Proceedings of the Institute of Acoustics

MAPPING THE AUDITORY SCENE: TEMPORAL PROXIMITY AND 3-TONE STREAMING

Sheila M Williams, Kevin L Baker & Roderick I Nicolson

Departments of Psychology and Computer Science, University of Sheffield, Sheffield, S10 2TN.

1. INTRODUCTION

This paper describes an investigation into the action of temporal proximity in auditory grouping which forms part of a programme of work to develop a computational model of Auditory Scene Analysis, informed by psychoacoustic experimentation [1].

By analogy with visual perception, Auditory Scene Analysis is the perceptual task of deriving a useful representation of reality from the sensory input arising from the acoustic signal [2]. Part of this process involves the identification of auditory streams, the perceptual entities which represent individual acoustic events or separate sound sources.

Following an empirical programme which aimed to establish the metric or metrics by which frequency proximity acts in streaming the components of the acoustic signal [3], we are now investigating the interactive effects of temporal variation upon such metrics. It is known that at short time intervals (less than 145 ms.) the perception of duration of an auditorily marked time interval may be dependent on its surrounding context [4, 5]. As the time intervals normally employed in experiments based upon the streaming paradigm fall within this range we are currently investigating the potential effects on streaming of varying the onset time intervals between the captor and targets.

The present study focuses on the temporal aspects of the stimulus patterns used in our previous streaming experiments [3] and, by extending the experimental series to include comparisons with patterns of much shorter durations, introduces new evidence to support the hypothesis that temporal coherence of sequential tones may be mediated by more than one mechanism [6].

2. BACKGROUND

2.1. Frequency/Pitch Proximity

In a previous series of experiments at Sheffield we adapted the streaming methodology of Bregman and Pinker [7] in order to study competitive grouping effects and begin to establish a metric or series of metrics which would describe the action of frequency proximity in auditory scene analysis. In their 1978 series of four experiments, subjects were presented with variations of a pattern of three tones, repeated in the sequence ABCABC etc. The tonal pattern consisted of a single tone followed by a complex tone of two partials, both lower in frequency than the previous tone. The single tone was varied in frequency relative to the complex, and the lower tone of the complex was varied in synchrony relative to the middle tone. The objective was to study the effects of these variations on whether the listener heard a rapidly repeating sequence of high tones accompanied by a slow lower tone or an alternating pattern of rich and pure tones.

2.1.1. Possible Metrics. Our first experimental programme [3] began with the assumption that the metric which operates to stream simple auditory scenes according to the frequency proximity of their component tones would be one of the metrics for frequency distance already established. Several different frequency difference metrics have been proposed calculated by different methods.

The Log scale is an accurate description of the mathematical relationships between the frequencies of tones which belong to the same pitch class. It has been shown to hold in simple streaming relationships where two tones are alternated at very rapid rates at increasing frequency disparities [8]. The relationship between the frequency disparity and the centre frequency of each tone pair at the point where two streams become heard in preference to a single stream, maps to a straight line graph on log-frequency scales.

The Bark scale [9] relates to the physiological properties of the auditory system and is derived by integrating the function which describes an estimate of the equivalent rectangular bandwidths (ERBs) of the cochlea at given frequencies. More recent studies have confirmed that it is a good model for the relationship between frequency and the place of maximum response on the basilar membrane within the cochlea [10].

The Mel scale [11] is a subjective metric based on equivalent pitch ratios. It is intended to represent equivalent pitch steps and was obtained by applying two different measures whose results were in close agreement. In one method, subjects were asked to adjust 3 tones at equal intervals between two boundary tones and in the other, a second tone had to be adjusted until it sounded half as high in pitch as a given standard tone.

Thus, the log scale represents a mathematical description of the relationships between the source frequencies, although it has also been validated in simple streaming contexts, the Bark scale is based on the structure and responses of the auditory periphery and the Mel scale is derived as a measure of what the listener perceives to be an equivalent pitch distance.

2.1.2. Selecting a Metric. In an attempt to verify which of the suggested metrics operated in competitive grouping, we wanted to determine at which point the captor stopped streaming with one potential target and began to stream with another. That is, which of two possible targets was perceptually nearer to the captor for streaming purposes. So, instead of presenting the captor to stream out a particular tone selected from the two target tones, we varied the captor in frequency between that of the two target tones and presented the stimuli at such a rate that streaming always occurs for most subjects.

