Temporal coherence versus harmonicity in auditory stream formation

Christophe Micheyl
Department of Psychology, University of Minnesota, Minneapolis, Minnesota 55455
cmicheyl@umn.edu

Heather Kreft
Department of Otolaryngology, University of Minnesota, Minneapolis, Minnesota 55455
plumx002@umn.edu

Shihab Shamma
Ecole Normale Supérieure, Paris 75230, France
sas@umd.edu

Andrew J. Oxenham
Departments of Psychology and Otolaryngology, University of Minnesota, Minneapolis, Minnesota 55455
oxenham@umn.edu

Abstract: This study sought to investigate the influence of temporal incoherence and inharmonicity on concurrent stream segregation, using performance-based measures. Subjects discriminated frequency shifts in a temporally regular sequence of target pure tones, embedded in a constant or randomly varying multi-tone background. Depending on the condition tested, the target tones were either temporally coherent or incoherent with, and either harmonically or inharmonically related to, the background tones. The results provide further evidence that temporal incoherence facilitates stream segregation and they suggest that deviations from harmonicity can cause similar facilitation effects, even when the targets and the maskers are temporally coherent.

PACS numbers: 43.66.Ba, 43.66.Fe, 43.66.Mk [QJF]
Date Received: October 10, 2012 **Date Accepted:** January 16, 2013

1. Introduction

Research on auditory scene analysis has identified various factors that govern the perceptual organization of sounds into "streams" (Bregman, 1990). However, the relative importance of these factors is still a matter of debate. In particular, while several psychophysical, neurophysiological, and modeling studies of auditory streaming performed during the last thirty years have focused on the importance of spectral—or tonotopic—contrasts for stream segregation (e.g., Hartmann and Johnson, 1991; for a review, see Shamma and Micheyl, 2010), recent work has emphasized the role of temporal coherence (Elhilali and Shamma, 2008; Elhilali *et al.*, 2009; for a review, see Shamma *et al.*, 2010). In this context, the notion of temporal coherence extends that of synchrony, and refers specifically to the *repeated* synchronous activation of auditory "channels" (or neural populations) tuned to different sound parameters, e.g., different frequencies, or different sound features, e.g., pitch and spatial location.

According to a *strong* version of the temporal-coherence theory (TCT), temporal coherence overrides all other stream-formation factors, such as frequency separation and harmonicity, and it produces obligatory stream integration, i.e., it precludes stream segregation (Shamma *et al.*, 2010). This strong interpretation leads to two strong predictions: First, listeners should be unable to segregate a repeating sequence

of "target" sounds from "background" sounds, if the target and background sounds activate auditory channels (or neurons) in a temporally coherent fashion; second, the inability to segregate the sequences should occur whether or not the targets and maskers are harmonically related.

This study sought to provide a test of these two predictions, using psychophysical performance measures. Listeners were given a task that required "hearing out" a sequence of target tones embedded in a multi-tone background, and judging the direction of a frequency change at the end of the sequence. Depending on the condition tested, the background tones were either temporally coherent or incoherent with the targets, and they were either harmonically related to the targets or not—with the exception of the final burst in each sequence, for which the frequencies of the masker tones were always randomly jittered. In general, repeating target tones can be "heard out" from a multi-tone background when the targets form a separate stream (e.g., Kidd *et al.*, 1994; Kidd *et al.*, 2003; Micheyl *et al.*, 2007). We therefore reasoned that listeners would show good performance only in conditions in which they were able to hear the target tones as a separate stream. Thus, the strong TCT predicts good performance in all conditions in which the target and background tones (up to the penultimate burst) were temporally incoherent, and poor performance in all conditions in which the target and background tones were temporally coherent. Moreover, the strong TCT predicts that harmonic relationships—or lack thereof—between the target and background tones would have either no effect, or a small effect compared to that of temporal coherence between targets and maskers.

2. Methods

2.1 Subjects

Eight subjects (aged 19 to 33 years) with normal hearing (audiometric pure-tone thresholds less than 20 dB hearing level (HL) between 0.25 and 8 kHz in octave steps) completed the experiment. Subjects provided written informed consent and were paid for their participation. All subjects except for one had received music lessons and/or played an instrument (for 7–18 years).

