EXPLORING THE EMERGENCE OF BEAT INDUCTION USING A 'SWARM OF ONSETS' GENERATIVE MODEL

Nolan LEM [1] and **Jens AHRENS**[1]

[1]*Division of Applied Acoustics*, **Chalmers Institute of Technology**, Gothenburg, Sweden

ABSTRACT

This study investigates the emergence of the beat percept in auditory textures that are built up from homogeneous sources. Sensorimotor synchronization (SMS) models of beat perception have examined how humans synchronize to musically relevant signals, however there have been few systematic inquiries into SMS in more soundscape-oriented phenomena where a listener is confronted by multiples of concurrent, interactive sources. In this experiment, participants were asked to synchronize (via tapping) to stimuli created from a generative model where forty metronome sounds are clustered around periodic, temporal centers using a Gaussian probability distribution function parameterized with eight levels of variance to produce 'swarms' of temporally aligned onsets. Participants encountered stimuli both in mono (dichotic) and binaural spatial presentations as we hypothesized that the spatial presentation might produce significant differences in SMS performance. As evidenced by inter-beat interval analysis and phase coherence analysis, participants were able to largely synchronize to six of the eight onset variance conditions despite tap frequency increasing with onset spread. Similarly, while we found a slight interaction between spatial format and stimuli onset variance, no significant differences were found between the overall tap responses to the binaural versus mono stimuli. This study is a first look into how the beat percept arises to sound textures whose constituent elements contain different levels of embedded synchrony.

1. INTRODUCTION

To be immersed in a sounding environment is to encounter and listen to many different collections of sounds that make up the auditory scene. By segregating and identifying the multitude of sources contained within the soundscape, we make sense of the auditory environment around us. For example, we can detect and isolate sound objects, appreciate complex musical phenomena, and identify prospective threats, all while forming a coherent auditory map of our surroundings that allow us to navigate a physical space [1–3].

While much research in sensorimotor synchronization (SMS) has investigated the human ability to entrain, synchronize, and coordinate behavior with a variety of musically-oriented signals, there has not been a systematic inquiry into how SMS emerges from more ambient, naturalistic sounds. Naturalistic auditory textures, such as rain, hail, crackling fire, stridulating insects, or the rustle of wind, are often made up of sounding elements that interplay with one other in ways both complex and predictable [4]. Here, variegated interactions at the local level can give rise to collective behaviors that result from different cyclic patterns in nature where biological materials respond to environmental demands. In short, different behavioral dynamics create sound in adaptive ways and this has been reflected in a number of studies such as insect communication, crowd dynamics, crowd applause, and traffic noise [5–8].

Taken as a class of auditory signals, many naturalistic sound textures exhibit signal stationarity which is associated with specialized forms of auditory processing [9]; McDermott et al. successfully recreated such sound textures by generating signals that mimic their statistical properties via cochlea filtering during natural listening experiences. Since sound textures are comprised of large numbers of concurrent sources, this line of research suggests perceptual systems may rely on forms of temporal averaging in order to recognize and detect such sounds. Other generative models for soundscape recreation take a different approach by synthesizing individual sound sources en masse, exploring techniques such as granular and concatenative synthesis, physical modeling, and other procedural algorithms that are more physically analogous to sound production in the real world (For a review, please see [4]). If individual sound sources interact and synchronize their actions, how does collective behavior shape the resulting soundscape? Human synchronization can offer insights to better understand how we might perceive and engage with these auditory worlds.

We can listen for auditory cues in the form of repeated sonic events in order to sense regularity and construct a notion of pulse, a sensorimotor skill that is critical for the performance of music and group synchronization in general [10, 11]. The well-known 'cocktail party effect' provides evidence that we are able to 'listen in' and latch onto constituent elements in a dense, auditory textures comprised of multiple homogeneous sources [12, 13]. Even within an auditory stream of spoken language, we have to parse linguistic signals through various mechanisms such as synchronizing to phoneme structure [14, 15]. In one study, short repetitions of synthetic soundscapes were re-

peated at different intervals to determine if listeners were able to detect this embedded regularity [16]. The authors found that recurring segments of naturalistic sounds facilitated the detection of sound examples within multiple mixtures and that listeners may unconsciously learn priors about a sound texture via repetition. Furthermore, there is limited research on how source spatialization affects the beat percept and SMS in general. While large-scale, immersive speaker arrays may provide more dynamic and enveloping listener experiences at the expense of degrading the perception of rhythmic information [17], it is not clear what effect listener envelopment has on the emergence of the beat percept from otherwise stationary signals.

