Chapter 45
Sequential and Simultaneous Auditory Grouping Measured with Synchrony Detection

Christophe Micheyl, Shihab Shamma, Mounya Elhilali, and Andrew J. Oxenham

Abstract Auditory scene analysis mechanisms are traditionally divided into "simultaneous" processes, which operate across frequency, and "sequential" processes, which bind sounds across time. In reality, simultaneous and sequential cues often coexist, and compete to determine perceived organization. Here, we study the respective influences of synchrony, a powerful grouping cue, and frequency proximity, a powerful sequential grouping cue, on the perceptual organization of sound sequences (Experiment 1). In addition, we demonstrate that listeners' sensitivity to synchrony is dramatically impaired by stream segregation (Experiment 2). Overall, the results are consistent with previous results showing that prior perceptual grouping can influence subsequent perceptual inferences, and show that such grouping can strongly influence sensitivity to basic sound features.

Keywords Auditory scene analysis • Stream segregation • Performance measures • Timing • Synchrony

45.1 Introduction

Research on auditory scene analysis has led to the identification of several "cues," which the auditory system can use to organize sounds perceptually (Bregman 1990; Darwin and Carlyon 1995). Synchrony, harmonicity, and frequency proximity are among the most important such cues. Synchronicity and harmonicity are used primarily to group simultaneous spectral components across frequency, whereas frequency proximity has an important role in binding sequential elements across time.

An important question currently facing psychophysicists is how these grouping cues interact with each other in order to determine the "correct" perceptual organization of acoustic scenes, which typically contain a multiplicity of such cues. Earlier studies have revealed that sequential grouping based on frequency proximity

C. Micheyl (✉)
Department of Psychology, University of Minnesota, Minneapolis, MN, USA
e-mail: cmicheyl@umn.edu

E.A. Lopez-Poveda et al. (eds.), *The Neurophysiological Bases of Auditory Perception*,
DOI 10.1007/978-1-4419-5686-6_45,

can counteract simultaneous grouping based on synchrony (e.g., Darwin et al. 1995; 1989; Shinn-Cunningham et al. 2007). For instance, a series of elegant experiments by Darwin and colleagues (e.g., Darwin et al. 1995; 1989) have demonstrated that "precursor" tones at the same frequency as a "target" component in a complex tone "captured" the target into a separate stream, thereby reducing its influence on the pitch or timbre of the complex.

The present study was inspired by these earlier findings and addressed two questions. The first question was whether sequential grouping affects listeners' ability to detect synchrony. The results of several previous studies suggest that listeners are unable to accurately perceive the temporal relationships between sounds across auditory streams. In particular, listeners cannot accurately discriminate the duration of temporal intervals between consecutive tones (Vliegen et al. 1999; Roberts et al. 2002), or correctly identify the temporal order of these tones (Bregman and Campbell 1971), under conditions where the tones are heard in separate streams. However, in all of these studies, the tones never overlapped in time. The situation might be quite different with synchronous tones, because synchrony detection appears to involve different mechanisms than temporal order identification, or temporal interval discrimination (Mossbridge et al. 2006). For instance, while synchrony detection could in principle be achieved using widely tuned neural coincidence detectors (Oertel et al. 2000), temporal interval discrimination require mechanisms for measuring the elapsed time between events. These various mechanisms, possibly taking place at different stages of processing in the auditory system, could be differently affected by sequential grouping. The finding that detrimental effects of sequential grouping generalize to synchrony detection would provide evidence that listeners' access to the output of coincidence detectors is strongly constrained by perceptual organization mechanisms.

The second question addressed in this study is whether across-frequency grouping based on synchrony predominates over sequential grouping based on frequency proximity. In order to answer this question, we measured listeners' thresholds for the detection of an asynchrony between two tones at different frequencies, A and B, preceded by a series of either synchronous or asynchronous "precursor" tones at the same two frequencies, A and B. We reasoned that, if across-frequency grouping due to synchrony predominates over segregation due to frequency separation, thresholds should be lower with synchronous precursors than with asynchronous precursors.