We used two sorts of stimuli, although our initial prediction was that we would obtain similar results in both cases. These are the sequential pattern, in which each of the tones is presented at a different time, and the synchrony pattern, in which the captor tone was presented alternately with the tone pair B and C synchronous.

In every case, tones B and C were fixed in frequency. For the first study tone B was set at 929Hz and tone C at 1973Hz. Tone A was presented at 15 levels varying between the frequencies of B and C. The target tones were not closely harmonically related and each captor tone was selected to be not within 3% of a simple harmonic ratio of either of the targets. As there was the possibility of an ordering effect in the sequential stimulus pattern, equal numbers of stimuli were prepared in the order BAC and CAB. We also used a variety of different tasks to investigate the effects of directing the subjects attention in particular ways, comparing two tasks in each study for both sets of target stimuli. Subjects were asked to judge the speed or relative frequency of the resulting streams. Each subject was only ever presented with one of the tasks.

We were surprised to find that the two stimulus patterns did give different results, $p<0.01$. The streaming midpoint for the synchrony stimulus pattern was significantly less than that for the sequential stimulus pattern. We were further surprised to find that neither of the results was particularly close to the values we had estimated for the midpoints using log, bark and mel scales. In each case our results showed that the streaming midpoint for simple competitive data was substantially lower than log, bark or mel midpoints.

2.2. Temporal Proximity

van Noorden [12] noted that perception of displacement from equal temporal intervals is more accurate for inter-stimulus onset intervals in excess of 120ms and that, at such rates of presentation, the amount of displacement necessary for the time shift to be perceived does not depend upon the temporal interval itself. For onset intervals below this rate, the minimum perceptible displacement appeared to be a function of both the inter-stimulus onset interval and the frequency disparity between the tones.

More recently, ten Hoopen and his colleagues have noted that in subjective estimates of short time periods, when an auditorily marked empty time interval which is as short as 120ms is preceded by a physically shorter neighbouring time interval, its duration is underestimated to a remarkable degree [4]. For longer intervals (240ms upwards) this effect is rarely found. Two alternative explanations were posited for this phenomenon. The first is that listeners have a tendency to perceive neighbouring time intervals as more similar in duration than they physically are and the second is that processing time is required to unbind the second marker from its nearest neighbour (the preceding marker) before it can be utilised as the perceptual start of the following interval. Thus, the disconnection time reduces the perceived duration of the interval [5].

Both of these investigations suggest that stimuli with onset intervals of less than around 120ms may result in qualitatively different perceptual phenomena than those with longer intervals. van Noorden's study [12] further demonstrated that in the simple alternating tone sequences, the temporal coherence boundary (the limit of the frequency disparity at which the alternating tones tended to be grouped as a single sequence) remained stable at around three semitones for periods less than 120ms, then rapidly increased to allow coherence over a range as great as 20 semitones for wider inter-stimulus onset intervals of 200-250ms.

2.3. Modelling the Processes.

In accordance with van Noorden's data, Beauvois and Meddis [6] have produced an auditory grouping model based on the ERB mechanism which accounts for both temporal and frequency proximity effects in predicting the assignment of alternating tones to one or two streams. The data from the model and experimental data from human subjects both demonstrate that temporal coherence occurs if the tones are below a certain frequency disparity or if their onset times are sufficiently far apart.

The outputs from the ERB model differ in each of these cases and Beauvois and Meddis suggest that this is due to the presence of an attentional mechanism which selects only the output channel with the largest excitation rate. In the case of tones presented very rapidly, the two sequential tones have to be sufficiently close to produce greatest excitation at a channel intermediate between the tones in order for them to be assigned to the same auditory stream. For tones presented more slowly, the attentional mechanism can switch between two output channels centred on the frequency of each of the tones and so the limits on their frequency disparity are less constrained.

For the rates of presentation of our data, 210 ms between onsets for the synchrony stimuli and 120 ms between onsets for the sequential data, the second of these mechanisms would predominate. This means that for shorter onset intervals between the tones in our stimuli, we might expect different streaming relationships to occur.

