Auditory Scene Analysis and Musical Experience

Robert G. Kellems

Indiana University

Abstract

Auditory scene analysis (ASA) is a commonly used model for thinking about how humans organize the sounds in their environment. Stream segregation, the process by which we group/separate sounds by their sources according to ASA, can play an important role in how we perceive music, with factors such timbre and pitch helping us to segregate the sounds in a piece of music as belonging to different instruments. Building off of a pre existing literature on expert perception in music, I conduct an experiment which is meant to test a) if there is a relationship between level of musical experience and success in a stream segregation task in the context of recorded music and b) if the presence of vocals affects performance in this stream segregation task in different ways with differing levels of musical experience. The results of this experiment seem to indicate that success in stream segregation in the context of recorded music increases with increased level of musical experience. The results also indicate that while no relationship exists between level of experience and success with stream segregation with vocals present, subjects at all or most levels of experience perform worse when vocals are present.

Auditory Scene Analysis and Musical Experience

Literature Review

In our everyday lives, we are constantly inundated with acoustic information via sound waves which cause our eardrums to vibrate. In order to make sense of our environment, we need some way to differentiate and individually understand the various sound waves which are entering our ears at any given moment. One popular model for understanding how humans do this is known as auditory scene analysis (ASA), according to which humans separate the different soundwaves hitting the eardrum by grouping and segregating them by source (i.e. which object/organism is creating the sound). These groupings of sound are referred to as auditory streams, and the process of segregating them is called stream segregation. To engage in stream segregation, we rely on both sequential (time-based) and simultaneous (frequency-based) grouping mechanisms (Bregman, 1996). For example, when someone listens to a piece of music featuring a singer and a guitarist, it is likely that they are (consciously or unconsciously, successfully or unsuccessfully) engaging in stream segregation by perceiving differences in rhythm and timing (sequential) and/or differences in pitch and timbre (simultaneous) in order to separate the sounds as belonging to two different instruments.

For ASA in the context of music, timbre (the "character" of a sound, as opposed to its pitch or intensity) plays a particularly important role in stream segregation, since an instrument's timbre is what differentiates its characteristic sound from that of other instruments. When two instruments within a song are very different timbrally, stream segregation is generally more likely to occur; instruments which are very similar timbrally (e.g. two instances of the same instrument) are more difficult to separate (Ashley et al., 2019). Considering that one's ability to engage in stream segregation may be influenced by learned patterns as well as raw perception

(Bregman, 1996), it is worth considering that one's familiarity with the timbres being presented to them in a stream segregation task may influence their performance. Therefore, musicians, who spend a great deal of time both playing and listening to different instruments, may have an advantage over nonmusicians in music-based ASA tasks (particularly stream segregation).

One study by Johnson et al. (2020) demonstrates that musicians tend to perform better on timbre-based stream segregation tasks even when traditional instruments are not involved. In their experiment, each subject (either a musician or nonmusician) was presented with three musical tones, each with the same fundamental frequency. All three tones had five harmonics on top of the fundamental frequency. For two of these tones, the amplitudes for each of these harmonics were identical, thus making the tones timbrally identical; for the remaining tone, the amplitude of the third harmonic was raised somewhat, giving it a slightly different timbre. The subject was then asked to identify which tone was different from the rest. The researchers found a statistically significant result indicating that the musician group performed better on this task than the nonmusician group, which would seem to suggest that musicians have an advantage over nonmusicians in timbre-related stream segregation tasks even without the effect of familiarity. With this (as well as other possible advantages relating to perception of rhythm, timing, and pitch) in mind, one would expect to see a relationship between level of experience as a musician and success in music-based ASA tasks, particularly those involving stream segregation.

There have been many other studies examining the relationship between level of musical experience and performance on music-related tasks as part of a broader literature on expert perception. For example, researchers have found that viola players with more experience tend to perform significantly better in memory tasks related to classical Western tonal music than those with less experience (Knecht, 2003), subjects tasked with playing “air piano” to piano recordings tend to display gestures which more accurately mimic the recording with regards to pitch and dynamics (among other factors) when they have more musical experience (Godøy et al., 2006), and the cues used to attribute certain emotions to a piece of music differ between experts and non-experts (Spitzer & Coutinho, 2014). One potential area of research which I did not see reflected in the literature of expert perception in music was the effect of vocals on stream segregation for experts versus non-experts; this, along with general performance in stream segregation tasks with experts versus non-experts, is what I aim to explore with this paper.

