An exploration of auditory stream segregation

Zhengshan Shi

Music 251 Final Project Paper

Jun.2013

Abstract: The human auditory system has the ability to segregate one auditory stream from another. How what we know affects what we actually perceive remains an interesting question. In this article, the author aims to explore the relationship between schema-based processing using previous knowledge of a melody and the perceptual organization of the sound. Two unfamiliar six-tone melodies were presented sequentially to the listeners, one of which was interleaved with a distractor melody. The listeners were required to tell whether the two melodies were similar or different. In order to explore the relationship between the previous knowledge, timbre information of the target melody was also presented. The paper explores whether it will affect auditory stream segregation.

Keywords: auditory scene analysis, schema-based processing, auditory stream segregation

Introduction:

Everyday we hear thousands of sounds, and the sound waves received by our ears are a mixture of all the effects and are distributed over time. However, the human brain has the ability to tell what kinds of sounds they are listening to. Instead of filling our head with a messy mixture of sound, the human auditory system intelligently segregates the streams arriving at the ears, both voluntary and involuntary. For example, when we are in a noisy room that is filled with peoples' conversations, we can voluntarily focus on one of the speakers among a mixture of conversations and background noises. This example is related to the famous "cocktail-party effect". Sounds are composed of the combination of frequency, amplitude and timbre, and the structure of our auditory system allows us to differentiate the sound sources coming to our ears.

In 1975, Leo van Noorden presented a sequence of alternating tones consisting of the letters "F" and "V" and discovered that the rhythm became ambiguous when the stream was galloping. Bregman(1990) concluded such experiments and proposed the term "Auditory Scene Analysis". Auditory Scene Analysis (ASA) describes the process by which the auditory system decomposes sound into elements through segregation and then recombines them into streams for better perception on the basis of Gestalt's principle of similarity, continuity, etc. The ASA includes two main processes: the primitive process and schema-based grouping. Primitive process uses sequential integration and spectral integration to decompose the incoming signal, describe the components according to their characteristics and group them based upon time and spectral information. The schema-based process employs listener's knowledge in varied acoustic environments. The attention in the auditory system, according to Bregman, can select a portion of the sensory information for more detailed processing. The auditory attention helps the brain break input stimuli into parts and segregate them for higher-order perception.

So how does previous knowledge of the melody help the process of auditory scene analysis, and how does it interact with the auditory attention? According to

previous research by Bay & McAdams (2002), if the whole melody is presented previously,, it's easier for humans to segregate from a complex sound mixture at a later point in time.. However, if only 'part' of the information were given, would that still improve stream segregation?

The current study is aimed at exploring the effect of timbre information when it is presented before the complete melody. A timbre cue was first presented to the listeners, then six unfamiliar tones were interleaved with six distractors, and were presented to the listeners. In the final stage, the listeners were given another melody and asked to tell if the two target melodies were identical. The target melody was a sine tone, and the distractors were synthesized sounds of clarinet and mandolin. A higher accuracy is expected for the mandolin distractor and with the cue information.

Method:

Stimuli:

Eight target melodies and eight distractors were constructed, each composed of six tones. Each of the eight target melodies had one original, and one modified version. For the modified version, one to two notes were changed within a range of 5 semi-tones. The mean note of target melodies is D5 (midi note number 74, frequency 587.33Hz). All of the melodies are diatonic.

The target melodies were synthesized using sine tones. The distractors were synthesized using two instrument timbres:: clarinet and mandolin. For each trial, a target melody was interleaved with either a clarinet distractor, or a mandolin distractor, and then the target melody was presented again, either altered or unaltered. Before the interleaved notes, there are three conditions: (1) no cue : nothing was presented before; (2) 'same cue' : a sine tone cue using the same first note as the target melody was presented; (3) 'different cue':, a sine tone cue using a different note as the first note of the target melody was presented.

For each target melody, there are twelve experimental variations with the following six conditions:

1. no cue and altered melody
2. no cue and unaltered melody
3. same cue and altered melody
4. same cue and unaltered melody
5. different cue and altered melody
6. different cue and unaltered melody

Two different instruments clarinet and mandolin contribute to a total of 12 variations. The experiment consists of a total of 96 trials and 2 practice trials in the experiment.

