Proceedings of the 25th Anniversary Conference of the European Society for the Cognitive Sciences of Music, 31 July-4 August 2017, Ghent, Belgium
Van Dyck, E. (Editor)

Statistical Summary Representations in Music-Like Perception

Harun Yörük[1], Esra Mungan[2]

Department of Psychology, Boğaziçi University,Turkey
[1]harun.yoruk@boun.edu.tr, [2]esra.mungan@boun.edu.tr

ABSTRACT

In the visual domain people tend to use ensemble coding to represent sets of objects by averaging their object features. Extraction of these statistical summaries appears to be a very fast and accurate process. Recent evidence suggested that listeners can also use ensemble coding in perception of auditory sequences with pure tones. In this study, we investigated statistical summary representations using more music-like stimuli. We found that nonmusician listeners performed above-chance when estimating the mean pitch frequency of a complex tone sequence with 6, but not 4 or 8 tones. Our study presents some evidence for statistical summary extraction in nonmusicians with complex tone sequences of moderate length. We discuss our results with respect to why complex tones might have brought some limit to statistical averaging. For higher ecological validity it is critical that studies on ensemble encoding with auditory stimuli start using complex rather than pure tones. This will also prepare grounds for a better understanding of various music-feature - related mechanisms in music perception.

I. INTRODUCTION

Statistical summaries or ensemble representations are higher-level representations that seem to occur before individual representations of objects. Our perceptual system extracts ensemble representations of sets of objects in order to overcome capacity limitations that would have been caused when representing all parts of a set separately and in detail. It is also more efficient to process visual information by taking advantage of the redundancies and regularities in the visual world (Alvarez, 2011). We seem to make these statistical summaries very rapidly and accurately (Chong & Treisman, 2003). Ensemble representation mechanisms have mostly been studied in the domain of visual perception, and observed for various features of visual objects such as; size (Ariely, 2001), orientation (Dakin, 2001; Parkes, Lund, Angelucci, Solomon & Morgan, 2001), color (Maule, Witzel & Franklin, 2014), brightness (Bauer, 2009), facial emotion (Haberman & Whitney, 2007), and facial identity (de Fockert & Wolfenstein, 2009). Further studies extended these findings from static displays to sequentially presented sets of objects (Albrecht & Scholl, 2010; Albrecht, Scholl, & Chun, 2012). It has been shown that, in a sequence of visual objects, contribution of early and later items to the summary statistics differ according to task demands. When a task requires rapid responses, earlier items dominate the statistical summary calculation and a primacy effect is observed. On the other hand, when the task requires updating for more recent items or when representations of early items are lost due to limited storage capacity of attention and memory, later objects gain importance and a recency effect is observed (Hubert-Wallander & Boynton, 2015).

Even though the literature about statistical summary representations is dominated with findings from visual cognition, the mechanism is not limited to vision. Statistical summary representations are also observed with auditory features. Albrecht, Scholl, and Chun (2012) showed that when people were presented a sequence of {pure tone-visual display} pairs, they successfully averaged the pitch frequencies of the pure tones as well as the disc diameters of the varying circles. Moreover, the absolute error percentages when averaging pure tones were even lower than those for the co-presented visual objects. This suggested that (1) statistical summary representations were not specific to vision but also present in auditory perception, and (2) the averaging mechanism was even more accurate for auditory perception.

In another study, Piazza and colleagues (2013) presented participants series of six logarithmically spaced and sequentially presented pure tones. Each 6-tone sequence was followed by a test tone that was always different from the pitch mean. Participants had to rate if the pitch frequency of a subsequent test tone was higher or lower than the mean frequency of the set. Findings showed that listeners were mostly correct in their answers suggesting that they were indeed capable of extracting an average frequency from an auditory sequence. They varied the number of tones in the set and observed that at least three tones were sufficient for accurate averaging. They also found that except for the last and somewhat for the first tone, listeners could not reliably identify members of the 6-tone sequence in a two-alternative forced choice task. The same was true when their task was to identify the position of a member within the sequence. In other words, participants seemed to extract the ensemble representation of the mean frequency without explicitly retrieving the individual tones.

