THE EFFECT OF VISUAL CUES ON MELODY SEGREGATION

Jeremy Marozeau
The Bionic Ear Institute,
Melbourne, Australia
jmarozeau@bionicear.org

David B. Grayden
University of Melbourne
Melbourne, Australia
grayden@unimelb.edu.au

Hamish Innes-Brown
The Bionic Ear Institute,
Melbourne, Australia
hinnes-brown@bionicear.org

Anthony N. Burkitt
University of Melbourne
Melbourne, Australia
aburkitt@unimelb.edu.au

ABSTRACT

Music often contains many different auditory streams and one of its great interests is the relationship between these streams (melody vs. counterpoint vs. harmony). As these different streams reach our ears at the same time, it is up to the auditory system to segregate them. Auditory stream segregation is based mainly on our ability to group different streams according to their overall auditory perceptual differences (such as pitch or timbre). People with impaired hearing have great difficulty separating auditory streams, including those in music. It has been suggested that attention can influence auditory streaming and, by extension, visual information. A psychoacoustics experiment was run on eight musically trained listeners to test whether visual cues could influence the segregation of a 4-note repeating melody from interleaved pseudo-random notes. Results showed that the amount of overall segregation was significantly improved when visual information related to the 4-note melody is provided. This result suggests that music perception for people with impaired hearing could be enhanced using appropriate visual information.

1. INTRODUCTION

While listening to music, listeners with normal hearing are generally able to hear the melody played by each instrument separately. This ability is commonly know as auditory streaming and refers to the process by which the human auditory system organises sounds from different sources into perceptually meaningful elements [1]. This ability is crucial to appreciate music, as its main interest relies on the relationship between voices (melody vs. counterpoint vs. harmony).

In a typical auditory streaming experiment [2], listeners are exposed to a sequence of alternating high and low notes – the sounds may be grouped together and perceived as coming from a single source (termed fusion) or perceived as streams from separate sources (termed fission). In the fusion case, the single stream is perceived as a "gallop". In the fission case, the sequence is perceived as two separate streams, or as a "Morse code" (Figure **1**).

Any salient perceptual changes, such as pitch and timbre [3], can influence auditory streaming [4]. If the same melody is played simultaneously by two different instruments, the perceptual segregation of the two auditory streams will be based mainly on their timbre differences. If the same instrument, such as a piano, plays two melodies simultaneously, the timbre of each melody is relatively similar [5]; therefore, the perceptual segregation is based mainly on the pitch difference between the two melodies.

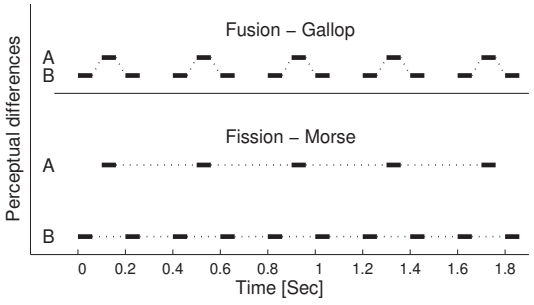

Figure 1. Illustration of typical auditory streaming experiment [Upper panel] The stimuli A are B perceptually close: the sequence is perceived as fused. [Lower panel] The stimuli A are B perceptually dissimilar: the sequence is perceived as two streams.

The mechanism of auditory streaming and the parts of the auditory pathway involved are still unclear. However, some evidence has shown peripheral cochlear [6] and central cortical components [2]. The "Peripheral Channelling" theory suggests that streaming depends primarily on the amount of overlap in the excitation pattern on the basilar membrane induced by the two stimuli, the more the two stimulus excitation patterns overlap, the more likely they are to be perceived as a single stream. On the other hand, streaming can be strongly affected by attention. Carlyon [2] showed that the minimum duration for two streams to be perceived as segregated depends on the amount of time the listener has paid attention to the sequence. As auditory attention can be affected by vision [7], this result suggests that a visual cue can have an influence on auditory streaming.

In normal hearing adults, coincident auditory and visual signals from the same object or event are combined in a process commonly known as multisensory integration. Multisensory integration can alter perception [8, 9] and facilitates information processing [10]. In adults with good hearing, stream segregation is also improved by multisensory integration. For example, studies have shown that the ability to segregate auditory speech from background noise is facilitated when visual cues (lip-movements) are available [11]. More recently, visual sequences have been shown to influence stream segregation of two ambiguous auditory sequences without linguistic components [12]. However, if incongruent information is presented to each sensory modality, the senses can also interfere with each other, causing altered or illusory percepts [8, 13].