Melodies and distractor sequences were composed of pure sine tones, 200 milliseconds long for each note. The interval was 500ms before the appearance of the comparison melody. If there is cue information included, the length of the cue was 800 milliseconds. Between each trial, there is a 2-second interval of silence. Figure 1 illustrates the time schedule of each trial:

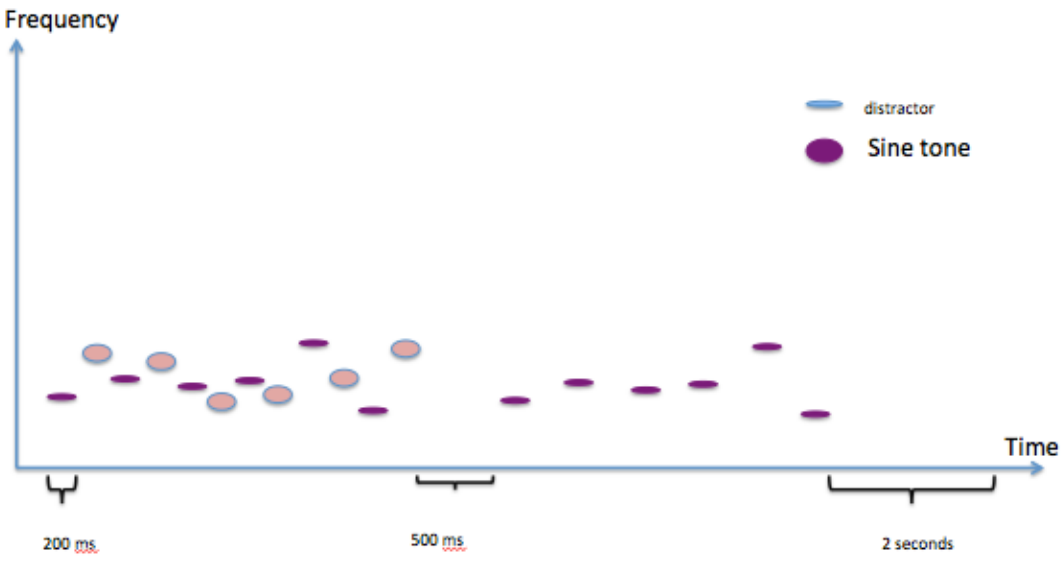
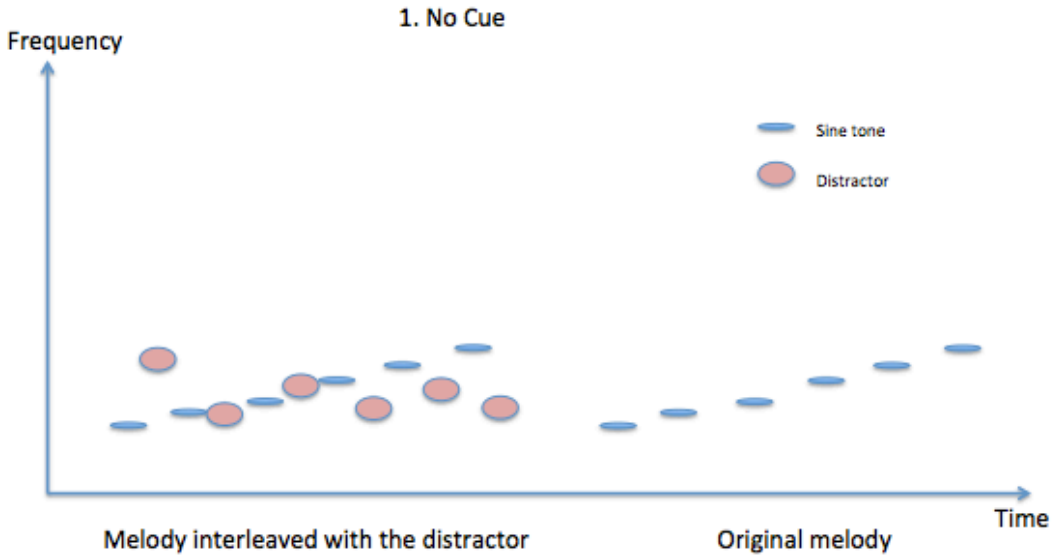


