Published as:

Marozeau, J. M., Innes-Brown, H., & Blamey, P. J. (2013). The effect of timbre and loudness on melody segregation. *Music Perception, 30*(3), 259-274.

http://dx.doi.org/10.1525/mp.2012.30.3.259

The Effect of Timbre and Loudness on Melody Segregation

Jeremy Marozeau, Hamish Innes-Brown, & Peter J. Blamey
The Bionics Institute, Melbourne, Australia

The aim of this study was to examine the effects of three acoustic parameters on the difficulty of segregating a simple 4-note melody from a background of interleaved distractor notes. Melody segregation difficulty ratings were recorded while three acoustic parameters of the distractor notes were varied separately: intensity, temporal envelope, and spectral envelope. Statistical analyses revealed a significant effect of music training on difficulty rating judgments. For participants with music training, loudness was the most efficient perceptual cue, and no difference was found between the dimensions of timbre influenced by temporal and spectral envelope. For the group of listeners with less music training, both loudness and spectral envelope were the most efficient cues. We speculate that the difference between musicians and nonmusicians may be due to differences in processing the stimuli: musicians may process harmonic sound sequences using brain networks specialized for music, whereas nonmusicians may use speech networks.

Received: April 12, 2011, accepted May 18, 2012.

Key words: auditory stream segregation, psychophysics, music training, cochlear implants, hearing impairment

Music is often composed of different melodic lines that are played together, either on the same or different instruments. These melodic lines, or *streams*, are often defined or separated by a number of perceptual parameters such as pitch, timbre, or loudness (reviewed by Bregman, 1990). One important aspect of listening to music is being able to hear these melodic lines separately and in relation to each other.

In a typical auditory streaming experiment (for a review of auditory streaming see Carlyon, 2004), listeners are presented with a rapid sequence of two pure tones differing in frequency (e.g., ABA-ABA). Depending on the frequency difference between the tones, different perceptual effects can occur (Figure 1). If the frequency difference is very small, the two tones can be perceived as one repeating tone. When the frequency difference is larger than the just-noticeable difference (JND) boundary, the listener will start to perceive the two tones as different. When the perceptual difference is still small, both tones are generally grouped into a single stream composed of two notes. Beyond a larger frequency difference, called the fission boundary (FB, van Noorden, 1975), listeners may start to hear the two tones as two independent streams of one repeating note (A-A-A-A and -B—B-). This perception is termed bistable (Pressnitzer & Hupe, 2006) because the perception will flip between one or two streams without an intermediate percept. Bistable perception can be influenced by top-down processes such as training, attention, or previous knowledge about the source (Bendixen, Denham, Gyimesi, & Winkler, 2010; Carlyon, Cusack, Foxton, & Robertson, 2001; Marozeau, Innes-Brown, Grayden, Burkitt, & Blamey, 2010). Finally, if the frequency separation is large enough, the listener will no longer hear the sequence as fused into a single stream, but will always hear it as segregated. This occurs at the temporal coherence boundary (TCB, van Noorden, 1975) and beyond. In addition to frequency, analogous boundaries can be defined for other acoustic differences between streams, such as intensity, or spectral centroid and F0 for complex tones.

Manipulation of the perception of auditory streams, via the FB and TCB, are tools composers use in order to convey musical intention. Fundamental frequency, and its perceptual correlate pitch, is the principal dimension used to encourage segregation between two melodic lines with similar timbre. Studies have shown that if the frequency separation between two melodies is greater than a semitone (i.e., no overlap), the melodies can be perceived as segregated (Dowling, 1973; Hartmann & Johnson, 1991; Marozeau, et al., 2010). However, other perceptual parameters can also be applied to create polyphony using a single instrument.