2.4. Auditory Scene Analysis.
While theories relating streaming to the physiological properties of the auditory periphery usually explain segregation of components in terms of a breakdown or failure of the frequency tracking mechanism, auditory scene analysis is seen as an active process in which components of the auditory scene are attributed to perceptual objects. Thus the alternating tone sequence is not either temporally *coherent* or *incoherent* but perceptually forms either one or two streams. If two streams are perceived, then temporal coherence exists between each tone and the repetition of itself in preference to a coherence which links successive tones in a rising and falling pattern.

The research described here forms part of a programme of investigation into the conflicts which arise in assigning auditory scene components to streams and the perceptual processes which operate to resolve these conflicts.

3. THE EXPERIMENTS

3.1. Introduction.
Our current experimental programme is intended to explore the interactive effect of temporal and frequency disparities on simple auditory scenes. The experiments described here were designed to compare the streaming midpoints found by Baker et al. [3] with those from stimuli with much shorter inter-stimulus intervals. In the original experiments two tonal patterns were used, each with different tone lengths and gaps to give similar subjective perceptions of two streams. This experiment compares the results obtained earlier with those obtained from similar stimuli with much shorter inter-stimulus onset intervals.

3.2. Subjects.
Five subjects, aged between 17 and 43 have so far taken part in these experiments.

3.3. Stimuli.
Both the synchrony and sequential patterns of tones were used (see figures 1 and 2 below). For the sequential pattern, equal numbers of patterns ordered BAC and CAB were presented.

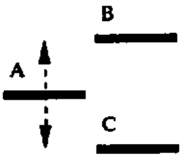

Figure 1. Synchrony pattern

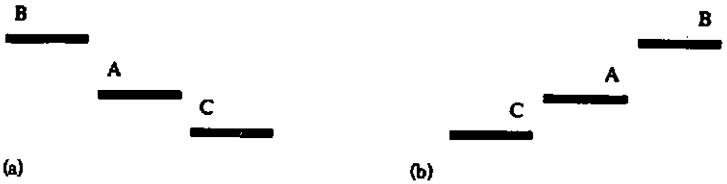

Figure 2. Sequential patterns (a) BAC and (b) CAB

Each tone consisted of a pure sine wave 30 ms steady state with 5 ms linear ramps either side. Inter-tone gaps were 10 ms long making the tone repetition time 50 ms. The value for the tones were those used in our previous study, with target tones at 929Hz and 1973Hz and captors at each of 15 values ranging betwen these extremes. In all, each subject listened to 90 trials, presented in random order and made up of 6 presentations with each captor value.

Each subject was asked to listen to the fastest beating stream and to indicate on a five point scale whether this was high or low in pitch when compared with the other tones they could hear. As the presentation rates used were close to the limit where temporal coherence may break down coherently, subjects were specifically asked to repond with a 3 if no stream appeared faster than any other.

3.4. Apparatus.
All stimuli were prepared and presented diotically at 20000 samples per second using a Viglen 386 PC running the MITSYN (1990) sound synthesis system. Each pattern of tones was stored as a signal file and presented as a stream through a play file which repeated the signal file 30 times. The signals were converted using a Data Translation DT2823 digital to analogue board and passed through a 8 KHz cutoff TTE passive filter to remove any digitization noise. The filtered signal was then attenuated to a level of 65dB (SPL) using an Advance type A64 stepped passive attenuator, and split to two channels for diotic listening. Subjects listened to the sounds through a pair of Beyer Dynamic DT48 headphones while seated in an Industrial Acoustics Co. acoustic chamber. A program written in Mitsyn Command Language was developed to present the stimuli in a random sequence and to collect and record the responses from the subjects' key-presses.

3.5. Results.
Preliminary results for the first five subjects have been obtained by plotting the average judgments for each subject as a function of the frequency of tone A. An estimate for the midpoints was made by using a curve-fitting technique to give a fifth order polynomial equation describing the best least squares fit of the data gathered for each data set. The equations describe the relationship between the average judgment and the frequency of tone A, the captor. We assume here that the streaming midpoint is indicated by an average judgment of '3' which may indicate that no coherent two-tone stream manifested itself or that the data at that point was equally likely to stream with the upper or lower target. The preliminary data suggests a midpoint of 1354 Hz for the synchrony pattern and 1346 Hz for the sequential pattern. This is close to the logarithmic mean of the data range.