Two hypotheses can be formulated:

1] A visual cue can enhance segregation of two streams by focusing the attention of the listener on one specific stream.

2] A visual cue will interfere with the auditory process and, therefore, weaken the segregation ability.

One of the most common complaints of people with hearing impairment, and especially cochlear implant recipients, is that music perception is very poor. A hearing impairment not only reduces the loudness of sounds, it also decreases the perceptual differences between sounds. This affects the ability of people with hearing impairment to segregate auditory streams and, therefore, weakens their appreciation of music. If hypothesis 1] is supported, this will indicate that visual support might be beneficial to improve music perception in people with hearing impairment.

The following experiment has been designed to extract baseline data for a study dedicated to better understanding auditory streaming ability in people with hearing impairment, and whether a visual cue may help them to improve their appreciation of music. The first step of this study is to test whether hypothesis 1] or 2] is supported for people with normal hearing and musical training.

2. EXPERIMENT

2.1. Methods

2.1.1. Stimuli

The auditory stimuli were all pure tones with a duration of 180 ms including 10 ms rise and fall raised cosine windows. The frequency of each tone ranged from midinote 45 (A2 or 110 Hz) to midinote 72 (C5 or 523.25 Hz). All tones were equalized in loudness at 70 phons according to the model of Moore [14]. A delay of 20 ms was introduced between each note. Stimuli were generated using Matlab 7.5 at a sampling rate of 44.1 kHz. The melody was composed of a repeating 4-note sequence (the target) interleaved with pseudo-random notes (the masker). The four notes of the target melody were middle C, F, D, and G (midinotes 60, 65, 62, 67). Figure **2(A)** shows the melody. Between each note of the target, a masker note was introduced, selected randomly within a range of 10 semi-tones. The note range of the masker varied across each block. Figure **2(B-D)** shows three possible sequences of notes depending on the note range of the masker. Sounds were presented through AKG K601 headphones at a comfortable level.

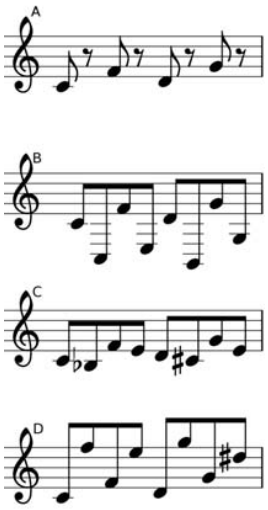

Figure 2. Musical sequence presented. The measure A] represents the target 4-note melody; B] shows a possible sequence when the note range of the masker is well below the target; C] a possible sequence when the note range of the masker overlaps the target; D] when the masker is well above the target.

Visual stimuli were presented in synchrony with the target and consisted of a piano staff and a quarter note. Each time a note from the target was played, it appeared as a quarter-note on the staff and disappeared when the note was turned off. The visual cues were presented on an LCD monitor and were generated through the software MAX/MSP 5.0. Figure **3** shows a screen shot of the visual cue.

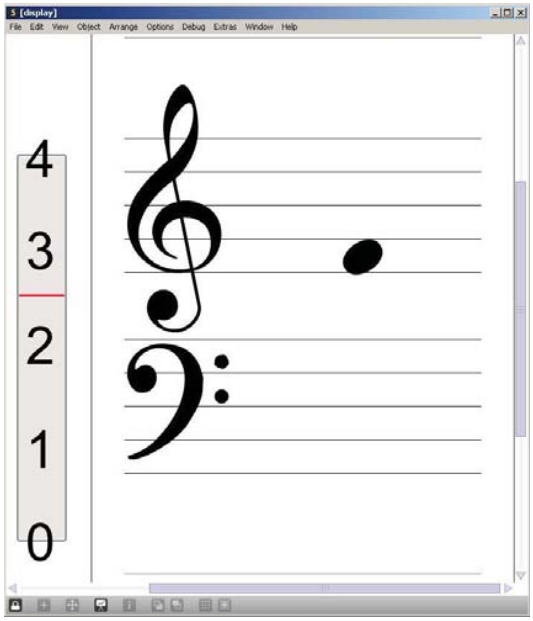

Figure 3: Screen shot of the visual cue. On the left the rating scale and on the right the staff with note indicated.