Figure 2 shows the first three bars of a keyboard suite: the Partita No.1 in *Bb*, Gigue, composed by J. S. Bach. This suite is composed of two interleaved voices. The first voice includes all the quarter notes and acts as

Music Perception, volume 30, issue 3, pp. 259–274, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2012.30.3.259

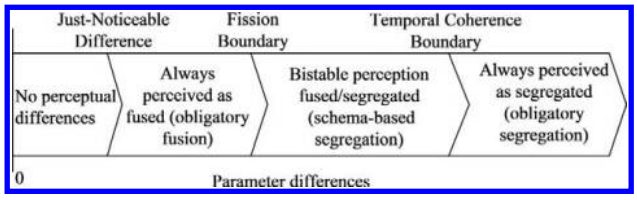

FIGURE 1. Summary of boundaries in auditory streaming.

the main melody, which can be divided into question phrases (beats 4 and 1) and answer phrases (beats 2 and 3). The second voice includes all the triplet notes and outlines the harmony. Based solely on fundamental frequency cues, the two voices will be segregated in bar 1 but not in bars 2 and 3. Therefore, other cues are required. First, the notes of the melody voice are approximately three times longer than the harmony voice. Second, an acoustic analysis of a performance of this work by Glenn Gould (Sony Classical) revealed a difference in the loudness level of between 5 to 10 phons (based on loudness model ANSI, 2007) between the two voices. Third, the same analysis revealed that as Gould applied more force to the keys while playing the melody voice, the timbre of the notes tended to be brighter. Based on a spectral centroid model (Marozeau & de Cheveigné, 2007; Marozeau, de Cheveigné, McAdams, & Winsberg, 2003), a difference of about 35% between the center frequencies of the spectral envelope of the two voices was observed once the effect of the fundamental frequency was removed.

These empirical data provide a first approximation of the physical differences needed to allow streaming, and show that acoustic differences beyond the fundamental frequency can affect the perceptual organization of sound sequences. In order to determine which acoustic cues can affect streaming, a variety of psychoacoustic studies have been performed.

It will be useful throughout to make a basic distinction between parameters of an acoustic signal, such as the intensity, temporal envelope, or the fundamental frequency (F0), and their related perceptual effects. The sensation of pitch, for example, is heavily influenced by F0, and the sensation of timbre depends on a combination of the temporal, spectral, and spectro-temporal envelopes. These terms will be elaborated further.

The Effect of Fundamental Frequency

Dowling (1973) asked participants to identify two interleaved familiar melodies. When the notes of both melodies completely overlapped in pitch range, participants were generally unable to identify them. As the amount of overlap decreased, participants began to identify the familiar melodies accurately. In a recent study by Marozeau et al. (2010), musicians and nonmusicians were asked to continuously rate the difficulty of segregating a simple four-note melody from a background of interleaved random distractor notes. The distractor notes were initially widely separated in pitch from the melody notes. As the separation between melody and distractor notes decreased, difficulty ratings generally increased. In order for nonmusicians to judge the 4-note melody as easy to perceive (less than 50% difficulty rating on a continuous scale), the melody and the distractor ranges had to be separated by at least one semitone. The musically trained participants, on the other hand, reported that they found it easy to segregate the melodies even with about 4 semi-tones overlap.

The Effect of Timbre

Timbre cues are also known to facilitate melody segregation between two different instruments playing in overlapping note ranges, as in a duo for piano and violin (Wessel, 1979). Timbre is a multidimensional perceptual quality that can be decomposed into spectral, temporal, and spectro-temporal dimensions (McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995). The effect of each dimension on stream segregation has been studied separately (for a review see Bregman, 1990) or collectively (Bey & McAdams, 2003; Iverson, 1995; Wessel, 1979). Bey and McAdams (2003) have shown that when a target melody is interleaved with a distractor melody, the ability to identify the target melody increases with the timbral distance between the

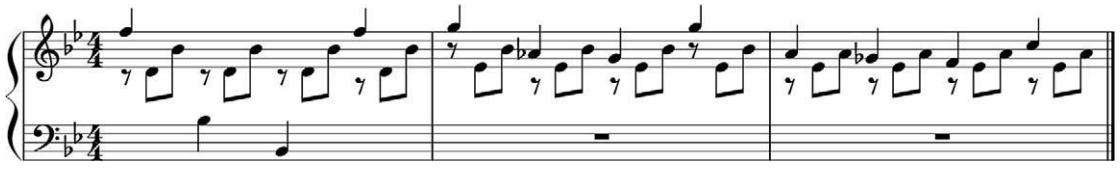

FIGURE 2. The first 3 bars of the Partita No.1 in *Bb*, Gigue, by J.S. Bach.

melodies. Although most studies show that the spectral dimension can affect segregation (Cusack & Roberts, 2000; Hartmann & Johnson, 1991; Iverson, 1995; Wessel, 1979), there is still some controversy about the temporal dimension (Hartmann & Johnson, 1991; Iverson, 1995; Wessel, 1979).