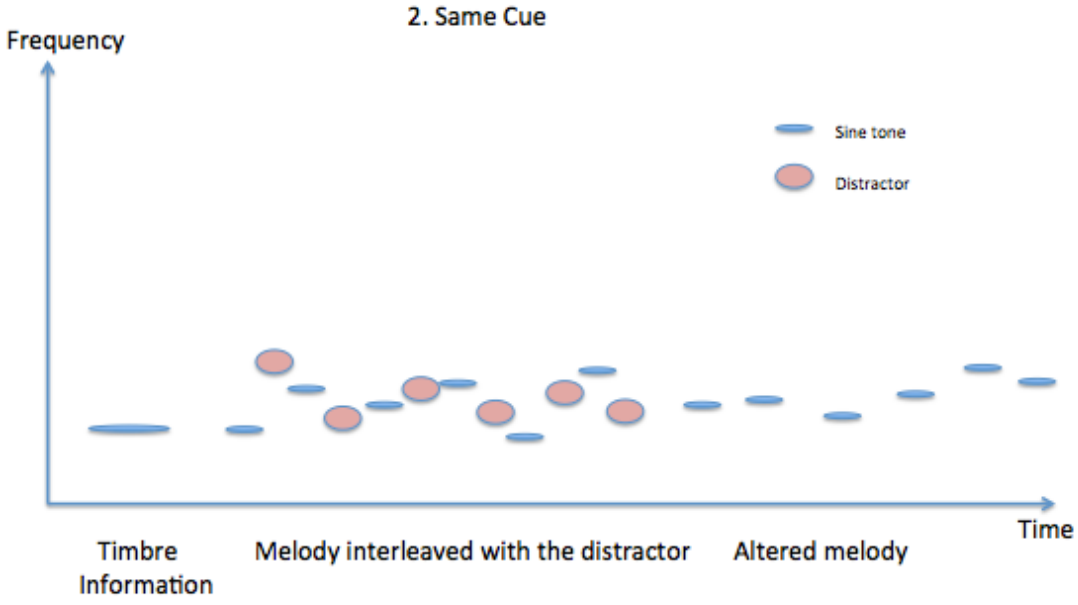
figure 1. time arrangement of each trial (without cue information)

Procedure:

Two unfamiliar six-tone melodies were presented successively to listeners, one interleaved with the distractor tones. Three experimental conditions were presented to each listener. The first condition is the 'no cue' condition, meaning that no extra information about the melody was given before each trial. In this condition, the listeners have no previous knowledge about the melodic information they'll need to extract. The second condition is called the 'same cue' condition where a sine tone of roughly 800 milliseconds was presented to the listeners which was the first note of the target melody. In this condition, the listeners were informed of the timbre and the first note of the melody they would need to extract. The third condition is called the 'different cue' condition in which a sine tone of a different pitch was presented -- In this case, only the timbre information was given. Before the whole experiment, two simple practice trials were presented to the listeners to improve their familiarity with the experiment. The distractors and the target melodies were designed to be within the same frequency range to avoid primitive stream segregation due to varied frequency ranges.

Between each trial, there is a two-second interval to guarantee that the listeners have enough time to prepare for the next trial. All trials were shuffled randomly before presenting to the listeners. The listeners were asked to hear a wave file approximately fifteen minutes in length, in which all of the 96 trials plus two practice trials were presented, and they were asked to type their answers into an excel sheet: marking an 's' representing if the two target melodies are identical, and 'd' if they are perceived as different melodies. The listeners were required to hear the whole experiment only once and without pause during the experiment's duration.. Figure 2 illustrates three different conditions of the experiment:


1. No Cue
Frequency
Sine tone
Distractor
Time
Melody interleaved with the distractor
Original melody


2. Same Cue
Frequency
Sine tone
Distractor
Time
Timbre Information
Melody interleaved with the distractor
Altered melody

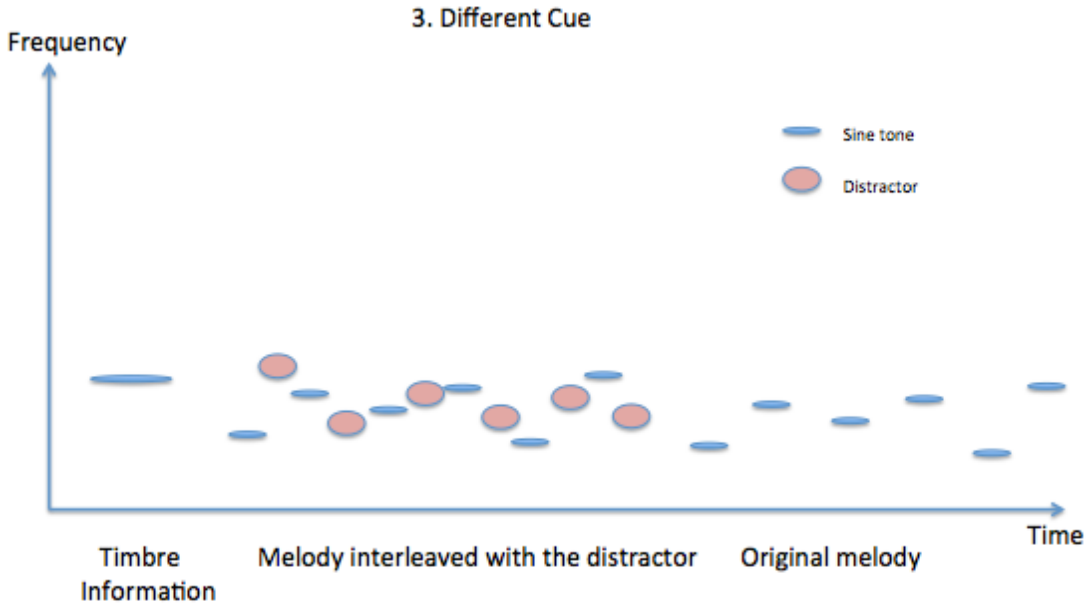


figure 2. Three different conditions: no cue, same cue and different cue

At the data analysis stage, the correct rate was first revised by hand, and then according to the order of the trial, it was labeled as different conditions in the experiment. Each trial was identified in sequence by numbering with the following format: instrument id + melody no. + variation no. 'C' and 'M' represent the clarinet distractor and the mandolin distractor; 1-8 represents eight different melodies, and 1-6 at the last slot stands for the variation of the experiment: (1) for no cue and altered melody, (2) for no cue and unaltered melody, (3) for same cue and altered melody, (4) for same cue and unaltered melody, (5) for different cue and altered melody, and (6) for different cue and unaltered melody. For example: sequence no.m_6-2 represents a mandolin distractor with an unaltered melody and no cue.

Then each randomly shuffled trial was switched back into normal order using the excel sorting functions so that the accuracy of different conditions could be collected quickly. For each instrument category, the following data was collected: total correct rate for each instrument, no cue correct rate, same cue correct rate, and different cue correct rate. Also, the overall correct rate was calculated regardless of instrument.

Apparatus:

Both the target melody and the distractor were synthesized using Chuck[1]. For the target melody, a sine oscillator was used for synthesis, and for the distractor, physical modeling instruments of a clarinet and mandolin were picked. All trials were generated in Chuck, and then exported into audacity[2] using the Sound flower[3] routing utility. Finally, a wave file of all 96 trials plus the two practice trials were obtained. The subjects were listening to the experiment on a macbook pro over a pair of Shure SRH440 headphones.

Subjects:

Twelve listeners took part in the experiment. They all reported having normal hearing. Six of the participants are musicians, and the remaining six are non-musicians. The criterion of 'musicianship' was classified by a training of over six years on classical instruments. Among the six musicians, three had particpated in formal ear training classes. For each listener, all of the trials were presented regardless of musician or non-musician status. As stated, each experiment lasted approximately 15 minutes, and the subjects were not permitted breaks.

Results:

The data from the twelve subjects was analyzed. However, two of the datasets were excluded after the analysis of the correct rate, because both subjects expressed a difficulty in recognizing the melody, and thus they failed to mark some of the trials. These subjects reported 'same' for almost 90% of the trials, and each case the subjects were non-musicians. Thus, the further analysis includes the data from the remaining ten subjects. These datasets contain no missing responses. Figure 3 plots the mean error rate (error numbers over 48 trials each) for two distractors: clarinet and mandolin. Figure 4 shows the average correct rate (correct numbers over 16 trials for each condition) with a classification of the cue information: no cue, same cue and different cue.