45.2 Experiment 1: Sequential Capture Overrides Synchrony Detection

45.3 Methods

Schematic spectrograms of the stimulus conditions tested in this experiment are shown in Fig. 45.1a. The basic stimulus elements were 100 ms pure tones at two frequencies, A, which was fixed at 1,000 Hz, and B, which was set 6 or 15

semitones above A. In the baseline, "No captor" condition (upper panel in Fig. 45.1a), only these two A and B tones were present. In one observation interval, the tones were synchronous; in the other, the B tone was delayed or advanced by Δt ms relative to the A tone. The task of the listener was to indicate in which observation interval the A and B tones were asynchronous.

Two other conditions were tested. In the "On-frequency captor" condition (middle panel in Fig. 45.1a), the A and B pair was surrounded by "captor" tones at the A frequency, with five captor tones before, and two captor tones after, the A–B pair. The captor tones were separated from each other, and from the target A tone, by a constant gap of 50 ms. Thus, in this condition, the target A tone formed part of a temporally regular sequence. In the final, "Off-frequency captor" condition (lower panel in Fig. 45.1a), the frequency of the captor tones was set to six semitones below that of the A tone. The listener's task was the same as in the baseline condition: to indicate in which of the two observation intervals presented on a trial the target A and B tones were asynchronous. A three-down one up adaptive procedure was used to measure thresholds, with Δt as the tracking variable. Each listener completed at least four threshold measurements in each condition. The data shown here are geometric mean thresholds across listeners.

In all experiments described here, the stimuli were generated digitally and played out via a soundcard (Lynx Studio L22) with 24-bit resolution and a sampling frequency of 32 kHz, and presented to the listener via the left earpiece of Sennheiser HD

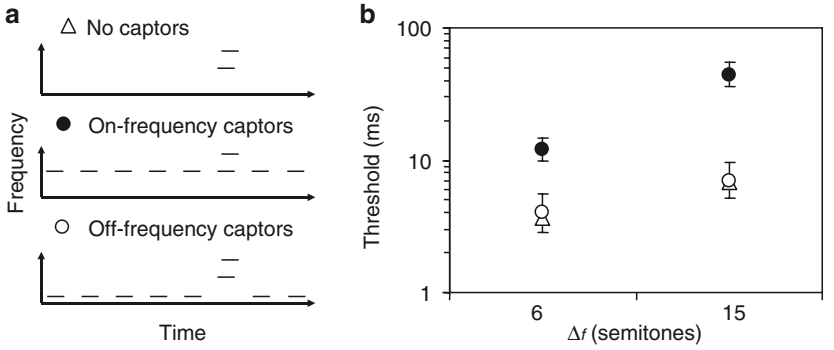

Fig. 45.1 (**a**) Schematic spectrograms of the stimuli in Experiment 1. The small horizontal bars represent 100-ms tones. In the baseline ("No captors") condition (*top panel*), the stimuli were a fixed 1,000-Hz tone, A, and a (6- or 15-semitone) higher-frequency tone, B. In one of the two observations intervals on a trial, the onset of the B tone was delayed (as shown here) or advanced (not shown) by Δt ms relative to that of the A tone; in the other observation interval, the two tones were synchronous (not shown). In the "On-frequency captors" condition (*middle panel*), the target A and B tones were preceded by five, and followed by two "captor" tones at the A frequency (1,000 Hz). Consecutive captors tones were separated from each other, or from the A tone, by a fixed, 50 ms silent interval. In the "Off-frequency captors" condition, the frequency of the captor tones was set to six semitones below that of the A tone, being equal to approximately 707 Hz. (**b**) Thresholds for the detection of an asynchrony between the target A and B tones in the different stimulus conditions shown on the left, for the two A–B frequency separations (6 and 15 semitones). Each data point was obtained by averaging thresholds across listeners. The error bars show geometric standard errors of the mean

580 headphones. Listeners were seated in a double-walled sound-attenuating chamber (Industrial Acoustics Company). The level of the tones was set to 60 dB SPL.

Eight listeners took part in this experiment. All had normal hearing (i.e., pure tone thresholds lower than 15 dB HL at octave frequencies between 500 and 6 kHz).