This perceptual study examines how the beat percept arises in sequences of concurrent sound onsets that are synchronized at different levels of temporal alignment. How synchronized do sound events have to be in order for humans to notice a sense of regularity, latch onto an entraining rhythm, and coordinate behavior in tandem with the incipient beat? By creating auditory textures built up from superpositions of sound onsets at different levels of synchrony, we can examine the point at which we begin to induce beat via tap synchronization. As such, this study is a first look into how the beat percept might arise within auditory textures comprised of homogeneous sound sources that are temporally allocated as (Gaussian) random variables over periodic moments in time. We hypothesized that people would use the embedded onset densities or 'swarms' as an auditory cue in order to synchronize to such auditory textures. Similarly, we anticipated that the spatial format of such sounds would influence human synchronization performance.

2. METHODS

2.1 Stimuli Preparation

Audio stimuli were generated from a generative Gaussian probability density function (PDF) that distributes 40 independent onsets in the form of metronome clicks around 15 "beat centers" as derived from 5 tempo conditions (60,68,77,88,100 bpm). For each tempo, the PDF was centered at the time associated with the period of each tempo condition, parameterized with 8 levels, or "gradations", of variance that correspond to a set of standard deviations spanning a linear range from 0.0 to 0.6 of the temporal period. Depending on the variance, this generative PDF allocates a 'swarm of onsets' more or less tightly around periodic centers corresponding to the tempo of each condition (each of the gradation conditions of variance scaled with tempo so as to keep the spread of distributions proportionate across the tempo condition). An audio sample of a single metronome click was used as the auditory source onset. Stimulus generation is illustrated in the Figure 1 which shows the Gaussian PDF parameterized at the 8 levels of variance for a single tempo condition (the lowest variance condition represents the isochronous case). Kernel Density Estimates (shown in orange) of the onset histograms are shown over each of the 15 beat centers for each variance condition. Correspondingly, the resultant audio waveform

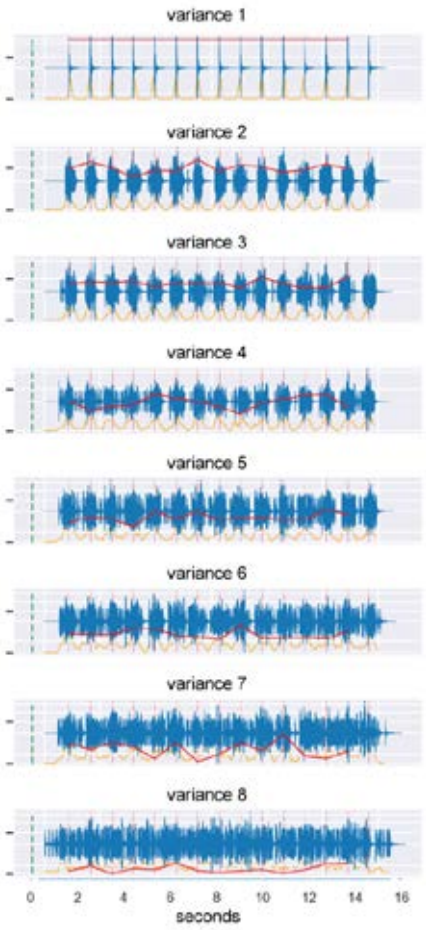

Figure 1. Examples of the stimulus waveforms for each of the variance conditions, or "gradations", for a single tempo condition. Kernel density estimates (shown in orange) approximate the shape of the generative Gaussian PDF centered on each of the 15 beat centers (shown as vertical red lines). The stimulus phase coherence, R_{onsets}, sampled at each temporal beat center, are shown in red.

is overlayed onto this plot; the red line shows the phase coherence magnitude (see Equation 1) , R_{onsets}, of the stimuli which is a summary statistic which indicates the global synchrony of the onsets over each beat region.