Given these first pieces of evidence for statistical summary representations in pure-tone sequences we aimed at extending these findings using (1) complex tones, (2) nonmusicians (Albrecht, Scholl, & Chun, 2012, do not report about their participants' musical background; participants in the Piazza et al. study, on the other hand, had moderate to higher levels of musical training), (3) a different task setup.

Our study used complex tones with synthesized piano timbre to obtain more music-like perception. Furthermore, we decided to use a two-alternative forced choice task as a slightly more direct way of testing statistical summaries. Piazza et al. (2013) used a test tone that was either higher or lower by 0.5 up to 5 semitones than the statistical average. Hence, participants in the Piazza et al. study never directly judged the statistical average per se, whereas in our setup they will (up against a distractor tone that is ± 2 semitones different from the target tone). Albrecht, Scholl, and Chun (2012), on the other hand, used a slider, as is commonly done in vision. But one potential issue with using a slider in a tone-context is that tones are typically part of a categorical rather than continuous representational system. Given that humans form tonotopic maps early on (cf. Trainor, 2005), finding lower error rates for auditory compared to visual stimuli when using a slider may be an artifact of differences in how pitch

frequency as opposed to disc diameter is represented. Even nonmusicians should, for instance, be less likely to stop the slider at a pitch frequency that does not exist in their representational repertoire (e.g. 445 Hz).

EXPERIMENT 1

In the first experiment we aimed to investigate if listeners could accurately estimate and encode the mean pitch frequency of a sequence of eight complex tones as a statistical summary representation. We also wanted to look at potential primacy or recency effects in the statistical summary extraction process. By adding a between-subjects primacy or recency distractor condition to the two-alternative forced choice test (2AFC), we aimed to see if mean representations might be (falsely) influenced by early or late items in the sequence. Given Albrecht et al's (2012) findings, we expected above chance performance for the group that only received the correct mean tone together with a distractor tone that was either 2 semitones above or below the correct average but never the average of the first or last four tones. For the groups who received the correct average tone together with either the first (primacy) or last (recency) 4-tone averages, we expected a potential primacy effect given the speed of the task (cf. Hubert-Wallander & Boynton, 2015). And in light of Piazza et al.'s (2013) finding that, despite good single-tone memory for the last tone, it was *ensemble* rather than single-note encoding that determined correct statistical summary, we did not expect a recency effect.

I. METHOD

A. Participants

Listeners were 45 Boğaziçi University students (16 participants in the "simple mean estimation" group, 14 participants in the "primacy distractor" group, and 15 participants in the "recency distractor" group). They all had less than 2 yrs of musical training (M = 1.1 yrs, SD = 1.6, for "simple mean estimation group", M = 1.7 yrs, SD = 4.4, for "primacy distractor group"; M = 0.2 yrs, SD = 0.5, for "recency distractor group"). They were compensated with course credits for their participation.

B. Stimuli

We prepared piano-like complex tones using Ableton Live 9 Suite software. Pitch frequencies of the tones were within the range of 82 to 1567 Hz (E2 to G6). Sequences consisted of eight isochronous tones of 500 ms duration each. They were generated by selecting a mean value per sequence and then composing a set with -7, -5, -3, -1, +1, +3, +5, and +7 semitones around the mean tone in random order. A 1500 ms interval was inserted after each tone sequence exposure. Two test tones appeared for 500 ms each, with a 500 ms ISI. For the "simple mean estimation group", tones in the distractor choice of the 2AFC test differed by ±2 semitones from the mean frequency and never corresponded to the mean of the first or last four notes. For the "primacy distractor group", the mean pitch frequency of the first four notes, for the "recency distractor group" the mean pitch frequency of the last four notes served as the distractor choice. None of the mean frequency (=target) tones used at test were present in the preceding complex tone sequence.