--Difference between different distractor

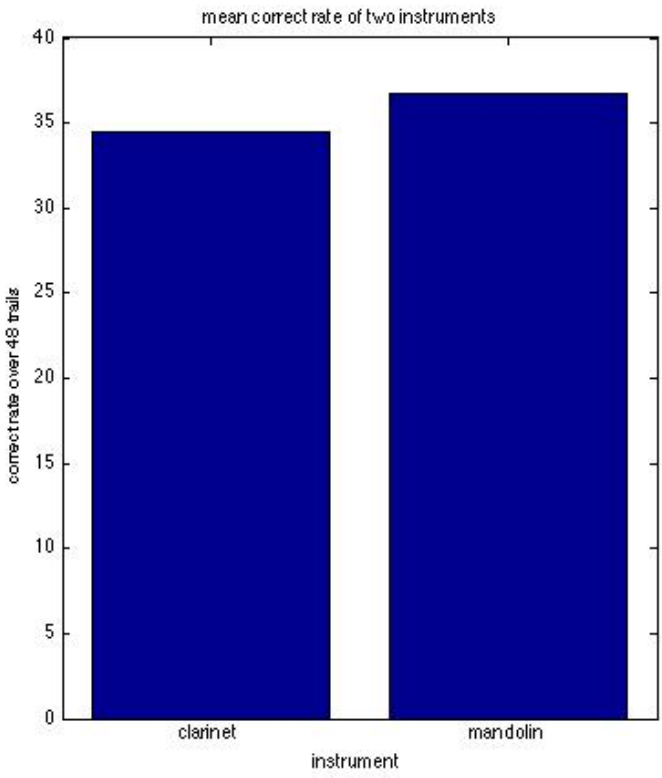


figure 3. average correct rate of two different distractors (over 48 trials)

Figure 3 illustrates the overall performance of two different distractors, the left bar is clarinet, and the right bar is mandolin. Overall, using mandolin as the distractor gives higher rate of correctness.

--Difference between cue information

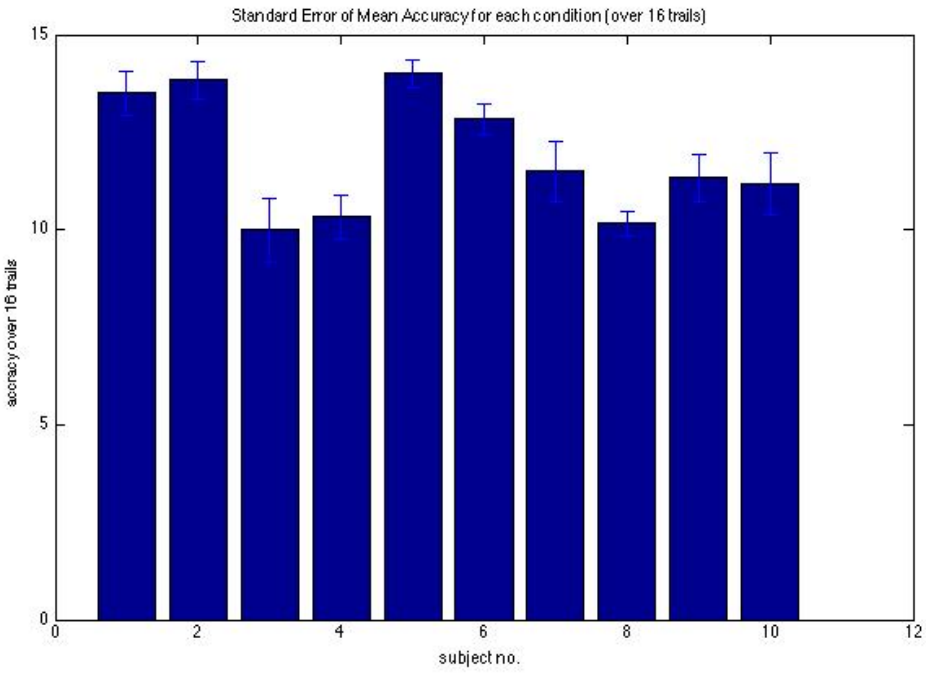


figure 4. standard error mean accuracy for each condition (10 subjects)

Figure 4 shows the standard error mean (SEM) of the accuracy over 16 trials in each condition (clarinet: no cue, same cue, different cue; mandolin: no cue, same cue, different cue). There is a relatively high fluctuation from subject to subject. . The data clearly shows a disparity between the sujects who were musicians and non-musicians. The overall performance in different conditions is illustrated in the following figure 5.