When the stimuli variance is low, the metronomes sounds are tightly distributed around the period associated with the tempo; conversely, when variance is high, the samples are more randomly distributed throughout the beat-to-beat temporal region. During stimulus creation, two versions of each variance gradation and tempo condition were created resulting in 80 stimulus sequences.

2.2 Test Procedure

14 volunteers from the University community participated in this study. Participants completed a questionnaire before the experiment began which collected demographic data. 4 women participated in the study and the mean age was 30.5 yrs (SD(age) = 11.2 yrs). 2 of the participants were left-handed and the entire cohort had 4.8 yrs as the average years of musical training (SD(musical training) = 6.2 yrs).

Participants were asked to tap on the spacebar of a 15" MacBook Pro Laptop using the dominant finger of their dominant hand with Sony MDR7506/1 headphones. The experimental stimuli were created using the SuperCollider music programming language. The stimuli were presented in randomly in two blocks, each block either consisting of sound sequences that were either dichotic (i.e., mono, the same sound presented to both ears) or binaurally using the SoundScape Renderer v. 0.6.1 with a head tracking unit (Supperware Head Tracker 1) [18]. The 40 onsets were distributed onto 40 virtual source positions located on the horizontal plane as a circle around the listener.

First, we collected a 15 second period of participant spontaneous tempo by asking them to tap at a rate that was comfortable to them without any auditory stimuli. Following this, participants attempted a practice block consisting of 4 auditory stimulus sequences before each experimental block. They proceeded onto the experimental tapping task which consisted of 80 audio files in each block which resulted in a total duration of around 20 minutes per block. The total experiment took approximately 40 minutes to complete.

2.3 Data Analysis

The parameters associated with each stimulus—namely the variance and the tempo — were initially collected and saved to disk during stimulus generation. The participant tap responses were analyzed in terms of their inter-tap-interval and the phase coherence magnitude, (R), and angle, (ψ). The inter-tap interval (ITI) is defined as the amount of time in between each of their successive taps which can be normalized with respect to the stimuli period (obtained from the tempo condition) so as to compare tap responses between tempo conditions; this is referred to as the normalized ITI (nITI). A nITI of 1 would mean that the participant tapped at the same rate as the stimulus beat centers (conversely, a nITI < 1 would indicate taps that occurred at a rate that was faster than the stimulus beat centers).

The phase coherence analysis is determined from calculating the stimulus onsets and participant tap response complex order parameters from which the phase coherence magnitude, R, and average angle, ψ, are obtained from Equation 1 where N is the number of onsets and ϕ_i is the phase in radians between consecutive beat centers (j is a complex number). R is a value between 0 and 1 which represents the synchrony of the group of onsets and ψ (0-2π) represents the average angle among all onsets.

$$R\, e^{j\psi} = \frac{1}{N}\sum_{i=1}^{N} e^{j\phi_i} \tag{1}$$

Unlike ITI analysis, phase coherence analysis provides information on how taps line up with the center of the beat and provides information about how taps may tend to precede ('lead') or fall behind ('lag') the center of the beat. Each participants tap and stimulus onset can be mapped to a point on a circle where the bounds of the circle represent the distance between consecutive beat centers where center of the beat is defined to be at π (onsets or taps $> \pi$ imply a lag and tap responses $< \pi$ imply a lead). For the tap and stimuli onset distribution, we can create a phasor with magnitude, R, at average angle, ψ, that reflects both the aggregate alignment of the participants' tap responses and the stimuli onsets. Rayleigh tests were performed to determine if the tap and onset distributions contained any significant directionality and lend support to the idea that the phase coherence magnitude and angle can be considered to be valid [19]. Similarly, comparisons between distributions can be made using Watson-Wheeler Tests, a circular statistic that tests for homogeneity on two or more samples of circular data in terms of the mean or variance [20].