2.1.2. Procedure

The experiment was divided into four blocks. As the experiment was designed to minimize the duration of testing, no training was provided. In order to reduce any possible bias due to training effects, the order of the four blocks was randomized across listeners. At the beginning of one block, the notes of the masker ranged between midinotes 45 (A2) to 54 (G3). After 12 seconds (60 notes), the note range of the masker increased by one semi-tone (from midinotes 46-55). The block stopped when the note range reached midinotes 72-81 (after 5.6 minutes). In another block, the masker notes range started at midinote 72-81, and then decreased at the same rate as the previous block, and stopped at midinotes 45-54 . These two blocks were each performed once with a visual cue, and once without. The overall experiment lasted less than half an hour.

The listeners were asked to focus their attention only on the target and to indicate their response by moving in real time the cursor of a midi controller. The response was a rating of how well the target melody was perceived. The cursor was graded from 0 to 4. A score of 4 indicated that the melody was easily perceived and a score of 0 indicated that the melody was not perceived at all. If the listeners perceived only half of the melody, or if they perceived the melody every other presentation, the listener was instructed to move the cursor to the position 2. Other positions of the cursor indicated any other intermediate percepts respectively. The midi controller was linked to a MAX/MSP patch that managed the experiment.

When a visual cue was presented, the listener was asked to keep their visual attention on the screen.

2.1.3. Listeners

Eight listeners participated, with ages ranging from 27 to 40 years of age. They all had some musical training and were able to read music. No one was paid for their participation.

2.2. Results

Figure **4** shows the results averaged across listeners and direction of the note range of the masker (from low to high register, or high to low). The data have been normalized to represent the probability of the fusion of the target and the masker into a single non-repeating melody (*i.e.* a score of 4 on the cursor was mapped to a value of 0 on the y-axis, and a score of 0 on the cursor to 1 on the y-axis). The x-axis represents the centre midinote range of the masker for any given range. The thin lines represent the raw data averaged over the eight listeners and the thick line a Gaussian-like fit:

$$Y_e = a * e^{-0.5*(\frac{b-N}{c})^2}, \quad (1)$$

where Y_e is the estimation of the model, N is the midinote, and a, b, and c are the free parameters of the model. This model was fit separately to the data with and without visual cues. The parameter a indicates the peak value of the model on the normalized scale, the parameter b is the location of the peak, and the parameter c is the bandwidth of the model. In order to assess the streaming ability as a function of the note range overlap between the masker and the target, an indicator, P50, was derived from both models. It represents the note range where the model was above a probability of 50% of fusion.

The overall shape of the functions with and without visual cues indicates as expected, that when the masker note range is well separated from the target, the two streams are clearly segregated, and when they overlap, the two streams are fused.

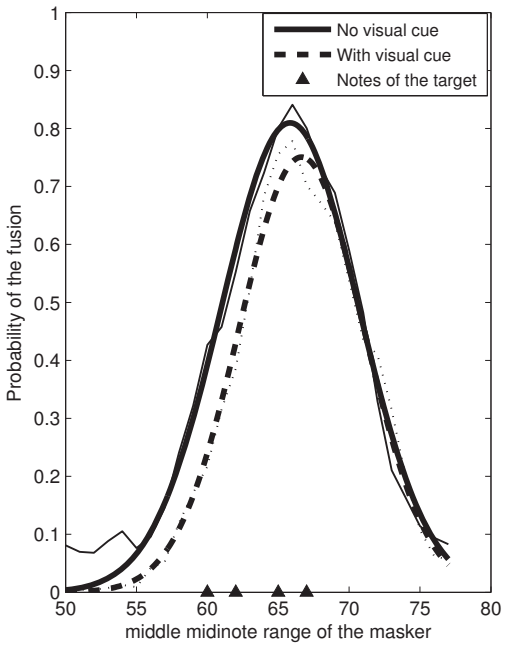

Figure 4: Average results (thin line) and Gaussian-like fit (thick line) of the results averaged across listeners and masker direction.

Student's t-tests were performed to compare the likelihood of fusion between the conditions (with and without visual cues). The analyses revealed that:

Parameter *a*]: The difference of amplitude observed in figure **4** between the conditions was not significant ($t(7) = 1.17$, $p = 0.1397$). This indicates that at the point of maximum difficulty, visual cues did not contribute significantly to auditory streaming.

Parameter *b*]: The average location of the peak was shifted to the right (higher note range) for the condition with visual cues. T-tests revealed a trend ($t(7) = 1.53$, $p = 0.0850$) and 7 out of 8 eight listeners showed a shifted peak. Thus, the visual cues showed a greater effect when the masker range was lower than the target range than when it was higher.

Parameter P50]: This parameter was highly significantly smaller for the condition with the visual cues ($t(7) = 3.02$, $p = 0.0097$), indicating that visual cue can allow a bigger note overlap between the target and the masker while still being perceived as streamed.