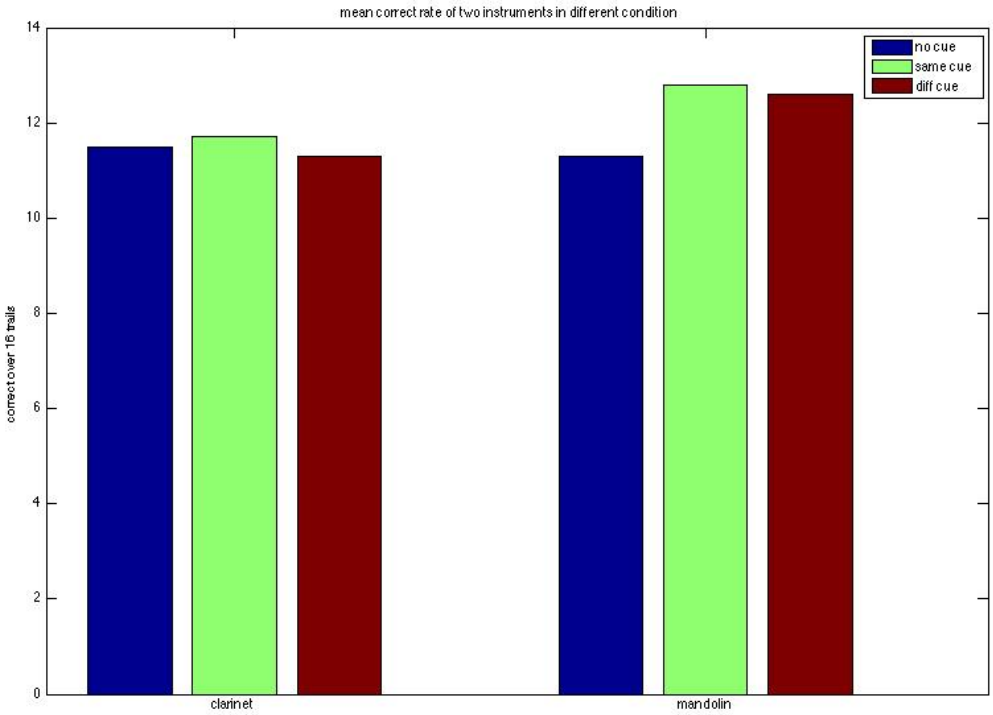


figure 5. performance with cue information

Discussion:

The study has shown that the ability to recognize interleaved melodies varies from person to person according to previous musical training or ear training experiences. It was also dependent upon the variables in different conditions. Recognition performance was higher when the difference in timbre (interval) of the distractor is greater. The interval between the mandolin and sine tone is greater than the interval between the Clarinet and sine tone. In a situation where the timbre cue of the target melody was given prior to the interleaved melody, although the average performance of recognition is higher, which is the same as the author's hypothesis, the difference between the 'same cue' and 'different cue' is not as great as previously proposed. The experiment shows that giving a cue of the same note or different note as the target melody doesn't affect the recognition greatly.

Using ANOVA, the following conclusions were drawn:: the main effect of Timbre (Clarinet/mandolin) was approaching to significance: $F(1,9) = 5.060$, $P = 0.051$, due to the greater performance in the mandolin condition compared with that in

the Clarinet condition. Although cue information in the mandolin condition is more significant than in the Clarinet situation, the overall effect of no cue / same cue/ different cue is not significant..