C. Procedure

The experimental procedure was prepared using PsychoPy software. There were 200 trials in total, which were divided into four blocks to allow listeners to have small breaks within the experiment. In each trial an eight-tone sequence was presented, followed by a 2AFC. In half of the trials the mean tone (=target), in half the distractor/primacy/recency tone was presented first. Whether the mean or distractor/primacy/recency tone was played first was randomized across each block of 50 trials. Participants were instructed to press "1" if they thought the first, and "2" if they thought the second test tone represented the mean frequency of the sequence.

Participants first received a demo trial with four examples of two-, four-, and eight-tone sequences. Each sequence was followed by their mean frequency tone to demonstrate what was meant by 'mean pitch frequency'. Participants then had a practice session of eight trials which were just like the experimental trials such that after each 8-tone sequence they were presented two choices from which to pick the correct answer. None of the sequences presented during the demo and practice sessions were used during the actual experimental sessions. Participants were told that they should respond fast without rehearsing the sequence in their mind. This instruction was given to minimize possible interference via top-down processes. The experimental trials began once participants reported to have understood the task. The 200-trial experimental session lasted about 40 min. At the end, participants were given a short post-experimental questionnaire about their musical background. They also received questions about whether and if so, which kinds of strategies they used during the task.

II. RESULTS AND DISCUSSION

We calculated the percentage of correctly choosing the mean frequency tone per sequence across trials. A one-sample t-test revealed that participants in the "simple mean estimation group" performed at 50% chance level (M = .50, p > .10), which suggested that they were not able to differentiate the mean from the ±2 semitone distractor (Figure 1). Participants of the "primacy distractor group", on the other hand, showed an above-chance choice for the mean tone over the primacy tone (M = .56, $t(13)$ = 2.52, p < .05). When the distractor was the mean of the recent four tones, participants once more showed chance performance (M = .48, p > .10). These findings could suggest that participants formed some ensemble representation based on the entire 8-tone sequence but with a resolution that falls within ±2 semitones of the mean given that the primacy and recency averages were always ±4 semitones apart from the target mean frequency. Interestingly this difference facilitated listeners' choice when the mean of the *first* four tones was presented as the distractor, meaning that the last four tones made them "shy away" from the primacy distractor. Yet, when the mean of the *last* four notes was presented as the distractor, they did confuse it to be the average despite the fact that it was as much as ±4 semitones off the actual mean tone. One could conjecture that

the final four tones had an attention-grabbing effect on the listeners which caused this confusion.

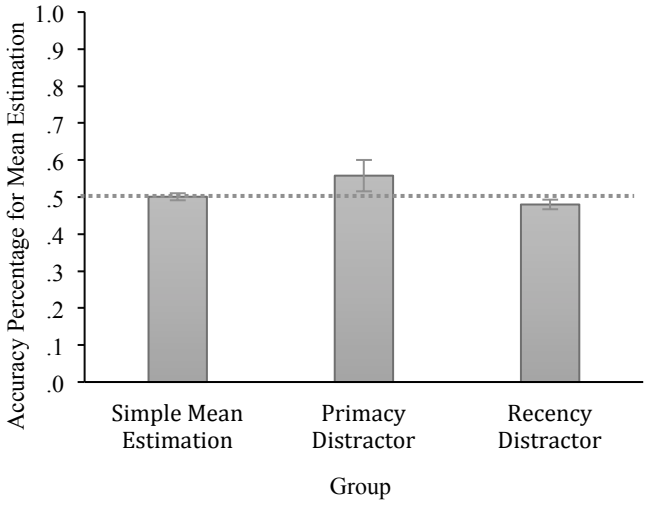

Figure 1. Accuracy percentage for choosing the mean tone target over the distractors.

EXPERIMENT 2

In two conditions of Experiment 1, choice for the mean tone was at chance level, but it was above chance when the correct mean was presented against a ±4 semitone primacy distractor. Yet, we cannot claim that the latter finding provides evidence for the existence of a statistical summary representation. Instead, participants probably picked the correct choice because the ±4 semitone primacy distractor appeared too "off" to be the mean.