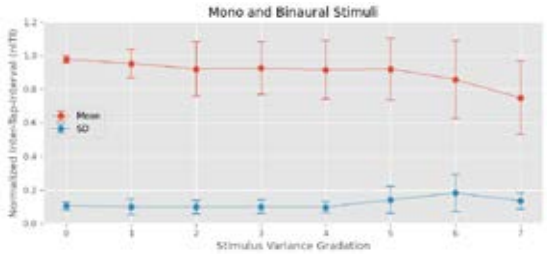

Figure 2. Participant mean (red) and SD (blue) of the normalized inter-tap-interval (nITI) over each stimulus variance condition. Error bars indicate the mean subject error.

3. RESULTS

Figure 2 shows the participant mean nITIs and SD of nITIs for each of the stimuli variance gradations (1-8) for the combined mono and binaural sounds. Here we see how nearly all of the participants were able to tap with a nITI$\approx$1 up until the 7th and 8th variance gradation at which point there is decrease in nITI which implies that the participants began to tap at a faster rate than the stimulus beat centers.

Figure 3 shows the mean nITI and SD distribution of all tapping trials as a function of the stimuli phase coherence. As the phase coherence of the stimuli decreases (at $R \approx 0.3$), many participants began to tap at a faster rate (lower nITI) with much more tap-to-tap variance (larger SD of nITI) across participants.

Table 1 summarizes the results of the three-way repeated measures ANOVA of the stimulus variance, spatial presentation, and tempo conditions for the mean nITI and SD respectively. For the mean nITI, the ANOVA revealed significant main effects of stimuli variance, tempo, and an interaction between stimuli variance and tempo. The main effect of stimuli variance was significant (F(7,84) = 5.84, $p < 0.001$) where post-hoc tests showed mean nITI progressively decreased from the stimulus variance gradation 1 to 8 (0.98, 0.95, 0.92, 0.93, 0.92, 0.92, 0.86, 0.75) and post-hoc analysis showed more significant mean nITI values as compared to higher stimulus variance gradations.

For the SD of the nITIs, the ANOVA revealed significant main effects only for the stimulus variance (F(7,84) = 3.83, $p = 0.001$) from gradation 1 to 8 (0.11, 0.10, 0.10, 0.10, 0.10, 0.14, 0.18, 0.14) where the mean SD of the subject

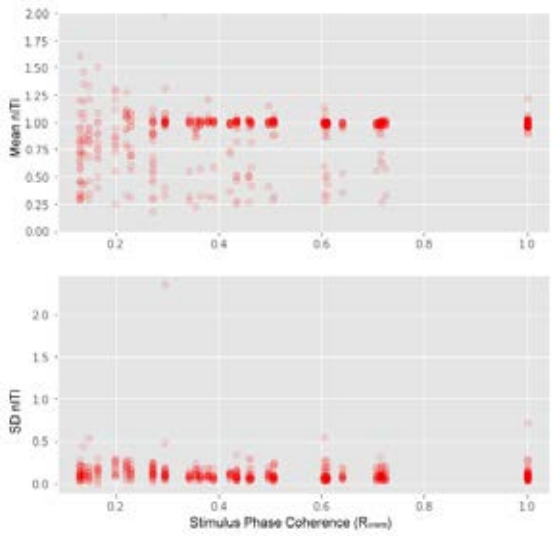

Figure 3. All mean (top) and SD (bottom) of the participant nITIs for each trial plotted as a function of the stimulus phase coherence. Mean nITIs < 1.0 imply that a participant tapped at a rate that was faster than the rate associated with the stimulus beat centers.

	Mean nITI	SD of nITI
Factor		
Variance Gradation	1-(3*,4*,5**,7***,8***) 2(-7*,8***) (3,4,5,6,7)-8***	2-8*** 2-(7*,8**) 3-(7*,8**) 4-(7*,8**) 5-(7*,8***)
Spatial Presentation	NS	NS
Tempo	60-(68*,77***,88***,100***) 68-100**	NS
Tempo x Gradation	Var 6: 60-(88*,100*) Var 7: 60-(77*,100**) Var 8: 60-(77*, 88*, 100*) 68-100**	NS

Table 1. Results of the three-way ANOVAs for the mean and SD of the participant nITIs. Note. Significance levels are indicated as *** $p < .001$, ** $p < .01$, * $p < .05$, NS = not significant. The levels for the variance conditions are (1-8) and the five levels for the tempo conditions were (60,68,70,88,100).

taps tended to increase over stimulus gradation.