2.2 Stimuli and task

On every trial, listeners were presented with a sequence of nine target tones (Mm. 1). The first eight target tones (or "precursors") had the same frequency, f_{ref}, selected at random (with equal probability) from a list of 16 values, which ranged from 0.5 octave below to 1 octave above 1 kHz in 0.1-octave steps. The frequency of the last (ninth) target tone ("test") was shifted randomly up or down (with equal probability) by 2% relative to f_{ref}. The task was to indicate the direction of the shift.

Mm. 1. TargetAlone. This is a file of type "wav" (218 Kb).

The target tones were accompanied by background tones [Fig. 1(A)]. The background tones were 100 ms in duration each and they were presented either synchronously with 100-ms targets ("Sync"), or 40 ms before each 60-ms target ("Async," for "asynchronous"); in all cases the offsets of the targets and background tones were synchronous. The background tones were spaced evenly on either a harmonic scale (H) or on a logarithmic scale (L), or they were "shifted" (S) complexes which were produced by shifting all harmonics (upward for four subjects, downward for the four others) by 25% of the F0 (as in Micheyl *et al.*, 2010). Manipulations L and S produced inharmonic maskers, but in the latter case, masker components were still spaced equally on the linear (Hz) frequency scale. In the H conditions, the target and background tones were harmonics of a fundamental frequency, f_0, which was equal to f_{ref}/N, where N was randomly set to 3, 4, 5, or 6, with equal probability; only harmonics with frequencies lower than $1.9f_{ref}$ were included in the stimulus. Depending on the condition being tested, N

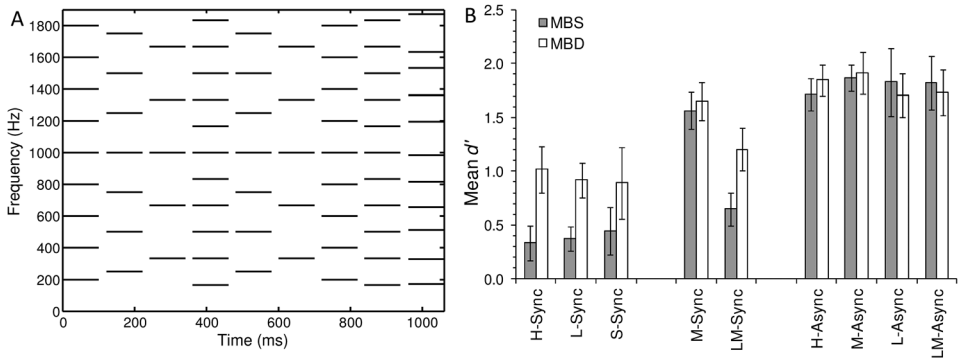

Fig. 1. Stimuli and results. (A) Schematic spectrogram of example harmonic MBD stimuli; see text for details. (B) Mean d' for the different test conditions. Error bars show standard errors of the mean (across eight listeners).

was constant within each trial (Mm. 2) (conditions denoted "MBS" for "multiple bursts same," following Kidd *et al.*, 1994; Kidd *et al.*, 2003), or was varied randomly across bursts within trial (Mm. 3) (conditions denoted "MBD" for "multiple bursts different"), with the constraint that two consecutive N values could not be the same, with the exception of the penultimate and final burst, for which N was always the same. It is important to note that since the components were sufficiently far part to be mostly individually resolved in the auditory system, in the Sync conditions, the MBS presentation mode presumably resulted in temporally coherent activation of all responsive frequency-selective channels or neurons (at least, up to the penultimate burst in each sequence), whereas the MBD presentation mode resulted in temporally incoherent activation of those channels (Shamma *et al.*, 2010). In Async conditions, target and masker channels were always activated in a temporally incoherent fashion.

Mm. 2. HSync_MBS. This is a file of type "wav" (218 Kb).

Mm. 3. HSync_MBD. This is a file of type "wav" (218 Kb).

In L conditions, the frequencies of the background tones were drawn from a list of values chosen pseudo-randomly so that (a) the target and maskers were equally spaced on a log scale; (b) the geometric-mean spacing of the components was equal to the geometric-mean spacing of the target and two adjacent maskers in the H condition.