Methods

Apparatus/Stimuli

The data for this experiment were collected using an online survey (created with Google Forms) that contained links to each of audio clips that the subjects were intended to listen to. These clips were hosted on Soundcloud, which allowed the subjects to repeat each clip and skip through portions of it as they desired. Each of the clips were (roughly) 20 second snippets taken from live music performances found on YouTube; although the sources for all of the clips originally had a visual component, the subjects were presented with the audio only. Since this experiment took the form of an online survey, I did not have control over the subjects’ listening environments beyond the instruction at the beginning of the survey to listen to the audio clips with headphones. Ideally, this would mean that all of the participants listened to the audio clips using some kind of device which sends sound directly into the ears (e.g. closed-back headphones, on-ear headphones, earbuds); however, without direct supervision, it is impossible to determine if every subject followed the instructions.

Four musical genres were represented in the eight clips included in the experiment, with a clip containing vocals and a clip not containing vocals for each genre. These four genres (bluegrass, jazz, metal, and classical) were selected because their performances generally consist entirely of live instrumentation; genres which heavily utilize tools like samplers and synthesizers (e.g. hip hop and electronic) tend to blur what counts as an instrument and thus weren't included. The inclusion of multiple genres was intended to account for differences in taste/experience between subjects; in other words, it would be unfair to metal musicians/listeners to only include samples of jazz music and vice versa. All of the vocals in the clips were in English in order to potentially enhance the effect of worsening performance (i.e. presenting semantic information to potentially distract from the music).

Design

Error rate, defined as the difference between a subject's estimate for the number of instruments in a clip and the actual number, was the only dependent variable that I tracked in this experiment. The independent variables were vocal presence (clip contained vocals or did not contain vocals) and musical experience (subject's subjective rating of their level of experience as a musician, on a discrete scale from 1 to 7). Since all of the subjects were exposed to clips both with and without vocals, vocal presence is a within-subjects variable. Since musical experience varies from subject to subject, it is a between-subjects variable.

Procedure

The survey began by instructing the subject to first provide their name and level of musical experience. The scale for musical experience was explained as follows: 1 describes someone with no musical experience whatsoever, 3 describes an amateur musician who practices/plays irregularly, 5 describes an intermediate musician who practices/plays somewhat regularly, and 7 describes an advanced/professional musician who practices/plays daily. The next instructions directed the subject to copy/paste the provided Soundcloud links into their browser, listen to the clips all the way through for as many times as needed, and then respond with the number of individual instruments they heard in each recording. The following rules were provided to eliminate potential points of confusion: multiple instances of the same instrument are to be counted as multiple instruments, an instance of a drum set is to be considered one instrument, and vocals count as an instrument. In addition, the subject was instructed to wear headphones while listening to the clips.

For each clip, the subject could choose any number between 1 and 10 to indicate how many instruments they heard; the range of actual values for the clips was 3 to 6, meaning that the subjects were able to both underestimate and overestimate the number of instruments for any clip. The clips were arranged in the same randomly chosen order for all subjects, although the design of the survey was such that a subject could provide their answers in any order.

Subjects

20 subjects participated in the experiment, with all of them completing it in full. Although ages were not recorded as part of the survey, I know that the age range was roughly 20 to 50 years old based on my knowledge of the participants. Of the 20 participants, 4 (20%) ranked themselves as a 1 in musical experience, 2 (10%) ranked themselves as a 2, 2 (10%) ranked themselves as a 3, 1 (5%) ranked themselves as a 4, 9 (45%) ranked themselves as a 5, and 2 (10%) ranked themselves as a 7, with no subjects in the 6 category.