45.4 Results and Discussion

The results of this experiment are shown in Fig. 45.1b. Significantly larger thresholds were observed in the "On-frequency captors" condition than in both the "No-captor" [$F(1, 7) = 19.96$, $p = 0.003$], and "Off-frequency captors" [$F(1, 7) = 13.47$, $p = 0.008$] conditions. In fact, at the largest (15-semitone) A–B frequency separation, thresholds in the presence of the on-frequency captors were occasionally at ceiling (100 ms). Thresholds in the "Off-frequency captors" and "No-captors" conditions were not statistically different, and generally low (3–5 ms), consistent with earlier findings (e.g., Zera and Green 1993). Finally, thresholds were generally larger at the largest (15 semitones) A–B frequency separation (Δf) than at the smaller one (6 semitones) [$F(1, 7) = 81.11$, $p < 0.0005$], an effect that was larger in the "On-frequency captors" condition than in the other two conditions [$F(1, 7) = 22.11$, $p = 0.002$].

The results of this experiment are consistent with earlier findings, which used "capture" effects to demonstrate an influence of sequential grouping (based on frequency proximity) on the perception of temporal relationships between sounds (Bregman and Campbell 1971; O'Connor and Sutter 2000). The current results reveal that the detection of synchrony (or asynchrony) is no more immune to sequential grouping influences than other temporal discrimination abilities (e.g., Broadbent and Ladefoged 1959; Roberts et al. 2002, 2008; Vliegen et al. 1999). Thus, while there is physiological evidence for the existence of "coincidence detectors" or "synchrony detectors" in the auditory system (Oertel et al. 2000), the present findings suggest that listeners' conscious access to the outputs of these detectors is constrained by perceptual organization mechanisms.

45.5 Experiment 2: Synchrony Overrides Sequential Grouping

45.6 Methods

The stimuli used in this experiment are illustrated schematically in Fig. 45.2a. They were sequences of A and B tones, where A and B represent different frequencies. The frequency of the A tone was kept constant at 1,000 Hz. The frequency of the B tone was set 6, 9, or 15 semitones above that of the A tone. Each sequence consisted of five "precursor" tones at each frequency (i.e., five A tones and five B

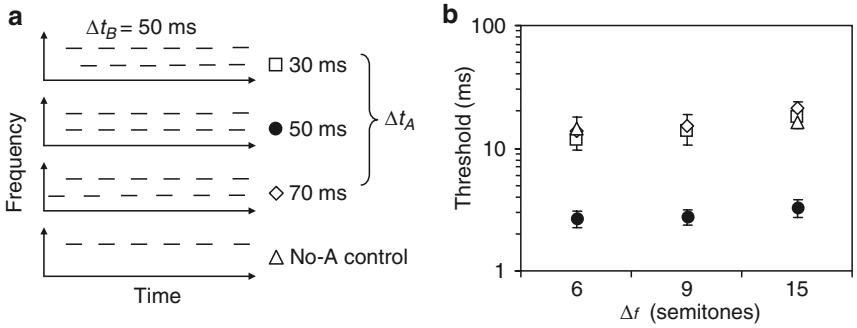

Fig. 45.2 (**a**) Schematic spectrograms of the stimuli in Experiment 2. The stimuli were sequences of 100-ms pure tones (shown here as small horizontal bars) at two different frequencies, A and B, separated by 6, 9, or 15 semitones. Except for the last two, tones at the higher (B) frequency were separated by a constant ITI, Δt_B, of 50 ms. Depending on the condition, tones at the lower (A) frequency were either present and separated by an ITI, Δt_A, of 30 ms (*top panel*), 50 ms (*second panel from top*), or 70 ms (*second panel from bottom*), or they were absent (No-A control condition, *lower panel*). In the condition where Δt_A and Δt_B were both equal to 50 ms, all A and B tones except the last two were synchronous. Depending on the observation interval, the last ("target") A and B tones were either asynchronous (as shown here) or synchronous (not shown here). In the former case, they were separated by a variable delay, Δt_{AB}, which was controlled by the adaptive threshold-tracking procedure. The task of the listener was to indicate the observation interval containing the delay Δt_{AB}. (**b**) Thresholds in Experiment 2. Thresholds measured in the different conditions illustrated in Fig. 45.2a are shown using different symbols (as indicated in Fig. 45.2a). Each data point was obtained by averaging thresholds across listeners. The error bars show geometric standard errors of the mean

tones), followed by two "target" tones (i.e., one A tone and one B tone). Each tone was 100 ms in duration, including 10-ms raised-cosine onset and offset ramps. The duration of the silent interval between two consecutive B precursors, Δt_B, was fixed at 50 ms. The inter-tone interval (ITI) between A precursors, Δt_A, varied across conditions; it was equal to 50 ms (in which case, the A and B precursors were synchronous, as shown in the middle panel of Fig. 45.2a), 30, or 70 ms (in which cases, the A and B precursors were asynchronous; see upper and lower panels in Fig. 45.2a).