3.6. Further experimentation.
As the preliminary results show a major disparity in the streaming midpoint from that obtained in our original experiments, we were concerned that the difference we had found there between the sequential and synchrony patterns might be solely attributable to the different inter-stimulus onset

times in the two patterns. We had, in each case, chosen the longest intervals which seemed to promote streaming into a single tone repetition stream and a two-tone stream. In the case of the sequential pattern, this was substantially shorter than for the synchrony pattern. We therefore repeated the experiment for the synchrony pattern using exactly the same intervals as originally used for the sequential pattern: ie. 12ms onset and offset ramps, 86ms steady state and 10ms gap between tones. Analysing the results for the first five subjects as described above, we estimate a midpoint of 1155Hz.

The following table shows the results, together with the results from the original frequency proximity experiments.

Freq. Range 929 to 1973 Hertz (Prelim.) Result

Inter-stimulus gap (in ms)	log	Bark	Mel	Seq	Sync
210	1354	1360	1390		1150*
120	1354	1360	1390	1250*	1155
50	1354	1360	1390	1346	1354

*from original experiment [3].

4. DISCUSSION

4.1. The Results.

The preliminary results of the experiments described above indicate that there is a substantial difference in the streaming midpoint for the range 929 to 1973 Hz between stimuli presented at an inter-stimulus onset time of 50ms and those presented at 120ms intervals. However, there is minimal difference (for the synchrony data pattern at least) between stimuli presented at 120ms and those presented at 210ms intervals. At the shortest of these intervals the midpoint is around the logarithmic mean which is also close to the Bark midpoint for this particular range. However at slower presentation rates the perceptual midpoint falls to around 1150 Hz at which level it appears to remian stable over some range. Further experimentation is clearly needed to trace out the pattern of these interactions but some interesting conclusions can be drawn from these results.

4.1.1. Inter-stimulus Intervals (Gaps). For both the 50ms and the 120ms onset disparities, every inter-tone gap was 10ms, so the recovery period (or masking period) was the same. Only the tones in the stimuli were extended and yet very different results were found. At 50ms intervals both synchrony and sequential patterns gave rise to very similar midpoints and these were close to the logarithmic mean originally predicted. At the 120ms intervals, the two stimulus patterns showed substantially different midpoints, neither of which were close to the mean.

4.1.2. The Effects of Masking.

A possible explanation suggested for our original data relates to the masking effects of each tone in the pattern upon the other elements in the configuration. Taking the possible effects of masking directly into account, we have to examine the 3 patterns separately. For the synchronous pattern in the original experiment, the possibility of backward masking is minimal due to the 70ms gaps between the elements of the pattern, also two-tone suppression is likely to be relatively low (f1/f2 = 1973/929 < 1.5), well outside the 1.5 maximum of the range suggested by Buser and Imbert

[13]. Forward masking might apply as an effect of the lower tone on the captor for low captor values and as an effect of the captor on the upper target for high captor values. In the middle of the range it is possible that these combined effects might serve to equalize the perceived loudness of both captor and upper target making them more similar to each other with respect to loudness than to the lower target tone. However, the combination of the 70ms gap and the relative distances between captor and targets towards the middle of the range would be unlikely to induce any major effect.

In the case of the 120 ms sequential patterns, two-tone suppression plays no part, but a significant difference between the two alternative tone orderings might be expected. The narrow gaps between the tones of 10ms might allow for a limited amount of backward masking to take place in the ACB pattern but this would be negligible compared with the expected forward masking predicted from the ABC pattern. Again, in the midrange, the lower target would be expected to suppress the captor which in turn might suppress the higher target. The effect would be low due to the relative frequency values in the center part of the range but stronger than in the case of the synchrony data due to the much smaller temporal distances between components. This possibility will be examined more thoroughly when data has been collected from more subjects.

For 120ms synchrony pattern presentations, the 10ms gaps would increase the likelihood of forward and backward masking, although it is possible that the shorter tone periods might balance out this effect. At shorter (50ms) intervals, the gaps remain the same and so the same effects might be predicted. However, it could be argued that the shorter tone lengths cause less stimulation of the auditory periphery and thus less masking effect.

In any case, it seems surprising that two-tone masking should be operating to differentiate the midpoints between the two stimulus patterns at one rate of presentation and not at another.