Hartmann and Johnson (1991) studied the influence of different acoustic parameters on melody segregation. Their results showed no significant contribution to stream segregation of the temporal envelope, but a significant effect of the stimulus duration was found. Iverson (1995) adapted the similarity task traditionally used to study timbre to stream segregation. His results suggested that auditory stream segregation is influenced by differences in static spectra and by dynamic attributes, including attack duration and spectro-temporal envelope. Singh and Bregman (1997), however, provided clear evidence of the effect of temporal envelope on stream segregation. The F0 difference required to allow segregation of complex ABA tones was significantly lower when there were differences in the rise and fall times between the two subsets of tones than when the temporal envelope was the same. Overall, results were mixed, possibly due to the overall ranges over which the stimuli were manipulated in each study. It is therefore necessary to test a wider range of levels, on multiple temporal features.

The spectral dimension of timbre can also influence stream segregation. In an ABAB type experiment, van Noorden (1975) showed that sounds with the same pitch but a different spectral envelope could allow stream segregation. This result has been confirmed by many other studies (Cusack & Roberts, 1999, 2000; Hartmann & Johnson, 1991; Iverson, 1995; Wessel, 1979). For example, when a melody consisting of complex sounds with approximately 2 dB attenuation per harmonic and a cut-off frequency around 5 kHz was interleaved with another melody played with tones that were pure sine waves, participants were able to identify two interleaved melodies 80% of the time (Hartmann & Johnson, 1991).

The Effect of Intensity

van Noorden (1975) measured the fission boundary of the intensity cue between two pure tones at the same frequency using an ABAB paradigm. Two normal hearing listeners were asked to adjust the intensity of tone A, interleaved with tone B presented at the same frequency as A and at a fixed level of 35 dB SL, in such a way that B tones could be just heard as segregated from A tones. When both tones were presented with a tone repetition time of 200 ms, an average of 4 dB difference was needed in order to hear the B tones in a separate stream. Hartmann and Johnson (1991) asked seven listeners to identify two well-known interleaved melodies. All the melodies were composed of eighth notes only. If the perceived difference in loudness between the melodies was more than 8 dB, both melodies were identified with an average accuracy of 70%. Thus, experiments on the effect of loudness cues on auditory streaming generally indicate that differences in intensity of between 4-8 dB between the A and B tones can allow them to be perceptually segregated. The overall level of both the A and B tones, on the other hand, does not seem to have an effect (Beauvois & McAdams, 1996).

Thus, fundamental frequency, intensity, and temporal and spectral envelopes can all affect stream segregation. However, only a few studies (Hartmann & Johnson, 1991) have been performed with the same listeners in all conditions, and thus it is difficult to compare the relative effectiveness of each cue. The purpose of the current study was to determine which of these streaming cues was the most efficient. In other words, to determine which cue required the least *perceptual* change in order to allow streaming.

The Effect of Music Training

The decay of streaming effects occurs more slowly in musicians compared to nonmusicians (Beauvois & Meddis, 1997), and in conditions with reduced spectral complexity, musicians can separate streams of notes that are closer in pitch than nonmusicians (Vliegen & Oxenham, 1999). Marozeau et al. (2010) found that in the absence of a visual score, musicians reported significantly less difficulty than nonmusicians segregating a repeating melody from interleaved distractor notes. Listeners with music training are also better able to separate concurrently presented sounds (Zendel & Alain, 2009). These authors found evidence from EEG recordings made during the task that the musicians' improvement in detecting mistuned harmonics was due to changes in early perceptual processing in addition to higher level cognitive processes.

One of our current goals was to study auditory streaming ability for musically valid stimuli and melody. Therefore, the musical background of each listener was taken into consideration to test its possible effect.

In order to determine which acoustic streaming cues are the most efficient, two experiments were performed. In Experiment 1, stream segregation difficulty ratings were determined separately for three acoustic cues: intensity, temporal envelope, and spectral envelope.