Lastly, the main effects of the tempo on the nITI was significant ($F(1,12) = 27.4$, $p < 0.001$) and the results of the post-hoc tests indicated that tap nITI tended to increase with increasing tempo conditions; post-hoc results showed bpm 60, 68, 77, 88, and 100 corresponding to 0.83, 0.89, 0.92, 0.93, 0.96 nITI respectively.

The phase coherence of the taps and the stimulus onsets were computed to derive a phasor representing the average angle across participants with respect to the center of the beat. These phase coherence distribution plots are shown for each variance gradation in Figure 4. Rayleigh tests reported all of the tap and stimulus onset distributions were shown to have significant directionality with the exception of the tap distribution of variance gradation 8. Watson-Wheeler tests showed significant mean differences between the distributions of the tap and onset distributions for every gradation (all comparisons per variance gradation

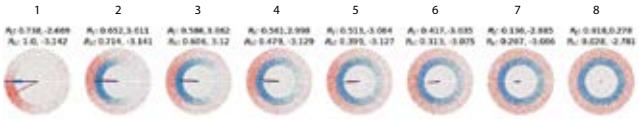

Figure 4. Phase coherence "swarm" plots showing the distribution of the participant taps (red) versus the stimulus onsets (blue). Phase coherence magnitude and average angle (R, ψ (rad)) are shown as phasors for the taps (R_t) and onsets (R_o).

$p < 0.0001$, except variance gradation 8, ($p = 0.003$)).

4. DISCUSSION

We observe three central tendencies among participant tap responses: 1.) all participants tapped in line with the region of the stimuli containing the highest onset density (the beat center) across variance gradation and 2.) participants slightly increased the mean frequency of their taps when synchronizing to stimuli with increased onset variance but tapping was slower at higher tempos and 3.) the spatial presentation (mono, or binaurally rendered) of the onset swarms did not significantly affect their tap responses. These results indicate that synchronization performance only degraded at gradation 7 and 8 suggesting that up until that variance gradation, participants were able to induce a beat percept from the stimulus sequences. While SMS probably depends on the source sound material and numerosity of sources and events, this study suggests that sound onsets normally distributed in time can induce the beat percept if they are sampled from a normal distribution with a SD that is ≈ 0.3 of the temporal period.

Previous research has observed tapping rate increases with stimulus asynchrony which might be a result of individual differences [21]. Surprisingly, participants taps were generally consistent throughout the course of the stimuli sequences since tap-to-tap variance (SD of nITI) was generally low and only significantly increased toward the last two stimuli variance gradation. While it is possible that subjects simply tapped at consistent rate without being entrained, their tapping rates during stimulus presentation were substantially faster than their collected spontaneous rates. For the last two variance gradations, approximately 40% of their taps still fell within a 10% window and 20% window of the target stimulus tempo for the last two stimulus variance gradations respectively. Why might people tap at faster rate when they lose the beat percept? Once individual onsets are no longer grouped as belonging to a particular beat center, individual onsets—taken sequentially rather than streamed — may begin to dominate the perception of the beat [22, 23]. Similarly, faster tapping might reflect strategies that incorporate active sensing where humans might sample from sensory spaces at a faster rate [24, 25].

No significant differences were found in tap responses to the mono (dichotic) presented sounds versus the binaurally presented ones. Initially, we hypothesized that the listener being enveloped by the sounds might allow them to segre-

gate specific sounds in the mix and use that sound as an auditory cue for synchronization. However, our findings support previous research showing how only four spatialized speakers can approximate spatial envelopment for broadband, stationary sounds such as those found in naturalistic soundscapes [26, 27].The findings of this study support the notion that SMS performance is not facilitated nor constrained by the presence of more spatialized sources.

