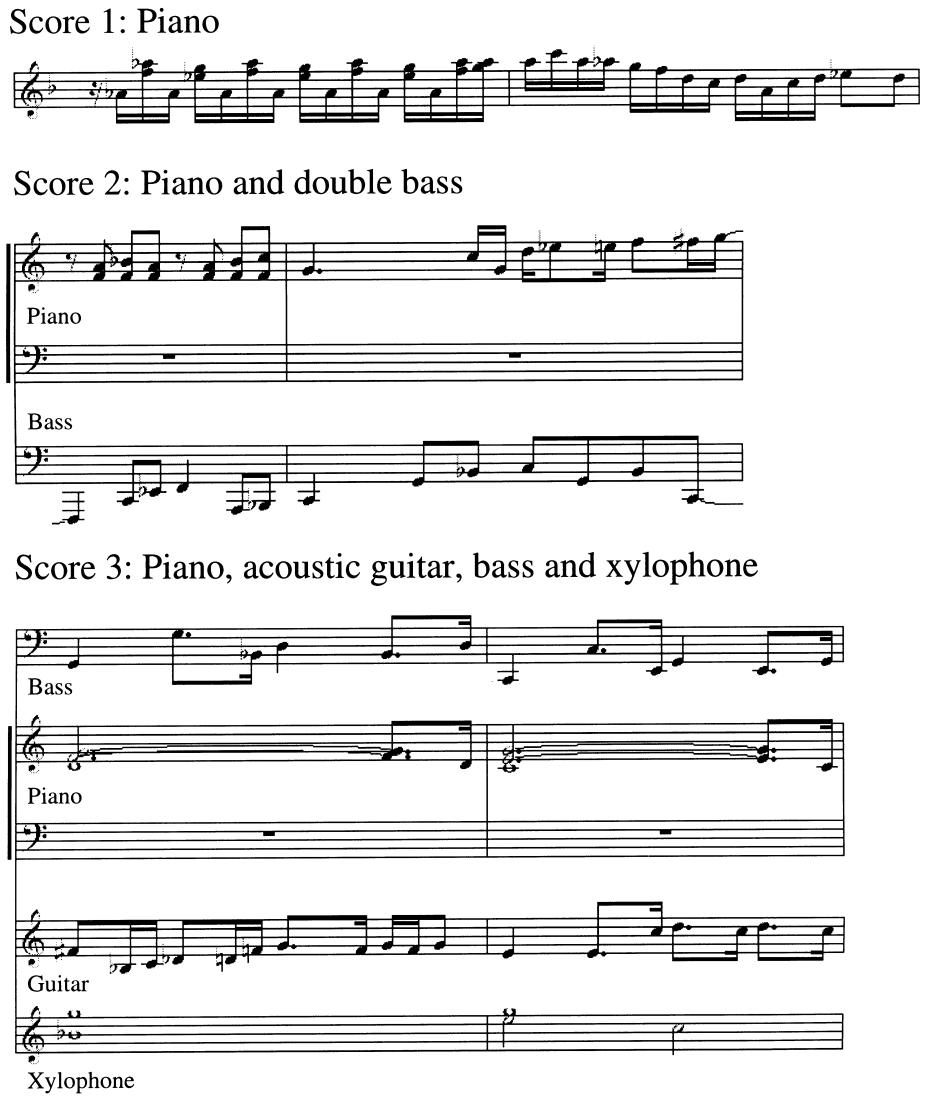

Fig. 7. Extracts from three musical scores used to test the computer model. Each score was performed by a sample-based MIDI synthesizer and recorded to a single-channel audio file, which was used as input to the model.

Johnson's findings. By increasing the duration of notes in one melody, the temporal proximity between notes is reduced and the model erroneously allocates consecutive notes (which belong to different melodies) to the same stream. In this condition, it is possible that listeners exploit a principle of similarity (Rogers and Bregman, 1993); more specifically, they group notes which have similar durations. Currently our model does not employ a principle of similarity, and hence grouping by temporal proximity leads to an inappropriate organisation.