"Mistuned-target" conditions (denoted by the letter M) were also tested. For those conditions, the stimuli were generated in the same way as for the H or L conditions, except that, prior to stimulus presentation, the frequency of the target tones was shifted upwards by 4% relative to its reference H or L position. Thus, in M conditions, the masker components were still harmonically related to each other, but they were no longer harmonically related to the target (Mm. 4). L-M conditions were also tested to provide a control: since for all L conditions the tones were inharmonic, no effect of mistuning was expected in those conditions.

Mm. 4. MSync_MBS. This is a file of type "wav" (218 Kb).

To limit the listener's ability to respond correctly based on a "global" percept evoked by the target and background tones in conditions where the target tones were not heard out as a separate stream, on the last (ninth) burst of each trial, the frequencies of each background tone was shifted by 2% in a random direction (up or down),

independently from the direction of the shift in the frequency of the target and of the other background tones.

All tones were presented at a fixed level of 60 dB sound-pressure level (SPL) per tone, with random starting phases, and 10-ms (Hanning) onset and offset ramps. Masker tones were separated by a silent gap of 20 ms; the onset times of the targets were adjusted relative to those of the masker tones to produce synchronous or asynchronous onsets, as described above. Example stimuli can be heard using the attached sound (wav) files.

2.3 Procedure

Fourteen subjects participated first in two "screening" tests, which were designed to check whether they could reliably perform the basic task. In the first screening test, the targets were presented with no background tones. Subjects were instructed to indicate whether the frequency of the last tone was shifted upward or downward relative to the preceding tones. The second screening test involved the stimuli from the M-Async MBD condition; we reasoned that this condition would be easiest for the subjects, since it contained two cues for distinguishing the targets from the maskers (mistuning and asynchrony). For this test and all subsequent tests, subjects were instructed to try to hear out the regularly repeating, constant-frequency target tone, and to indicate whether its frequency was shifted upward or downward at the end. Each screening test involved 100 trials. Subjects were excluded from the study if they were unable to achieve at least 90% correct by the third 25-trial block of the first screening test, and 70% correct by the second block of the second screening test. Five subjects were excluded based on these criteria. Nine remaining subjects performed two additional practice tests (100 trials each) prior to the main experiment. These practice tests involved stimuli similar to those of the M-Sync MBD and M-ASync MBS conditions. These conditions were specifically selected for the practice because they contained cues that facilitated the detection of the target which, we reasoned, would help listeners familiarize themselves with the task. To check for practice (learning) effects, listeners were re-tested on H-Sync (MBS and MBD) conditions near the end of the study; no statistically significant learning effect was found.

During the main experiment, conditions involving harmonic maskers were tested first; conditions L and S were tested in a subsequent phase. Eight subjects completed four blocks of 25 trials in each condition, yielding 100 trials/condition in total. Data from one subject, who could not complete 100 trials in all conditions, were not included in the final analysis, leaving eight subjects. After subjects completed two blocks (50 trials) for each condition, they performed one block (25 trials) with targets only (no background) before they carried on with the regular test conditions. Visual feedback as to response correctness was provided after each trial.

2.4 Apparatus

Stimuli were generated digitally and played out via a LynxStudio L22 soundcard (32-kHz sampling, 24-bit resolution), and presented diotically (Sennheiser HD 580 headphones).

2.5 Statistical analysis

Individual correct-response counts were transformed into d'. The resulting d' values were analyzed using planned comparisons, including paired t-tests (for comparisons across conditions) and unpaired t-tests (to determine if sensitivity was significantly higher than zero, i.e., chance).

3. Results and discussion

Figure 1(B) shows the mean d' (across subjects) for the different test conditions. The results can be summarized as follows. First, for the H-Sync condition, d' was significantly higher, on average, for the MBD than for the MBS presentation mode [$t(7) = 3.23$, $p = 0.014$].

This suggests that harmonicity-based grouping was not sufficiently powerful to overcome segregation due to temporal incoherence. A similar effect was observed for the L-Sync condition, which is also consistent with the TCT (Shamma *et al.*, 2010), and in line with earlier findings using inharmonic MBD and MBS stimuli (Kidd *et al.*, 1994; Kidd *et al.*, 2003). For the frequency-shifted (S-Sync)—and thus, inharmonic—masker condition, a trend in the same direction was visible, but the difference was not statistically significant [$t(7) = 1.97$, $p = 0.090$]; therefore, the results of this condition are inconclusive.