In contrast to our findings, listeners in Albrecht, Scholl, & Chun (2012) study were able to accurately average the pitch frequency of an eight-tone sequence. It could be the case that when using complex tones, an eight-tone sequence was too complex for statistical summary. The confusion experienced with a recency distractor of ± 4 semitones off the real mean also hints to such overload in a sequential series. In the second experiment we reduced the sets to six and four complex tones to test whether using shorter complex tone sequences would yield evidence for statistical summary representations. We had to drop the primacy and recency conditions because in the six-tone condition of this experiment, the mean tone of the first and second half of the sequence was always present in the listened sequence, which would have created a confounding. Using a primacy and recency condition was also not feasible for the four-tone condition, since Piazza et al.'s study has shown that averaging required at least 3 tones.

I. METHOD

A. Participants

Listeners were 13 Boğaziçi University students (data from one participant deleted because of 15 years of musical training, which left 12 participants), who head 1.50 yrs (SD = 1.50) of musical training on average. They were compensated with course credits for their participation.

B. Stimuli and Procedure

Hundred four-tone and 100 six-tone sequences were prepared in a similar way as described in Experiment 1. Otherwise, everything was exactly the same except that there were no primacy/recency distractor conditions. All participants went through two blocks of 100 four-tone and then 100 six-tone sequences.

II. RESULTS AND DISCUSSION

Results of the second experiment showed that choice for the mean pitch frequency of the complex tone sequence over the distractor was significantly above chance level for the six-tone sequence (M = .54, $t(12)$ = 2.96, p < .05), but not for the four-tone sequence (M = .51, p >.10) (Figure 2). These results implied that listeners could estimate the mean pitch frequency of a complex tone sequence at above-chance level when it consisted of six but not four tones. Moreover, the precision of that statistical summary representation seemed to be good enough to differentiate it from a mean that deviated by only ±2 semitones.

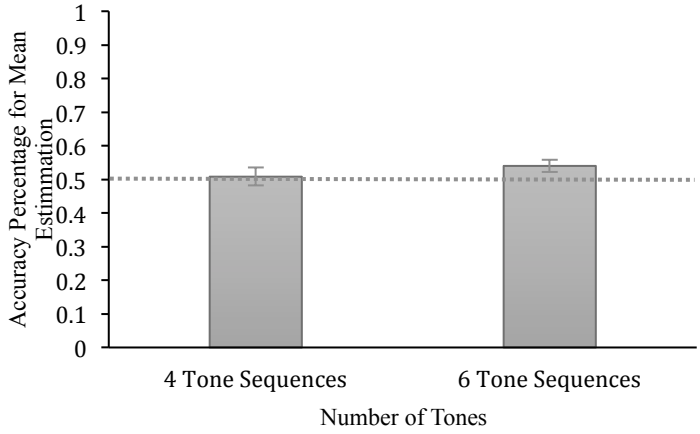

Figure 2. Accuracy percentage for choosing the mean tone target over the distractor tone for 4 and 6 tone sequence conditions.

III. CONCLUSION

To our knowledge, this is the first study to have used complex tones. Overall, results of our study showed that participants were able to estimate mean pitch frequency of a complex tone sequence for a six- but not four- or eight-tone sequence.

Our finding of above chance performance for six-tone sequences replicates Piazza and colleagues' (2013) results despite differences in setup, participant music background and stimulus type. Whereas listeners in Piazza and colleagues' (2013) study were given a non-target test tone and asked to estimate if it was higher or lower in frequency compared to an 'imagined' mean frequency tone of the preceding sequence, in our study, listeners were directly presented with the correct choice among two alternatives. This implies that listeners not only estimated the mean pitch frequency of an auditory sequence relative to another tone, but that they could also choose the exact average over a distractor tone that was only two semitones away from it.