The phase coherence angle of the group's tapping response to gradation 1, a completely isochronous stimuli, shows a lag (≈0.47 rad, ≈75 ms for the 60 bpm condition, ≈45 ms for the 100 bpm condition) between the mean of the subject tap and the center of the beat is well within the range of negative mean asynchrony (NMA) observed in other SMS studies which for humans to coordinate their taps with a slight delay behind an entraining stimulus onset [28]. ψ_{taps} slightly preceded ψ_{onsets} during variance gradation 2, 3, and 4 suggesting that listeners may have been more adaptive in their tapping; error correction models have accounted for such anticipatory responses in auditory pacing sequences with expressively modulating tempos [29]. However, as the stimulus variance increases, we observed a reduction of phase coherence magnitude across and within participant tap responses. Surprisingly, while the R_{taps} value remained small (≈ 0.14), ψ_{taps} still remains pointing towards the center of the beat until the last variance gradation which suggests that overall, there was still a tendency to tap toward either the center of the beat or in between two beat centers. Another indication of the participant entrainment is observed in the transition from variance 4 to 6, where $R_{taps} > R_{onsets}$: the participants placed their taps with more temporal alignment than the onsets themselves.

Ultimately, our findings support the notion that onset density is indeed used by the listener in order to construct a sense of beat, however tentative that may be.

It should be noted that the waveform of the stimuli associated with larger variances did contain a weakly periodic amplitude envelope suggesting that volume might be the entraining auditory cue as shown in other research in amplitude modulated noise perception [30] but that the majority of participants were still able to coordinate their taps with onset density when phase coherence magnitude, R, was small (≈0.3). Even among a backdrop of a homogeneous auditory texture, we suspect that participants are able to track the accumulation of onsets as they coalesce around temporal centers and that the auditory mechanism involved in auditory averaging of stationary signals transitions to parsing onsets into periodic temporal regions, a process that shares many features with beat-bin based models of microtiming in groove-based music [31]. In fact, this model is built into to the stimuli generation a priori: PDFs generate onsets that align with more or less spread around temporal centers, and beat perception in this context may rely on integrating onsets within durational boundaries that arise from making causal inferences from priors in order to structure beat relevant information in the real-time [32].

The SMS behaviors observed in this study might not apply to phenomena exhibiting other statistical distributions given that our stimuli produced onset swarms normally distributed in time; natural phenomena exhibiting other statistical distributions might induce other forms of synchronization. This analysis did not capture any aspects related to the time structure of participant taps—that is, there was no analysis to determine if their taps evolved over the course of the stimuli sequences. We observed that tap variance was relatively low for most of the variance gradations but further inquiry into the nature of when participants began to tap or how their taps evolved on a tap-to-tap basis would be interesting to evaluate more thoroughly. While we did capture the participants' head tracking movement data during the experiment, we noticed that they did not tend to move their head during the tapping task and therefore did not analysis this data. Because the 40 virtual sources sufficiently enveloped the listener, head movements would not likely have facilitated beat induction.

5. CONCLUSIONS

This study provided a first look into how a sense of beat might arise from interactions among sound sources that makeup a larger auditory texture. We used a probability density function parameterized at different levels of variance to span a range of synchronous onset sequences. While spatial source location did not significantly affect tap performance, ITI and phase coherence analysis suggested that participants synchronized their behavior with the recurring onset density and that the gradual loss of the beat percept was associated with tapping at an accelerated rate. Further inquiry into SMS using onset distributions might examine the effects of source numerosity; for example, listeners may well use such spatialized cues when the number of perceived sound sources in a space is lower. This would be particularly true if the sources themselves were less homogeneous in timbre, as more salient sound objects might direct attention in such a way as to significantly alter their SMS performance [33, 34].

Acknowledgments

This research was supported by the Anna Ahrenberg Foundation.