4.3. The Two Mechanisms Debate.

The results shown here indicate a clear difference between the streaming of three-tone patterns at 50ms inter-stimulus onset presentations and at 120ms presentations. Both the higher streaming midpoint and the co-incidence of midpoints for the two stimulus patterns distinguish the faster presentations from the slower presentations we had previously applied. The value of 50ms for the inter-stimulus onset period was specifically chosen as this was the fastest rate of presentation considered for two-tone streaming by Beauvois and Meddis [6] in their model of the auditory periphery. They concluded that at such rates, streaming occurs due to the combined effects of alternating tones on a channel whose peak response is to a frequency value intermediate between that of the two tones and, for the purposes of their model, assumed this to be at the log midpoint between the tones. At slower rates, such as the 120ms presentations employed here, streaming is maintained by the attentional mechanism switching from one to the other channel centred on the actual frequency value of the stimulus tones. The switching mechanism permits two-tone streaming over much larger frequency ranges. Thus, it may be that this switching follows different criteria for temporal proximty. However the range of tones employed in the current study is rather larger than that to which their model was applied. Nevertheless, the differences exhibited here must offer support for the two mechanisms hypothesis.

5. CONCLUSION

A preliminary analysis of our data indicates that different rates of presentation produce clear perceptual differences for three-tone streaming in the rage 929 to 1973 Hz. Very fast rates of presentation indicate a streaming midpoint at around the logarithmic mean and do not distinguish between target tones which are synchronous and those which do not overlap in time. Slower rates show significantly different streaming midpoints for the synchrony and sequential stimulus patterns, both of which are substantially below the logarithmic mean. These results support the hypothesis that more than one mechanism is operating to promote auditory streaming by frequency proximity although auditory masking effects may also play a role in the analysis of simple three-tone auditory scenes.

6. ACKNOWLEDGEMENTS

This research was supported by the SERC/MRC/ESRC Joint Council Initiative into Cognitive Science/HCI, grant number SPG8921799.

7. REFERENCES

[1] S M WILLIAMS, R I NICOLSON & P D GREEN, 'STREAMER: Mapping the Auditory Scene', Proc. IOA Autumn Conf. in Speech and Hearing, Windermere, (1990).
[2] A S BREGMAN, 'Auditory Scene Analysis', MIT Press, (1990).
[3] K L BAKER, S M WILLIAMS & R I NICOLSON, 'Frequency Proximity in Auditory Grouping: Eliciting the metric', 25th Int. Congress of Psychology, Brussels, (1992).
[4] Y NAKAJIMA, T SASAKI, R G H VAN DER WILK & G TEN HOOPEN, 'A new illusion in time perception - I', 1st Int. Conf. on Music Perception and Cognition, Kyoto, Japan, (1989).
[5] G TEN HOOPEN, G VIS, G HILKHUYSEN & Y NAKAJIMA, 'A new illusion in time perception - II', 1st Int. Conf. on Music Perception and Cognition, Kyoto, Japan, (1989).
[6] M W BEAUVOIS, R MEDDIS, 'A computer model of auditory stream segregation', The Quarterly Journal of Experimental Psychology, 43, 517-41, (1991).
[7] A S BREGMAN, S PINKER, 'Auditory Streaming and the Building of Timbre', Canadian Journal of Psychology, 32 , pp.19-31, (1978).
[8] G A MILLER & G A HEISE, 'The trill threshold', Journal of the Acoustical Society of America, 22, 637-8, (1950).
[9] B C J MOORE & B R GLASBERG, 'Suggested formulae for calculating auditory-filter bandwidths and excitation patterns', JASA, 74, 750-753, (1983).
[10] D D GREENWOOD, 'A cochlear frequency-position function for several species - 29 years later', JASA, 87, 2592-2605, (1990).
[11] S S STEVENS & J VOLKMANN, 'The relation of pitch to frequency: a revised scale', American Journal of Psychology, 53, 329-353, (1940).
[12] L P A S VAN NOORDEN, 'Temporal coherence in the perception of tone sequences', Unpublished doctoral dissertation, Eindhoven University of Technology, (1975).
[13] P BUSER & M IMBERT, "Audition", Cambridge: MIT Press, (1991).