An explanation of the better performance with the 'mandolin' distractor is due to the primitive analysis of frequency and timbre. Computer-synthesized clarinet sounds very similar to a sine tone. According to previous research, Bregman concludes that sounds of similar timbres will group together. So if the distractor is mandolin, successive sounds of the sine tone will be segregated from those of the mandolin, even when they are playing in the same frequency range. The filter model of Broadbent(1971), suggests the concept of categorization:: the filter groups the input signal according to different characteristics and then passes this information to higher levels. The filter now allows some stimuli to be responded to, but of less evidence than others, rather than the former theory that it blocks out all unattended messages. Where as for the clarinet, our auditory system groups the clarinet tones with the sine tones so that it becomes more difficult to segregate. This phenomenon in the auditory system comes from Gestalt's principle of similarity and proximity on visual grouping: we tend to cluster similar members of a set together.

An explanation about why the cue information helps stream segregation during the mandolin condition is this: the cue information attracts the attention of the auditory system and thus helps focus on the target melody of the same timbre. Another possible reason is that, although there is a 2-second interval between each trial, it is not enough time for the brain to wipe the image of the previous melodies while the presentation of the 800-ms sine tone has an effect on clearing the short-term memory so that the listeners will be less likely to be affected by previous trials.

Interestingly, in the clarinet condition, cue information doesn't help the accuracy of the recognition significantly. A possible explanation is still due to the similarity between the sine tone and the synthesized clarinet tone. According to Scharf(1998), hearing must be sensitive to new sound, must be ready for sounds from anywhere since it is omni-directional, and we don't have any predictive information for what is coming next. After we receive the sound, we could ascertain the characteristics of the

sound and then segregate one from another according to different features such as spatial location, frequency, et al.... Humans could voluntarily select which sound to attend to, while ignoring others based upon their own selection or personal experiences. In the meantime, we retain little memory for the information that we don't pay attention to. The auditory system seems to process the incoming signal first and then engages in a period of selection. Thus, when tones of similar timbres were presented, the ear has no primitive frequency band to select from, and we are not prepared for the information to come, which leads to a lower performance of recognition.

Overall, the experiment was in-line with the the author's hypothesis while leading to some interesting findings. After the analysis of the data, the author found that the following aspects needed to be improved for the next experiment:

(1) The 'practice trial'.

The two 'practice trials' are too short for the listeners to get used to the nature of the experiment.. In the design of next experiment, the author should record a set of aural instructions pertaining to the experiment and a demonstration of the practice trial. Also, prior to the practice trial, there should be some controlled experiment to identify if the subject has a given level of ability to identify whether two seperate melodies are identical or different. Using two tones without any interleaving is a suggested method.

(2) Timing issue

The total 96 trials were presented strictly on a fixed timescale to the listeners: a two second interval between subsequent trials.. In order to prevent a situation where the listeners did not have enough time to identify and write down their answers, a subject determined pause in the experiment should be allowed throughout the entire process.

Conclusion:

The auditory system not only helps us localizing each source, but also endows us with the ability to discriminate between simultaneous sounds. Perceptual features of the sound attribute to the identifying orientation, selection, segregating and combination period of the auditory system. In this paper, the author is trying to explore how schema-based processing affects auditory attention and auditory stream segregation. The experiment was designed to explore whether previous knowledge will help the performance of auditory stream segregation.

After the auditory system receives the sound, it identifies the characteristics of the sound and then segregates one from another according to various features such as spatial location, frequency, and others. From the experiment we found that if the distractor is of a farther timbre interval from the target melody, then it is easier for the brain to segregate the two streams. When a previous knowledge – timbre information in this experiment - is given, the performance of stream segregation is better compared with the condition that no prior knowledge is given to the subject..

The research on auditory stream segregation and auditory attention has significance in that it explores humans' ability to perceive sounds from the stage of detection to selection. It is also important in the area of cross-modal attention or dual-task research. Although the auditory system differs from the visual system in the way it perceives and selects objects, concepts in both can be integrated. The interaction and integration of sensory information including the auditory visual system, or even touch information could help the study of human sensing and cognition, and even aid research deeper into the memory. With the development of technology and theory, we can foresee a bright future of auditory attention and the cross-modal attention research.

Future work based on this experiment will be the exploration of the spatial information on stream segregation. Rather than interleaving two tones in different timbre, a future task is to interleave two tones in different spatial position. Instead of giving the timbre information before, a spatial cue will be given before the

presentation of the interleaved melody. This relates to the study of head-related transfer function and 3D spatial auditory perception. The goal is to explore what kind of knowledge is important to draw auditory attention and thus help improve the understanding of auditory stream segregation. A further goal is to explore how we could model the human auditory system and apply its stream segregation to the area of source separation and computational auditory scene analysis when done using computers..