6. REFERENCES

[1] A. J. King, "Sensory experience and the formation of a computational map of auditory space in the brain," *BioEssays*, vol. 21, no. 11, pp. 900–911, Oct. 1999. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1521-1878(199911)21:11⟨900::AID-BIES2⟩3.0.CO;2-6

[2] K. T. Gagnon, M. N. Geuss, and J. K. Stefanucci, "Fear influences perceived reaching to targets in audition, but not vision," *Evolution and Human Behavior*, vol. 34, no. 1, pp. 49–54, Jan. 2013. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1090513812000918

[3] C. Douglas, J. Noble, and S. McAdams, "Auditory Scene Analysis and the Perception of Sound Mass in Ligeti's Continuum," *Music Perception: An Interdisciplinary Journal*, vol. 33, no. 3, pp. 287–305, 2016.

[4] D. Schwarz, "State of the art in sound texture synthesis," *Proceedings of the International Conference on Digital Audio Effects, DAFx*, no. May, pp. 221–232, 2011, iSBN: 9782954035109.

[5] M. Greenfield, "Cooperation and conflict in the evolution of signal interactions," *Annual Review of Ecological Systems*, vol. 25, pp. 97–126, 1994.

[6] M. Moussaïd, D. Helbing, and G. Theraulaz, "How simple rules determine pedestrian behavior and crowd disasters," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 17, pp. 6884–6888, 2011.

[7] Q. Meng and J. Kang, "The influence of crowd density on the sound environment of commercial pedestrian streets," *Science of the Total Environment*, vol. 511, pp. 249–258, 2015, iSBN: 1759-4685 (Electronic)\r1759-4685 (Linking).

[8] Z. Neda, E. Ravasz, T. Vicsek, Y. Brechet, and A. Barabasi, "Physics of the rhythmic applause," *Theoretical and Mathematical Physics*, vol. 104, no. 3, pp. 1104–1107, 1995, iSBN: 1063-651X.

[9] J. H. McDermott, M. Schemitsch, and E. P. Simoncelli, "Summary statistics in auditory perception," *Nature Neuroscience*, vol. 16, no. 4, pp. 493–498, 2013, iSBN: 1546-1726 (Electronic)\r1097-6256 (Linking).

[10] E. W. Large and C. Palmer, "Perceiving temporal regularity in music," *Cognitive Science*, vol. 26, pp. 1–37, 2002.

[11] B. H. Merker, G. S. Madison, and P. Eckerdal, "On the role and origin of isochrony in human rhythmic entrainment," *Cortex*, vol. 45, no. 1, pp. 4–17, 2009, publisher: Elsevier Srl ISBN: 0010-9452. [Online]. Available: http://dx.doi.org/10.1016/j.cortex.2008.06.011

[12] S. R. Alain, Claude; Arnott, "Selectively attending to auditory objects," *Frontiers in Bioscience*, vol. 5, no. 3, p. A505, 2000, iSBN: 1093-4715 (Electronic) 1093-4715 (Linking). [Online]. Available: https://www.bioscience.org/2000/v5/d/alain/list.htm

[13] J. Xiang, J. Simon, and M. Elhilali, "Competing streams at the cocktail party: Exploring the mechanisms of attention and temporal integration," *Journal of Neuroscience*, vol. 30, no. 36, pp. 12 084–12 093, 2010, iSBN: 9550091023.

[14] J. Morton, S. Marcus, and C. Frankish, "Perceptual centers (P-centers)," *Psychological Review*, vol. 83, no. 5, pp. 405–408, 1976.

[15] R. C. Villing, B. H. Repp, T. E. Ward, and J. M. Timoney, "Measuring perceptual centers using the phase correction response," *Attention, Perception, and Psychophysics*, vol. 73, no. 5, pp. 1614–1629, 2011, iSBN: 1341401101101.

[16] J. H. McDermott, D. Wrobleski, and A. J. Oxenham, "Recovering sound sources from embedded repetition," *Proceedings of the National Academy of Sciences*, vol. 108, no. 3, pp. 1188–1193, 2011, iSBN: 0027-8424. [Online]. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.1004765108

[17] T. Mouterde, N. Epain, S. Moulin, and E. Corteel, "On the factors influencing groove fidelity in immersive live music events," *AES Convention*, 2023.