5.2. Segregation of musical sounds

The model has also been evaluated by investigating its ability to segregate polyphonic music into its constituent melodic lines. In order to quantify the performance of the model, a metric is used that matches the onset time and fundamental frequency of notes (i.e., groups of strands) against the original musical score. Using this metric, four types of error may occur:

1. Either the onset time or fundamental frequency do not correlate with a note in the score.
2. A note may be grouped with events belonging to the wrong melodic line (i.e., the onset time and fundamental frequency have been correctly identified, but the corresponding group of strands has been allocated to the wrong stream).
3. A note may be deleted. This may occur when there is considerable overlap between the spectra of two simultaneous acoustic events, causing a single group of strands to be produced.
4. A note may be inserted; in other words, the organisational strategy may inappropriately partition a set of synchrony strands into two groups.

Three musical examples were used to test the model: extracts from their scores are shown in Fig. 7. The first piece is a solo piano recording, but the second is more complex, consisting of piano and double-bass parts. Although the latter piece is polyphonic and multi-timbral, the two instruments play in distinct registers. The final piece consists of four parts; piano and guitar (playing in the same octave), bass and occasional xylophone motif.

Fig. 8 shows the model performance for the three musical examples. Each score was performed by a sample-based MIDI synthesizer and recorded to a single-channel audio file, which provided the input to the model. The model performs well in the first two conditions, although a number of insertion errors occur. These are due to the limited high-frequency resolution of the gammatone filterbank, which may cause unrelated harmonics to be integrated into a single strand. Such strands oscillate between the frequencies of the two harmonics, destroying harmonicity and common frequency movement cues. As a result, these spurious components do not group strongly with other strands in the auditory scene, and are interpreted as notes in a separate stream. Furthermore, the incorrect grouping of these harmonics may affect the timbre of remaining notes, causing some of them to be allocated to inappropriate streams.

The final condition is very challenging, and this is reflected in the poor performance of the model; there are many insertion and deletion errors. This is not surprising, given that the piano and guitar are playing in the same register; many of the guitar notes are masked by the piano, causing the model to miss them completely. Also, harmonic components belonging to the guitar and piano occa-

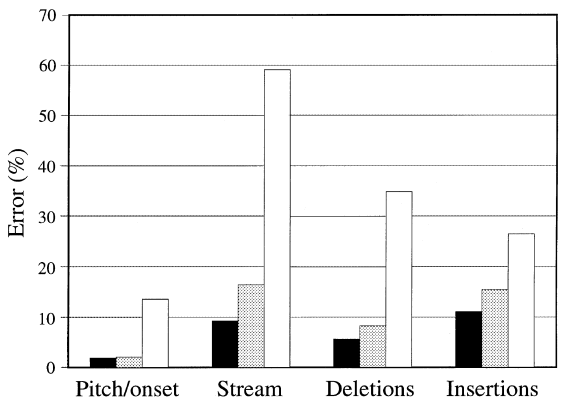

Fig. 8. Segregation performance for three musical stimuli, consisting of solo piano (black bars), piano and bass (grey bars) and a complex four-part arrangement (white bars). Errors in each category are expressed as a percentage of the number of notes in the musical score.

sionally fuse, producing a group of strands with a novel timbre that is placed in a new stream.

6. Conclusions

This paper has presented an architecture for CASA that accommodates the interaction of context-sensitive and retroactive grouping mechanisms at several levels of abstraction. Competition between grouping principles is managed implicitly using a blackboard metaphor; knowledge sources influence the score of hypothesised organisations, and the hypothesis with the highest score is selected when an organisation must be fixed. It has also been demonstrated that the model is able to integrate top-down and bottom-up processing, by allowing high-level predictions about meter and melody to influence the organisation at lower levels of the blackboard.