A surprising feature of these results is that d' was statistically higher than zero in the H-Sync and L-Sync conditions, even for the MBS presentation mode—a situation in which subjects had no reliable external cue to distinguish the target tones from the masker tones [$t(7) = 4.08$, $p = 0.005$; $t(7) = 3.37$, $p = 0.012$]. A possible explanation for this is that the target usually fell into a frequency (or harmonic-rank) region which the listeners tended to orient their attention toward or, was weighted more heavily by the listeners in determining the overall percept. Consistent with the latter explanation, for the H stimuli, the target always corresponded to harmonics 3-6, and harmonics 3-5 usually have a stronger influence on the pitch of complex tones than components outside of this region (Ritsma, 1967).

Secondly, and importantly, d' was markedly higher for the M-Sync MBS condition than for the H-Sync MBS condition [$t(7) = 6.88$, $p < 0.0005$]. In fact, it was not significantly lower than for the M-Sync MBD condition [$t(7) = 1.31$, $p = 0.232$]. This shows that mistuning the target was enough to facilitate its detection despite temporal coherence—since in the M-Sync MBS conditions the targets were temporally coherent with the maskers—and that the facilitating effect of mistuning on discrimination performance can be as large as that of temporal incoherence. One interpretation of this outcome is that mistuning the targets by 4% was enough to induce their perceptual "pop out" (Moore *et al.*, 1985; Oh and Lutfi, 2000), and was sufficient to allow their segregation from the maskers; once the targets could successfully be "heard out," it no longer mattered whether or not they were temporally coherent with the background.

By contrast, for the log-spaced maskers, a large difference between MBS and MBD was still observed even with the frequency-shifted targets [$t(7) = 7.86$, $p < 0.0005$]. This is not surprising if one considers that the log-spaced stimuli were already inharmonic prior to the shift; thus, shifting the target by 4% did not introduce an inharmonicity cue in this condition, and the listeners could therefore benefit greatly from temporal incoherence. Nonetheless, a small but statistically significant benefit of shifting the target was observed for these log-spaced maskers (compare L-Sync and LM-Sync) for both MBS [$t(7) = 3.51$, $p = 0.010$] and MBD [$t(7) = 2.75$, $p = 0.029$], suggesting that the 4% target shift did introduce a slight but useable cue, other than inharmonicity. It is unclear what this cue is, but based on previous findings (see, e.g., Roberts and Brunstrom, 1998, reviewed in Micheyl and Oxenham, 2010), one may speculate that the human auditory system is somewhat sensitive to violations of regular log-spacing, and that this could explain why performance was slightly higher for LM-Sync conditions than for L-Sync conditions.

Lastly, and consistent with previous findings (Kidd *et al.*, 1994; Kidd *et al.*, 2003; Turgeon *et al.*, 2005), the introduction of (40-ms) onset asynchronies between the target and masker tones had a beneficial effect on sensitivity for all conditions [$2.67 < t(7) < 9.87$, $p \leq 0.032$], including conditions with a mistuned target and harmonic maskers [M-Sync vs M-Async: $t(7) = 2.67$, $p = 0.032$]. This is seen by comparing the rightmost group of bars with the bars for corresponding synchronous conditions, on the left. This indicates a general benefit of temporal incoherence resulting from the introduction of onset asynchronies between the targets and the maskers, over and above any benefit of inharmonicity or of violations from regular frequency spacing on a log scale. For these asynchronous target-masker conditions, no significant difference was found between the MBS and MBD presentation modes. This suggests that the 40-ms onset asynchronies were already sufficient to maximize segregation based on temporal incoherence, so that the introduction of further temporal incoherence (via

randomization of the maskers) produced no additional benefit. Although the benefits of onset asynchronies between target and masker can be understood in terms of a temporal-incoherence-detection mechanism (Shamma *et al.*, 2010), alternatively, or in addition, subjects may have used the shorter target duration as a cue to distinguish the targets from the maskers. Moreover, it is possible that reducing the target duration reduced the strength of harmonicity-based grouping somewhat (Moore, 1987); however, this cannot entirely explain the benefit of introducing an asynchrony between targets and maskers because such a benefit was observed even for the log-spaced conditions, in which the targets and maskers were not harmonically related to each other.