[18] J. Ahrens, M. Geier, and S. Spors, "The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *Audio Engineering Society Convention 124*, May 2008. [Online]. Available: https://www.aes.org/e-lib/browse.cfm?elib=14460

[19] B. R. Moore, "A Modification of the Rayleigh Test for Vector Data," *Biometrika*, vol. 67, no. 1, p. 175, Apr. 1980. [Online]. Available: https://www.jstor.org/stable/2335330?origin=crossref

[20] K. V. Mardia, "Statistics of Directional Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, 1975.

[21] N. Lem and T. Fujioka, "Individual differences of limitation to extract beat from Kuramoto coupled oscillators: Transition from beat-based tapping to frequent tapping with weaker coupling," *PLOS ONE*, vol. 18, no. 10, p. e0292059, Oct. 2023. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0292059

[22] A. Bendixen, S. L. Denham, K. Gyimesi, and I. Winkler, "Regular patterns stabilize auditory streams," *The Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3658–3666, Dec. 2010. [Online]. Available: https://pubs.aip.org/jasa/article/128/6/3658/904282/Regular-patterns-stabilize-auditory-streams

[23] R. Brochard, C. Drake, M.-C. Botte, and S. McAdams, "Perceptual organization of complex auditory sequences: Effect of number of simultaneous subsequences and frequency separation." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 25, no. 6, pp. 1742–1759, Dec. 1999. [Online]. Available: https://doi.apa.org/doi/10.1037/0096-1523.25.6.1742

[24] S. C.-H. Yang, D. M. Wolpert, and M. Lengyel, "Theoretical perspectives on active sensing," *Current Opinion in Behavioral Sciences*, vol. 11, pp. 100–108, Oct. 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352154616301255

[25] C. E. Schroeder, D. A. Wilson, T. Radman, H. Scharfman, and P. Lakatos, "Dynamics of Active Sensing and perceptual selection," *Current Opinion in Neurobiology*, vol. 20, no. 2, pp. 172–176, Apr. 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959438810000322

[26] O. Santala and V. Pulkki, "Directional perception of distributed sound sources," *The Journal of the Acoustical Society of America*, vol. 129, no. 3, pp. 1522–1530, 2011. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.3533727

[27] T. Pihlajamäki, O. Santala, and V. Pulkki, "Synthesis of Spatially Extended Virtual Sources with Time-Frequency Decomposition of Mono Signals," *J. Audio Eng. Soc.*, vol. 62, no. 7, 2014.

[28] G. Aschersleben and W. Prinz, "Synchronizing actions with events: The role of sensory information," *Perception & Psychophysics*, vol. 57, no. 3, pp. 305–317, Apr. 1995. [Online]. Available: http://link.springer.com/10.3758/BF03213056

[29] M. C. Van Der Steen, N. Jacoby, M. T. Fairhurst, and P. E. Keller, "Sensorimotor synchronization with tempo-changing auditory sequences: Modeling temporal adaptation and anticipation," *Brain Research*, vol. 1626, 2015.

[30] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1364–1380, Nov. 1979. [Online]. Available: https://pubs.aip.org/jasa/article/66/5/1364/778563/Temporal-modulation-transfer-functions-based-upon

[31] A. Danielsen, K. Nymoen, E. Anderson, G. S. Câmara, M. T. Langerød, M. R. Thompson, and J. London, "Where is the beat in that note? Effects of attack, duration, and frequency on the perceived timing of musical and quasi-musical sounds," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 45, no. 3, pp. 402–418, 2019.

[32] M. T. Elliott, A. M. Wing, and A. E. Welchman, "Moving in time: Bayesian causal inference explains movement coordination to auditory beats," *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1786, 2014.

[33] M. Elhilali, J. Xiang, S. A. Shamma, and J. Z. Simon, "Interaction between attention and bottom-up saliency mediates the representation of foreground and background in an auditory scene," *PLoS Biology*, vol. 7, no. 6, 2009.

[34] S. A. Shamma, M. Elhilali, and C. Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends in Neurosciences*, vol. 34, no. 3, pp. 114–123, 2011, publisher: Elsevier Ltd ISBN: 1878-108X (Electronic) 0166-2236 (Linking). [Online]. Available: http://dx.doi.org/10.1016/j.tins.2010.11.002