The precursors were followed by a pair of "target" A and B tones, which were either synchronous, or asynchronous. In the latter case, the B tone randomly led or lagged the A tone by an amount, Δt, which was varied adaptively by the tracking procedure used to measure the threshold. Fig. 45.2a illustrates the case of a lagging B tone. In all cases, the interval between the A target and the preceding A precursor was the same as that between two consecutive A precursors. Importantly, the A and B sequences were always positioned in time relative to each other in such a way that the target A and B tones were synchronous in one of the two observation intervals presented during a trial, and shifted by plus or minus Δt in the other observation interval. In addition, a control condition was run, in which the A tones were turned off, and the B tones were generated in exactly the same way as described above.

Thresholds for the detection of an asynchrony between the target A and B tones were measured using a two-interval, two-alternative forced-choice (2I-2AFC) procedure with an adaptive three-down one-up rule. Listeners had to indicate, on each trial, which of the two presented tone sequences (separated by a silent gap of 500 ms) contained asynchronous target A and B tones at the end. The order of presentation of the two sequences was randomized. At the beginning of each adaptive run, the tracking variable, Δt, was set to 20 ms. It was divided by a factor c after three consecutive correct responses, and multiplied by that same factor after each incorrect response. The value of c was set to four at the beginning of the adaptive run; it was reduced to two after the first reversal in the direction of tracking (from decreasing to increasing), and to $\sqrt{2}$ after a further two reversals. The procedure stopped after the sixth reversal with the $\sqrt{2}$ step size. Threshold was computed as the geometric mean of Δt at the last six reversal points. Each listener completed at least four threshold measurements in each condition. The data shown here are geometric mean thresholds across listeners.

Nine listeners with normal hearing (i.e., pure-tone hearing thresholds of 15 dB HL or less at octave frequencies between 500 and 8,000 Hz) took part in this experiment.

45.7 Results and Discussion

The results of this experiment are shown in Fig. 45.2b. In the condition in which the A and B precursor tones were synchronous (50 ms ITI at both frequencies), thresholds (indicated by filled circles) were generally small (around 3 ms), even at the largest A–B frequency separation tested (15 semitones). In contrast, in conditions in which the nominal duration of the ITI in the A-tone stream was shorter (30 ms) or longer (70 ms) than that the ITI in the B-tone stream, so that the precursor A and B tones were presented asynchronously and at different tempi, thresholds were considerably larger (10–20 ms). The difference was highly statistically significant [$F(1, 10) = 7.394$, $p < 0.001$].

The finding of relatively large thresholds in the conditions in which the precursor A and B tones were asynchronous is consistent with other results in the literature (e.g., Bregman and Campbell 1971; Vliegen et al. 1999; Roberts et al. 2002), which indicate that listeners cannot accurately judge the relative timing of sounds across streams. In fact, the thresholds measured in those two conditions were not significantly different from those measured in the control condition, in which the A tones were turned off (upward pointing triangles), and the only cue available for task performance was a temporal irregularity in the B stream (i.e., a longer or shorter ITI between the last two B tones than between previous tones). This suggests that in conditions in which the A and B tones formed two separate streams, performance was based on a within-stream cue, rather than on across-stream timing comparisons.

The finding of consistently low thresholds in conditions involving synchronous precursor tones indicates that in this condition, listeners were able to make accurate

timing judgments between the A and B tones. This suggests that in these conditions, the A and B tones formed a single stream. Thresholds increased somewhat with the A–B separation [$F(1, 10) = 5.73$, $p < 0.001$], indicating that synchrony-based grouping did not completely override the effect of frequency separation. However, even at the largest frequency separation tested (15 semitones), they were still quite small, and considerably smaller than in the other (asynchronous precursors or no-A-tones control) conditions. This result is noteworthy, because it demonstrates that spectral components separated by more than an octave can still be grouped in perception if they are synchronous or quasi-synchronous; this provides further evidence that the auditory system can accurately detect synchrony, or lack thereof, across widely separated frequencies (Mossbridge et al. 2006).