Results

Experience Level & Error Rate

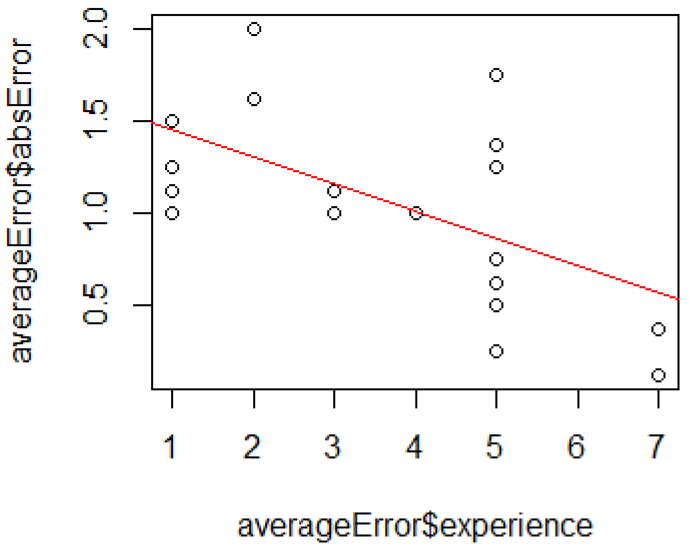
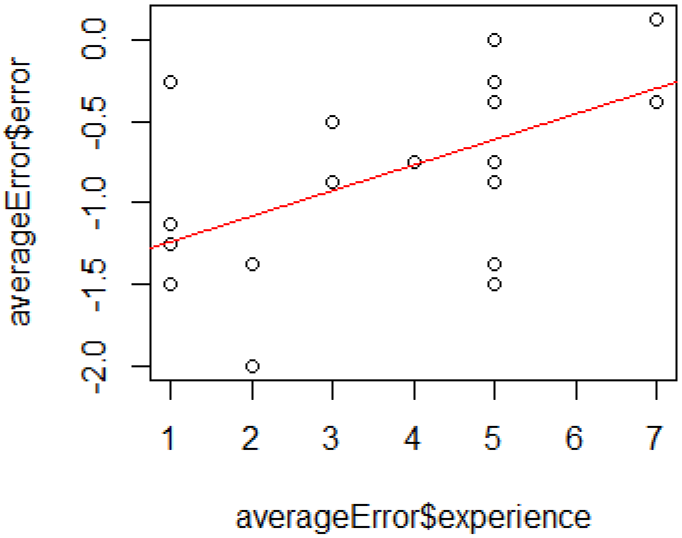


Figure 1: Average absolute error for each subject plotted as a function of experience level.

The regression line of average absolute error (the absolute value of the error rate) on experience level has an intercept of 1.6086 and a slope of -0.1483; on average, as experience increases by one, average absolute error decreases by 0.1483. The Pearson correlation coefficient calculated from these data is -0.5739361, which results in a statistically significant p-value of 0.008141.

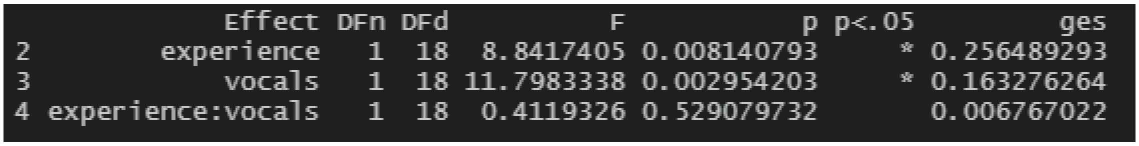
The regression line of average error on experience level has an intercept of -1.3984 and a slope of 0.1587; on average, as experience increases by one, average error increases by 0.1587. The Pearson correlation coefficient calculated from these data is 0.5262375, which results in a statistically significant p-value of 0.01715.

Figure 2: Average error for each subject plotted as a function of experience level.



Vocal Presence

Table 1: Results of an ANOVA with absolute error as the dependent variable.



```
             Effect DFn DFd          F           p p<.05         ges
2        experience   1  18  8.8417405 0.008140793     * 0.256489293
3            vocals   1  18 11.7983338 0.002954203     * 0.163276264
4 experience:vocals   1  18  0.4119326 0.529079732       0.006767022
```

As shown in Table 1, I found a main effect of experience level on absolute error rate, F(1, 18) = 8.84, p = 0.0081 with average absolute error rate generally decreasing with increased experience levels; experience = 1 (M = 1.22), experience = 2 (M = 1.81), experience = 3 (M = 1.06), experience = 4 (M = 1.00), experience = 5 (M = 0.96), experience = 7 (M = 0.25). I found a main effect of vocal presence on absolute error rate, F(1, 18) = 11.80, p = 0.0030 with higher absolute error on clips which included vocals (M = 1.25) than on clips which did not include vocals (M = 0.83).