Another critical point is that our task must have been an easier task than Piazza et al.'s task. In Piazza et al., we see the exact data of only one participant (all remaining data are presented in the form of slope values of each participant's psychometric curve), which shows that s/he was considerably better if the test tone deviated by more than two semitones. When the test tone deviated by ± 2 semitones, the participant seemed to be confused as to whether the tone was above or below the average. Given that we only used ±2 semitones distractors, having obtained above-chance results is worthy of attention. If we had used distractors that were ±3 or 4 semitones off the mean frequency, our above-chance rates would have probably been much higher.

The more perplexing finding is that we found above chance performance for six- but not four- or eight-tone sequences. In other words, we found an inverted U function type trend hinting to some capacity limitation with longer sequences[1] and some non-capacity related limitation with shorter sequences. We are not aware of any study in ensemble encoding that consistently manipulated length of temporally unfolding sequences. What Piazza et al. did in their Experiment 2 was to present parts of the six-tone sequences to see whether partial information was enough for listeners to correctly judge test tones against the mean of the entire 6-tone sequence. They found that participants produced best results with all six tones and incrementally worse results with less. Hence, they concluded that listeners must have used all tones to produce a statistical summary rather than used some heuristic approximation via a few tones within the set. Yet, this did not test listeners' ability to do statistical summaries of shorter sequences per se. Given that complex unlike pure tones are tones that have harmonics, it could be that four notes or less are not sufficient to understand its representative mean. It is also quite likely that different statistical processes may be at work for complex as opposed to simple pure tones. One might speculate that some extraction of geon-like sub-wholes (see Biederman, 1995; also cf. Bigand, Gérard, & Molin, 2009) across harmonics might be necessary for statistical representations in a sequence of complex tones which may not emerge in sequences that are too short. On the other hand, simple, single-dimension pure tones may allow for direct statistical averaging, hence better performance also with 8-tone sequences (Albrecht, Scholl, & Chun, 2012).

With this study, we provided some insight about statistical summary representations in music-like perception. Our findings provide some base for the possibility that nonmusicians could use ensemble coding while listening to almost musical stimuli. Our future goal is to look for evidence of statistical summary representations in more melody-like stimuli. For example, listeners could be asked to average the pitch height of a melody sequence that consists of transposed variations of a melodic motif, or the mean tempo of a melody sequence that consists of faster and slower versions of the same melodic motif. This would provide us with more information about representations are formed when faced with different musical surface features.

[1] The ensemble encoding literature in vision rejects a capacity-related explanation (cf. Alvarez,2011) but many working memory models suggest that auditory and visuospatial domains are different (e. g. Baddeley, 2003). Moreover, we deal with the specific case of statistcial representation of *temporally unfolding* stimuli.

ACKNOWLEDGMENT

We thank Ece Kaya and Alperen Karan for their help with constructing the experimental stimuli.

REFERENCES

Albrecht, A. R., Scholl, B. J., & Chun, M. M. (2012). Perceptual averaging by eye and ear: Computing summary statistics from multimodal stimuli. *Attention, Perception, & Psychophysics, 74*, 810-815.

Albrecht, A. R., & Scholl, B. J. (2010). Perceptually averaging in a continuous visual world extracting statistical summary representations over time. *Psychological Science, 21*, 560-567.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157-162.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences, 15*(3), 122-131.

Bauer, B. (2009). Does Stevens's power law for brightness extend to perceptual brightness averaging? *The Psychological Record, 59*(2), 171.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*(4), 393-404.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *JOSA A, 18(*5), 1016-1026.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology, 62*(9), 1716-1722.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology, 17*(17), R751-R753.

Hubert-Wallander, B., & Boynton, G. M. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of Vision, 15*(4):5, 1-12

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: Metric and categorical effects on ensemble perception of hue. *JOSA A, 31*(4), A93-A102.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739-744.

Piazza, E. A., Sweeny, T. D., Wessel, D., Silver, M. A., & Whitney, D. (2013). Humans use summary statistics to perceive auditory sequences. *Psychological Science, 24*(8), 1389-1397.

Trainor, L. J. (2005). Are there critical periods for musical development? *Developmental Psychobiology, 46*(3), 262-278.