Since the early work by Cooke et al. (1993), a number of other workers have proposed blackboard architectures for CASA. In a recent thesis, Ellis (1996) has described a 'prediction-driven' system, which reconciles acoustic evidence with an internal model of environmental sound sources. Ellis uses a blackboard to manage multiple hypotheses about the organisation of the auditory scene, and to handle competition between them. However, his approach does not fully exploit the potential of blackboard systems to co-ordinate processing at different levels of abstraction, and his evaluation is less concerned with psychophysical plausibility than ours. Similarly, the multi-agent system of Nakatani et al. (1998) and the Bayesian network approach of Kashino et al. (1998) both address the issue of combining evidence from multiple grouping principles, but are different in their motivation to the system described here. Other related work includes the IPUS sound-understanding system described by Lesser et al. (1995), which employs a sophisticated blackboard architecture. However, the IPUS system addresses complex signal analysis in general, and is not strongly motivated by an auditory account.

Where possible, the parameters of our model have been estimated from available psychophysical data. Although this approach is principled, it may prove problematic if there is inconsistency between different psychophysical findings. In practice, during parameter estimation for the model there were no occasions where parameter values could not be set to accommodate the relevant psychophysical data. However, it is likely that we will have to address this problem during future development of the model; as more grouping principles are added, there will be a greater likelihood of inconsistencies during parameter estimation.

Perhaps a unique aspect of our model is its scalability, a property that Bregman (1998) has identified as crucial for CASA systems. The parameters associated with the models' grouping principles have been estimated from psychophysical data. Consequently, the model is able to demonstrate a qualitative match to a wide range of psychophysical phenomena. However, the architecture can also be scaled up to process complex musical stimuli. In short, the architecture allows us to exploit principles derived from simple psychophysical experiments, and to investigate how those same principles apply to the segregation of real-world sounds.

Although the model has currently been evaluated with musical signals, it is intended to be a general architecture for CASA. Little modification of the model would be required in order to process mixtures of speech and other environmental sounds. For example, although the generation of timbre tracks is currently optimised for musical stimuli, a representation of timbre could provide information about voice quality. The most significant changes would be required at the highest level of the blackboard, where speech-specific knowledge sources would be provided (for example, acoustic-phonetic models or higher-level knowledge about semantics and pragmatics). Future development of the model will focus upon the implementation of such knowledge sources.

Acknowledgements

We would like to thank three anonymous reviewers for their helpful comments, which significantly improved the presentation of the paper. DG

was supported by a BBSRC Research Studentship and GJB was supported by EPSRC grant GR/K18962.

References

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

Bregman, A.S., 1998. Psychological data and computational ASA. In: Rosenthal, D.F., Okuno, H.G. (Eds.), Computational Auditory Scene Analysis. Lawrence Erlbaum, NJ.

Bregman, A.S., Tougas, T.Y., 1989. Propagation of constraints in auditory organisation. Perception and Psychophysics 46, 395–396.

Brown, G.J., Cooke, M.P., 1994a. Computational auditory scene analysis. Computer Speech and Language 8, 297–336.

Brown, G.J., Cooke, M.P., 1994b. Perceptual grouping of musical sounds: A computational model. Journal of New Music Research 23, 107–132.

Cooke, M.P., 1993. Modelling Auditory Processing and Organisation. Cambridge University Press, Cambridge.

Cooke, M.P., Brown, G.J., Crawford, M., Green, P., 1993. Computational auditory scene analysis: listening to several things at once. Endeavour 17, 186–190.

Deutsch, D., 1980. The processing of structured and unstructured tonal sequences. Perception and Psychophysics 28, 381–389.

Ellis, D.P.W., 1996. Prediction-driven computational auditory scene analysis. PhD thesis, Massachusetts Institute of Technology.

Engelmore, R., Morgan, T., 1988. Blackboard Systems. Addison-Wesley, Reading, MA.

Erman, L.D., Hayes-Roth, F., Lesser, V.R., Reddy, D.R., 1980. The HEARSAY-II speech-understanding system: integrating knowledge to resolve uncertainty. Computing Surveys 12, 213–253.