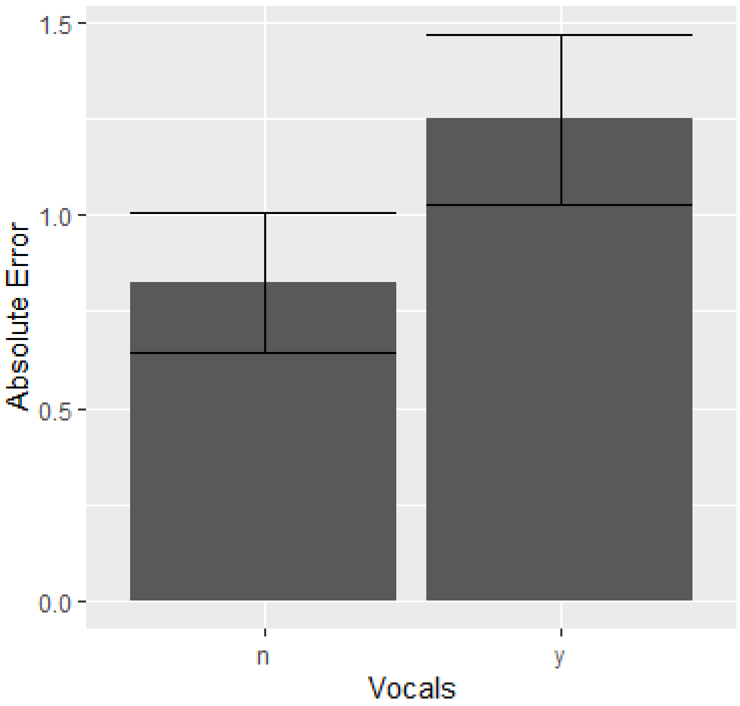


Figure 3: Bar graph showing difference in mean absolute error for vocal presence conditions.

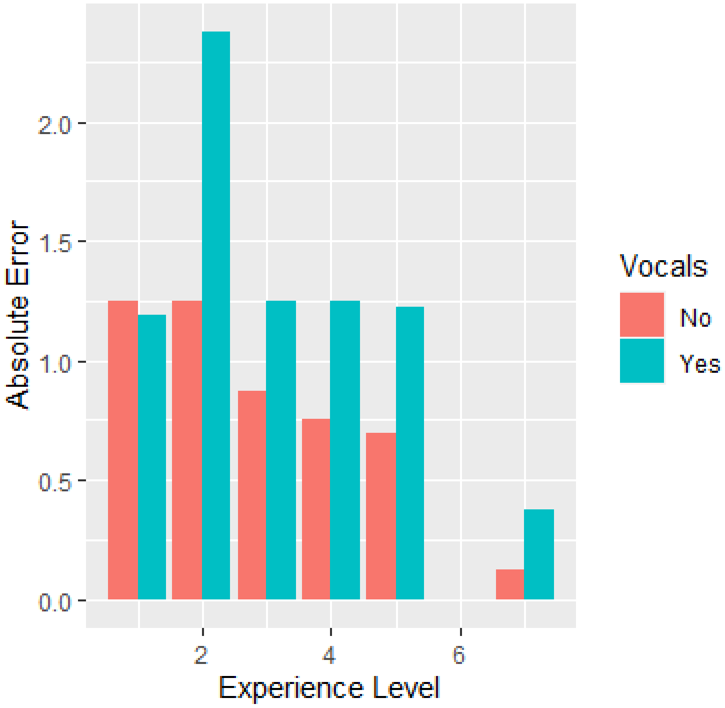


Figure 4: Bar graph showing differences in mean absolute error for vocal presence conditions at different experience levels.

Discussion

Based on the results in the **Experience & Error Rate** section, we can conclude that absolute error rate in a stream segregation task (identifying the number of instruments in a piece of music) is negatively correlated with level of musical experience; in other words, performance generally improves with increasing levels of musical experience. It is likely that this correlation is related to the findings of Johnson et al. (2020) in that the different timbres coexisting in these clips likely gave some important cues for stream segregation. However, when dealing with pieces of recorded music rather than just simple tones, it also seems likely that rhythm, timing, pitch, and location (where each instrument is located in the mix) could have played similarly important roles in triggering stream segregation, and thus future research could perhaps be directed at singling out these elements of ASA to find the relationship between them and musical experience. In addition, taking note of each subject's level of experience with particular timbres and/or combinations of timbres (perhaps by asking about genre preferences/familiarity) could yield interesting results relating learned patterns to performance on stream segregation tasks.