Overall, the results suggest that visual cues had a significant effect on auditory streaming.

3. DISCUSIONS

3.1. The effect of visual cues on auditory streaming

The results from this experiment clearly indicate that visual cues can have an effect on auditory streaming. Listeners, when presented with visual cues, could segregate a melody from an interleaved masker with a larger note range overlap than without visual cues.

This result strongly supports hypothesis 1] and refutes hypothesis 2]. This indicates that visual information might help people with impaired hearing to restore some of their appreciation of music.

3.2. The asymmetry of the effect.

The parameter *b* indicates that the effect of the visual cue was greater when the two lower notes were masked. Figure **4** shows no effect of the visual cues when the two higher note were masked. As the melody was tonal, and in a C key, the C note of the melody appeared clearly as the down beat. Listeners reported that if the C note was perceived, it was easier to extract the rest of the melody. The results suggest that the visual cues should emphasise the down beat. Further research is needed to validate this hypothesis.

3.3. The mechanism of auditory streaming

In conditions with and without visual cues, the auditory information was the same. Therefore, if streaming depends primarily on the excitation pattern on the basilar membrane, as suggested by the "Peripheral Channelling" theory, both conditions should lead to the same results. However, our data showed an effect on streaming of attention driven by visual cues. This result, therefore, supports the theory that streaming is at least partially supported by some cortical components.

3.4. The effect of musical training

All the listeners that participated had some musical training. It might be possible that the results could be different with people with normal hearing and no musical training. However, if this is the case, it implies that training could improve the weight of visual information in auditory segregation. Further testing is needed to validate this hypothesis.

4. ACKNOWLEDGMENT

The authors wish to thanks Ayla Barutchu and Prof. Peter Blamey for helpful comments on the early version of the paper and Aimee Clague for proofreading the document. Financial support was provided by the Jack Brockhoff Foundation, Goldman Sachs JBWere Foundation, Soma Health Pty Ltd, Mr Robert Albert AO RFD RD, Miss Betty Amsden OAM, Bruce Parncutt & Robin Campbell, and The Frederick and Winnifred Grassick Memorial Fund. The Bionic Ear Institute acknowledges the support it receives from the Victorian Government through its Operational Infrastructure Support Program.

REFERENCES

[1]. Bregman, A. S., *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: The MIT Press, 1994.

[2]. Carlyon, R. P., "How the brain separates sounds". *Trends Cogn Sci*. 8: p. 465-71, 2004.

[3]. Wessel, D., "Timbre Space as a Musical Control Structure". *Comput Music J*. 3: p. 45-52, 1979.

[4]. Moore, B. C. and Gockel, H., "Factors Influencing Sequential Stream Segregation". *Act. Acustica*. 88: p. 320 – 332, 2002.

[5]. Marozeau, J., de Cheveigne, A., McAdams, S. and Winsberg, S., "The dependency of timbre on fundamental frequency". *J Acoust Soc Am*. 114: p. 2946-57, 2003.

[6]. Hartmann, W. M. and Johnson, D., "Stream segregation and peripheral channeling". *Music Percept*. 9: p. 155-184, 1991.

[7]. Driver, J. and Spence, C., "Crossmodal attention." *Curr Opin Neurobiol*. 8: p. 245-253, 1998.

[8]. McGurk, H. and MacDonald, J., "Hearing lips and seeing voices". *Nature*. 264: p. 746-8, 1976.

[9]. Barutchu, A., Crewther, G. S., Kiely, P., Murphy, M. J. and Crewther, D. P., "When /b/ill with /g/ill become /d/ill: Evidence for a lexical effect in audiovisual speech perception". *Eur J Cogn Psychol*. 20: p. 1-11, 2008.

[10]. Miller, J., "Divided attention: evidence for coactivation with redundant signals". *Cogn Psychol*. 14: p. 247-79, 1982.

[11]. Sumby, W. H. and Pollack, I., "Visual contribution to speech intelligibility in noise." *J Acoust Soc Am*. 26: p. 212-215, 1956.

[12]. Rahne, T., Bockmann, M., von Specht, H. and Sussman, E. S., "Visual cues can modulate integration and segregation of objects in auditory scene analysis". *Brain Res*. 1144: p. 127-35, 2007.

[13]. Shams, L., Kamitani, Y. and Shimojo, S., "Visual illusion induced by sound". *Brain Res Cogn Brain Res*. 14: p. 147-152, 2002.

[14]. American National Standard. "Procedure for the Computation of Loudness of Steady Sounds" *ANSI S3.4-2007*, 2007.