Foley, J.D., van Dam, A., Feiner, S.K., Hughes, J.F., 1996. Computer Graphics: Principles and Practice. Addison-Wesley, London.

Glasberg, B., Moore, B.C.J., 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Research 47, 103–138.

Godsmark, D., 1998. A computational model of the perceptual organisation of polyphonic music. PhD thesis, Department of Computer Science, University of Sheffield.

Godsmark, D., Brown, G.J., 1996a. A blackboard model of auditory organisation I: primitive grouping. Proceedings of the Institute of Acoustics 18, 11–18.

Godsmark, D., Brown, G.J., 1996b. A blackboard model of auditory organisation II: timbral similarity. Proceedings of the Institute of Acoustics 8, 83–90.

Godsmark, D., Brown, G.J., 1998. Context-sensitive selection of competing auditory organisations: a blackboard model. In: Rosenthal, D.F., Okuno, H.G. (Eds.), Computational Auditory Scene Analysis. Lawrence Erlbaum, NJ.

Handel, S., 1995. Timbre perception and auditory object identification. In: Moore, B.C.J. (Ed.), Handbook of Perception and Cognition: Hearing. Academic Press, San Diego.

Hartmann, W.M., Johnson, D., 1991. Stream segregation and peripheral channeling. Music Perception 9, 155–184.

ISO, 1988. Normal equal-loudness level contours for pure tones under free-field listening conditions (ISO 226), International Standards Organisation.

Iverson, P., Krumhansl, C.L., 1993. Isolating the dynamic attributes of musical timbre. Journal of the Acoustical Society of America 94, 2595–2603.

Kashino, K., Tanaka, H., 1992. A sound source separation system using spectral features integrated by the Dempster's Law of combination. Annual Report of the University of Tokyo Engineering Research Institute 51, 67–72.

Kashino, K., Nakadai, K., Kinoshita, T., Tanaka, H., 1998. Application of Bayesian probability network to music scene analysis. In: Rosenthal, D.F., Okuno, H.G. (Eds.), Computational Auditory Scene Analysis, Lawrence Erlbaum, NJ.

Kluender, K.R., Jenison, R.L., 1992. Effects of glide slope, noise intensity, and noise duration on the extrapolation of FM glides through noise. Perception and Psychophysics 51, 231–238.

Lesser, V.R., Nawab, S.H., Klassner, F.I., 1995. IPUS: An architecture for the integrated processing and understanding of signals. Artificial Intelligence 77, 129–171.

Mellinger, D.K., 1991. Event formation and separation in musical sound. PhD thesis, Stanford University.

Miller, G.A., Heise, G.A., 1950. The trill threshold. Journal of the Acoustical Society of America 22, 637–638.

Nakatani, T., Okuno, H.G., Goto, M., Ito, T., 1998. Multi-agent based binaural sound stream segregation. In: Rosenthal, D.F., Okuno, H.G. (Eds.), Computational Auditory Scene Analysis. Lawrence Erlbaum, NJ.

Nii, P., 1988. Blackboard systems. In: Engelmore, R., Morgan, T. (Eds.), Blackboard Systems. Addison-Wesley, Reading, MA.

Patterson, R.D., Holdsworth, J., Nimmo-Smith, I., Rice, P., 1988. Implementing a gammatone filterbank. APU Report 2341, Cambridge, Applied Psychology Unit.

Rogers, W.L., Bregman, A.S., 1993. An experimental evaluation of three theories of auditory stream segregation. Perception and Psychophysics 53, 179–189.

Rosenthal, D.F., 1992. Emulation of human rhythm perception. Computer Music Journal 16, 64–76.

van Noorden, L.P.A.S., 1975. Temporal coherence in the perception of tone sequences. PhD thesis, University of Eindhoven.

Warren, R.M., 1984. Perceptual restoration of obliterated sounds. Psychological Bulletin 96, 371–383.