4. Conclusions

The results of this study provide further evidence, using performance in a task involving judgments of the direction of frequency shifts, that temporal coherence is an important factor in the perceptual organization of sounds into streams. However, they also indicate that deviations from harmonicity can substantially facilitate the perceptual segregation of a repeating sequence of target tones embedded among masker tones, *even when the targets and the maskers are temporally coherent with each other.* This outcome qualifies a strong interpretation of the TCT, according to which temporal coherence leads to obligatory across-frequency grouping, and temporal incoherence is both necessary and sufficient for sound segregation. Instead, the results suggest that temporal (in)coherence is one of several important factors—including spectral separation, presentation rate, harmonicity, and spectral separation—that, together, contribute to the perceptual organization of sound sequences into streams.

Acknowledgments

This work was supported by NIH Grant No. R01 DC 07657. We thank two anonymous reviewers for helpful comments on an earlier version of the manuscript.

References and links

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organisation of Sound* (MIT Press, Cambridge, MA).

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (**2009**). "Temporal coherence in the perceptual organization and cortical representation of auditory scenes," Neuron **61**, 317–329.

Elhilali, M., and Shamma, S. A. (**2008**). "A cocktail party with a cortical twist: How cortical mechanisms contribute to sound segregation," J. Acoust. Soc. Am. **124**, 3751–3771.

Hartmann, W. M., and Johnson, D. (**1991**). "Stream segregation and peripheral channeling," Mus. Percept. **9**, 155–184.

Kidd, G., Jr., Mason, C. R., Deliwala, P. S., Woods, W. S., and Colburn, H. S. (**1994**). "Reducing informational masking by sound segregation," J. Acoust. Soc. Am. **95**, 3475–3480.

Kidd, G. J., Mason, C. R., and Richards, V. M. (**2003**). "Multiple bursts, multiple looks, and stream coherence in the release from informational masking," J. Acoust. Soc. Am. **114**, 2835–2845.

Micheyl, C., Divis, K., Wrobleski, D. M., and Oxenham, A. J. (**2010**). "Does fundamental-frequency discrimination measure virtual pitch discrimination?," J. Acoust. Soc. Am. **128**, 1930–1942.

Micheyl, C., and Oxenham, A. J. (**2010**). "Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings," Hear. Res. **266**, 36–51.

Micheyl, C., Shamma, S., and Oxenham, A. J. (**2007**). "Hearing out repeating elements in randomly varying multitone sequences: a case of streaming?," in *Hearing—From Basic Research to Applications.*, edited by B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey (Springer, Berlin), pp. 267–274.

Moore, B. C. J. (**1987**). "The perception of inharmonic complex tones," in *Auditory Processing of Complex Sounds*, edited by W. A. Yost and C. S. Watson (Erlbaum, Hillsdale, NJ), pp. 180–189.

Moore, B. C. J., Peters, R. W., and Glasberg, B. R. (**1985**). "Thresholds for the detection of inharmonicity in complex tones," J. Acoust. Soc. Am. **77**, 1861–1867.

Oh, E. L., and Lutfi, R. A. (**2000**). "Effect of masker harmonicity on informational masking," J. Acoust. Soc. Am. **108**, 706–709.

Ritsma, R. J. (**1967**). "Frequencies dominant in the perception of the pitch of complex sounds," J. Acoust. Soc. Am. **42**, 191–198.
Roberts, B., and Brunstrom, J. M. (**1998**). "Perceptual segregation and pitch shifts of mistuned components in harmonic complexes and in regular inharmonic complexes," J. Acoust. Soc. Am. **104**, 2326–2338.
Shamma, S. A., Elhilali, M., and Micheyl, C. (**2010**). "Temporal coherence and attention in auditory scene analysis," Trends Neurosci. **34**, 114–123.
Shamma, S. A., and Micheyl, C. (**2010**). "Behind the scenes of auditory perception," Curr. Opin. Neurobiol. **20**, 361–366.
Turgeon, M., Bregman, A. S., and Roberts, B. (**2005**). "Rhythmic masking release: Effects of asynchrony, temporal overlap, harmonic relations, and source separation on cross-spectral grouping," J. Exp. Psychol. Hum. Percept. Perform. **31**, 939–953.