45.8 Conclusions

The results of this study indicate that while sequential grouping can prevent synchrony-based grouping, and dramatically impair listeners' ability to detect asynchrony (Experiment 1), on the other hand, synchrony can prevent stream segregation based on frequency separation, and lead to perceptual grouping of spectral components at remote frequencies (Experiment 2). How can these apparently contradictory results be reconciled?

A possible explanation is based on Bregman's (1990) "old plus new" heuristic. According to this explanation, whichever grouping cue is introduced first dominates the subsequent organization of the auditory scene. Thus, in situations where sequential grouping cues precede simultaneous grouping cues, sequential grouping overrides simultaneous grouping (Experiment 1); conversely, in situations where simultaneous grouping cues (such as synchrony) are introduced before sequential grouping cues, simultaneous grouping predominates (Experiment 2).

Some findings in the psychoacoustic literature suggest that the "old plus new" heuristic cannot be the whole story, however. For instance, Dau et al. (2009) found that comodulation masking release (CMR) was eliminated when the flankers were followed by spectrally similar "post-cursors," with which they formed a sequential stream, separate from the signal that the listener had to detect. This suggests that if we had presented only postcursor tones (no precursors) in Experiment 1, sequential capture may have also occurred, and qualitatively similar (albeit perhaps weaker) results would have been obtained.

Because grouping effects induced by postcursors cannot be explained simply in terms of the old-plus-new heuristic, a more general account of perceptual grouping is needed. Elhilali et al. (this volume) describe a computational model, which can account for these and other psychophysical results on simultaneous and sequential grouping in auditory scene analysis.

Acknowledgments This work was supported by the National Institutes of Health (R01 DC 07657). Cynthia Hunter is acknowledged for assistance with data collection. More details

concerning the methods and results of Experiment 2 may be found in Elhilali et al. (2009). Experiment 1 formed part of a broader study, the results of which are described in Micheyl et al (2010).

References

Bregman AS (1990) Auditory scene analysis: the perceptual organization of sound. MIT Press, Cambridge

Bregman AS, Campbell J (1971) Primary auditory stream segregation and perception of order in rapid sequences of tones. J Exp Psychol 89:244–249

Broadbent DE, Ladefoged P (1959) Auditory perception of temporal order. J Acoust Soc Am 31:151–159

Darwin CJ, Carlyon RP (1995) Auditory grouping. In: Moore BCJ (ed) Hearing. London, Academic Press, pp 387–424

Darwin CJ, Hukin RW, al-Khatib BY (1995) Grouping in pitch perception: evidence for sequential constraints. J Acoust Soc Am 98:880–885

Darwin CJ, Pattison H, Gardner RB (1989) Vowel quality changes produced by surrounding tone sequences. Percept Psychophys 45:333–342

Dau T, Ewert S, Oxenham AJ (2009) Auditory stream formation affects comodulation masking release retroactively. J Acoust Soc Am 125:2182–2188

Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma AS (2009) Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron 61:317–329

Micheyl C, Hunter CH, Oxenham AJ (2010) Auditory stream segregation and the perception of across-frequency synchrony. J Exp Psychol Hum Percept Peform. In press

Mossbridge JA, Fitzgerald MB, O'Connor ES, Wright BA (2006) Perceptual-learning evidence for separate processing of asynchrony and order tasks. J Neurosci 26:12708–12716

O'Connor KN, Sutter ML (2000) Global spectral and location effects in auditory perceptual grouping. J Cogn Neurosci 12:342–354

Roberts B, Glasberg BR, Moore BC (2002) Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. J Acoust Soc Am 112:2074–2085

Roberts B, Glasberg BR, Moore BC (2008) Effects of the build-up and resetting of auditory stream segregation on temporal discrimination. J Exp Psychol Hum Percept Perform 34:992–1006

Oertel D, Bal R, Gardner SM, Smith PH, Joris PX (2000) Detection of synchrony in the activity of auditory nerve fibers by octopus cells of the mammalian cochlear nucleus. Proc Natl Acad Sci U S A 97:11773–11779

Shinn-Cunningham BG, Lee AK, Oxenham AJ (2007) A sound element gets lost in perceptual competition. Proc Natl Acad Sci U S A 104:12223–12227

Vliegen J, Moore BC, Oxenham AJ (1999) The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. J Acoust Soc Am 106:938–945

Zera J, Green DM (1993) Detecting temporal onset and offset asynchrony in multicomponent complexes. J Acoust Soc Am 93:1038–1052