We can also see from *Figure 2* and its related data that the average error rate (without finding the absolute value) for almost every subject was negative, implying that subjects generally underestimated the number of instruments in the clips. This finding is exactly what one should expect to see if stream segregation truly is how we perceive music, since failure to find the correct number of instruments would be dependent on the inability to separate streams to a fine level of detail and the subsequent erroneous grouping of multiple instruments into the same stream.

The results in the **Vocal Presence** section reaffirmed the previously explored relationship between absolute error rate and musical experience while also demonstrating an even stronger relationship between absolute error rate and vocal presence. Although the original goal of the section was to attempt to find a relationship between vocal presence and experience level (hence the use of an ANOVA), there seems to be no significant difference between how subjects of different experience levels perceive music with vocals; instead, it seems that subjects of almost all levels tend to perform worse in the stream segregation task when vocals are present. Interestingly, the only level for which this wasn't true was 1, the subjects who considered themselves to be totally inexperienced. Based on intuition, I thought that people with less musical experience would tend to hyperfocus on the vocals and thus show inhibited stream segregation, since a) their lack of knowledge relating to instruments may draw them toward the human voice, which is created by a more familiar structure and/or b) the presence of semantic information in the form of English lyrics (something which may be of more interest to them) may distract from the instrumentation. This isn't to say that these things aren't happening for level 1 subjects (the slightly higher error rate for the no vocals condition could be a fluke), but rather that all the other levels seem to be affected adversely by vocals, in which case these explanations make less sense. It may be the case that humans in general tend to focus on the sound of the human voice when listening to music; further research based on this idea could yield some interesting results.

The overall design of this experiment was far from perfect, and certain changes could be made to possibly get more accurate results. Perhaps the most glaring issue was the lack of control over each subject's listening environment, since different pairs of headphones may boost/hinder certain frequencies and thus lead to different listening experiences. Ideally, each subject would listen to the clips on the same device with the same pair of headphones, but it would have been difficult to do so and also get an acceptable number/range of subjects. The distribution of musical experience levels in my sample was also unbalanced, which could have been remedied by simply finding more participants for the levels which were lacking (e.g. 6). It is also worth noting that this experiment could not hope to capture the normal listening habits of its participants, since they are being explicitly asked to listen for instruments in the mix. It is possible that many participants would perform better/worse (most likely worse) if their normal level of attention when listening to music was applied for this experiment, but it would be very difficult to design an experiment in which this is possible. Despite these issues, it seems that this experiment has led to some interesting results relating to ASA in the context of music and suggests potential future avenues for research in this area.

References

Ashley, R., Timmers, R., McAdams, S., & Goodchild, M. (2019). Musical Structure: Sound and Timbre. In *The routledge companion to music cognition* (pp. 281–301). essay, Routledge, Taylor & Francis Group.

Bregman, A. S. (1996). *Auditory scene analysis*. MIT Press.

Godøy, R. I., Haga, E., & Jensenius, A. R. (2006). Playing "Air instruments": Mimicry of sound-producing gestures by novices and experts. *Lecture Notes in Computer Science*, 256–267. https://doi.org/10.1007/11678816_29

Johnson, N., Shiju, A. M., Parmar, A., & Prabhu, P. (2020). Evaluation of auditory stream segregation in musicians and nonmusicians. *International Archives of Otorhinolaryngology*, *25*(01). https://doi.org/10.1055/s-0040-1709116

Knecht, M. G. (2003). Music expertise and memory: The relationship between music expertise and memory of music patterns, within various degrees of contextual constraint. *Music Education Research*, *5*(3), 227–242. https://doi.org/10.1080/1461380032000126328

Spitzer, M., & Coutinho, E. (2014). The effects of expert musical training on the perception of emotions in Bach's Sonata for unaccompanied violin no. 1 in G minor (BWV 1001). *Psychomusicology: Music, Mind, and Brain*, *24*(1), 35–57. https://doi.org/10.1037/pmu0000036