References:

Bey, C., & McAdams, S. (2002). Schema-based processing in auditory scene analysis. *Percept. Psychophys. 64, 844-854.*

Bregman, A.S. (1990). *Auditory Scene Analysis. The Perceptual Organization of Sound*, Cambridege, Mass: MIT Press.

Broadbent. D.E. (1952). Listening to one of two synchronous messages. *Journal of Experimenta*l Psychology, *44,* 51-55.

Broadbent. D.E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology, 47*, 191-196.

Egeth, H. (1992). Dichotic listening: Long-lived echoes of Broadbent's early studies. *Journal of Experimental Psychology: General, 121(2)*, 124.

Lupyan, G. (2009). Cognitive Influences on Attention. Ed. B. Goldstein, *The Sage Encyclopedia of Perception.* (pp.70-74). Sage Publications, Inc.

Pressnitzer, D., Suied, C., & Shamma, S.A. (2011). Auditory scene analysis: the sweet music of ambiguity. *Frontiers in Human Neuroscience*, 5, 158.

Scharf, B. (1998). Auditory attention: The psychoacoustical approach. In H.Pashler. *Attention.* Hove, UK: Psychology Press.

Spence, C. J., & Driver, J. (1996). Covert spatial orienting in audition: Exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 555-574.

Styles.E.A. (2006). *The Psychology of attention.* Hove England; New York: Psychology Press.

Treisman, A.M.. (1969). Strategies and models of selective attention. *Psychological Review, 76(3)*,282-299.

Appendix: Table

Subject No. 1, Musician

	Total	no cue	good cue	ok cue
Clarinet:	42/48	15--16	14--16	13--16
Mandolin:	39/48	12--16	15--16	12--16
Total:	81/96	27/32	29/32	25/32

Subject No. 2, Musician

	Total	no cue	good cue	ok cue
Clarinet:	43/48	13--16	14--16	16--16
Mandolin:	40/48	13--16	14--16	13--16
Total:	83/96	26--32	28--32	29--32

Subject No. 3, Musician

	Total	no cue	good cue	ok cue
Clarinet:	27/48	11--16	8--16	8--16
Mandolin:	33/48	11--16	9--16	13--16
Total	60/96	22/32	16/32	21/32

Subject No. 4, Musician

	Total	no cue	good cue	ok cue
Clarinet:	28/48	9--16	9--16	10--16
Mandolin:	34/48	10--16	12--16	12--16
Total:	62/96	19/32	21/32	22/32

Subject No. 5 Musician

	Total	no cue	good cue	ok cue
Clarinet:	41/48	13/16	14/16	14/16
Mandolin:	43/48	13/16	15/16	15/16
Overall	84/96	26/32	29/32	29/32

Subject No. 6 Musician

	Total	no cue	good cue	ok cue
Clarinet:	37/48	11--16	13--16	13--16
Mandolin:	40/48	13--16	13--16	14--16
Overall:	77/96	24--32	26--32	27--32

Subject No. 7 NonMusician

	Total	no cue	good cue	ok cue
Clarinet:	30/48	13/16	12--16	8--16
Mandolin:	36/48	11--16	12--16	13--16
Overall:	66/96	24--32	24--32	21--32

Subject No. 8 NonMusician

	Total	no cue	good cue	ok cue
Clarinet:	30/48	11--16	9--16	10--16
Mandolin:	31/48	10--16	11--16	10--16
Total:	61/96	21/32	20/32	20/32

Subject No. 9 NonMusician

	Total	no cue	good cue	ok cue
Clarinet:	33/48	11--16	12--16	10--16
Mandolin:	35/48	10--16	14--16	11--16
Total:	68/96	21/32	26/32	21/32

Subject No. 10 NonMusician

	Total	no cue	good cue	ok cue
Clarinet:	31/48	8--16	12--16	11--16
Mandolin:	36/48	10--16	13--16	13--16
Total:	67/96	18/32	25/32	24/32

Table1. Data for individual subject, good cue = same cue, ok cue = diff cue