This experiment was performed using the same methods and participants as a previous experiment that focused on the effect of fundamental frequency (Marozeau, et al., 2010). In Experiment 2, the amount of perceptual change that was induced by each cue was measured by varying the three physical parameters in a dissimilarity rating paradigm. Finally, by combining the results from both experiments, it was possible to determine the relative perceptual change required for each cue to affect the streaming difficulty ratings.

Experiment 1

METHOD

Participants. Thirty-seven adults participated in this study. On the same day, each listener also participated in a complementary study on the effect of visual cues on auditory streaming ability (Marozeau et al., 2010). Listeners were divided into two groups according to their music training.

Musicians. The first group was composed of 18 adults (8 male, 10 female) with normal hearing and music training. Normal hearing was defined as audiometric thresholds less than 20 dB HL (American National Standard Institute, 1996) at octave frequencies from 250 to 8000 Hz. Each listener was categorized using a hierarchical cluster analysis designed to maximize the differences between groups of musicians and nonmusicians on four normalized musical activity variables: (a) sight-reading ability self-ratings, (b) general musical aptitude self-ratings, (c) the number of hours of musical practice per week, and (d) years of music training (for more information see Marozeau et al., 2010). Average age was 31 years with a standard deviation of 7.2 years.

Nonmusicians. The second group was composed of 19 adults (9 male, 10 female) with normal hearing and minimal music training. Average age was 32.2 years with a standard deviation of 7.9 years.

No one was paid for their participation but travel and lunch expenses were reimbursed.

STIMULI

Two types of sequences were used: the target was a repeating melody and the distractor consisted of pseudorandom notes. The two sequences were termed the *target melody* and the *distractor*. The two sequences were presented interleaved to the participants (see Figure 3). The target melody was a 4-note repeating melody with the following physical parameters:

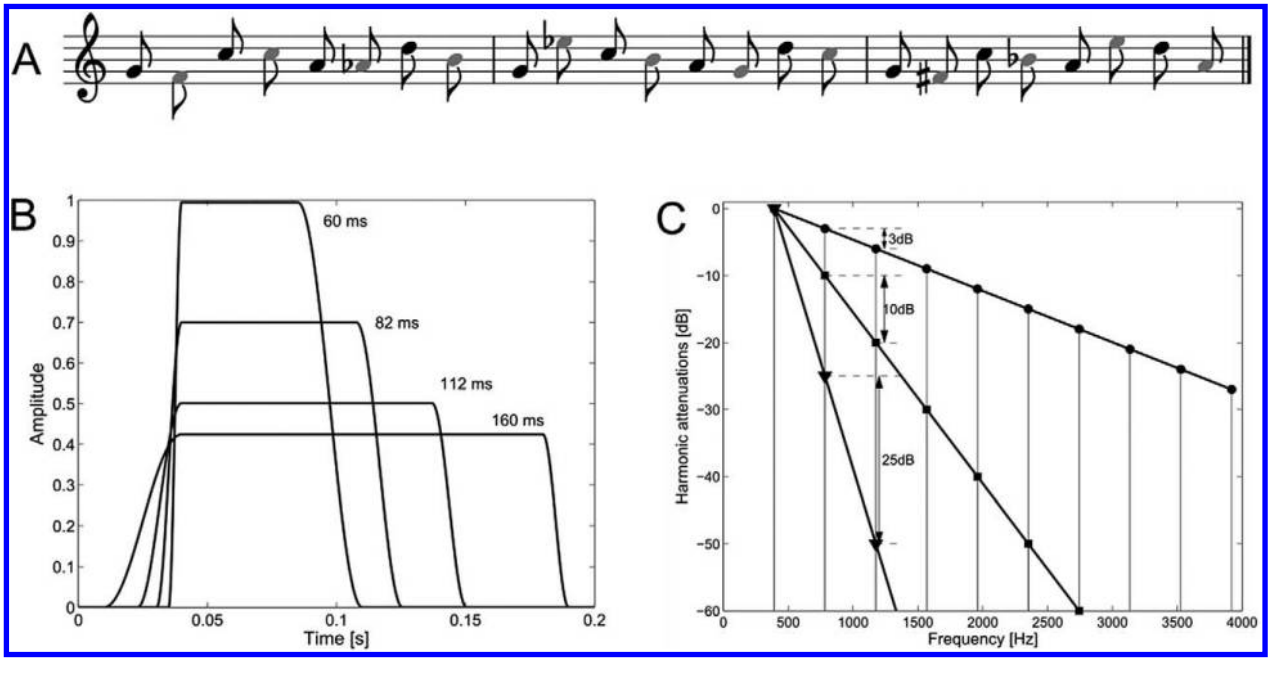

FIGURE 3. A) Example of note sequences. The repeated target melody notes are in black and the interleaved random distractor notes are in light grey. B) Different temporal envelopes varying in impulsiveness from 60 ms FDHM to 160 ms FDHM (the target notes). Only 4 of the 20 possible temporal envelopes are shown here. C) Different spectral envelopes varying from 3 dB of attenuation per harmonic (the target notes - circle) to 25 dB of attenuation per harmonic (triangles). Only 3 out of the 20 possible spectral envelopes are shown here.

1. The F0 frequencies of the four target melody notes were: 392, 523, 440, and 587 Hz. These frequencies correspond to G, C, A, and D in musical notation or to 67, 72, 69, and 74 in midinote notation (black notes in Figure 3A). The target melody was composed of intervals large enough to be perceived by many people with poor pitch discrimination[1] while being small enough for the notes to be grouped into a single melody (instead of two interleaved melodies composed of the two low notes and the two high notes).
2. The temporal envelope of each note was composed of a 30 ms raised-cosine onset, 140 ms sustain, and 10 ms offset, for a total duration of 180 ms or an impulsiveness of 160 ms, defined as the full duration of the sound at half of the maximum amplitude, FDHM (see Figure 3B).
3. The spectral envelope of each note consisted of 10 harmonics, successively attenuated by 3 dB (see Figure 3C).
4. The intensity of each note was adjusted in order to reach 65 phons (i.e., as loud as a 1 kHz tone at 65 dB SPL) according to loudness models (American National Standard Institute, 2007; Glasberg & Moore, 2002).

The distractor notes consisted of uniformly randomized notes selected from a range of an octave overlapping the target melody (349 to 659 Hz, 65 to 76 midinote). The distractor notes were varied on one parameter at a time in each of three different conditions, while the other parameters were kept constant. The parameters were varied gradually within each condition from the same level as the target melody (level 0) to a level that was designed to allow easy segregation (level 19). The three conditions were:

1. *Temporal envelope*. Twenty levels of distractor impulsiveness were tested, logarithmically spaced between 160 and 60 ms FDHM (see Figure 3B). The attack, sustain, and release times of the envelope were varied logarithmically in order that the most impulsive stimulus had the sharpest attack and the longest release, and the least impulsive stimulus had the slowest attack and the shortest release. As the overall duration of the stimulus was shortened, its overall amplitude was increased in order to keep its loudness level constant at 65 phons (ANSI, 2007; Glasberg & Moore, 2002).
2. *Spectral envelope*. Twenty levels of attenuation were tested, logarithmically spaced between 3 and 25 dB attenuation per harmonic (see Figure 3C).
3. *Intensity*. The amplitude of each distractor note was varied in twenty 2-phon steps in order to set the loudness level from 65 to 27 phons.

The intensity of each stimulus was individually adjusted using a loudness model (American National Standard Institute, 2007; Glasberg & Moore, 2002) to ensure that the sensation of loudness was constant as the spectral envelope and temporal envelope were varied. The stimuli were constructed using Matlab 7.5 and sequences were generated using MAX/MSP 5. Sounds were played through an M-AUDIO ProFire 610 48-kHz 24-bit sound card and loudspeaker (Genelec 8020A, Iisalmi, Finland, selected for its flat frequency response) positioned on a stand at the listener's ear height, 1 m from the listener's head. The experimental protocol was approved by the Human Research Ethics Committee of the Royal Victorian Eye & Ear Hospital. Written, informed consent was obtained from all participants prior to their participation in the study.

PROCEDURE

The experiment consisted of blocks of trials in which the target melody was repeated continuously, interleaved with distractor notes. In each block, all twenty levels of the distractor note parameter were tested. The distractor notes either started with the parameter at the highest level (i.e., the least similarity between target melody and distractor) and decreased, called the DEC block, or began at the lowest level (i.e., both target melody and distractor shared the same physical parameters) and increased, called the INC block. At the start of DEC blocks, the target melody and distractor notes were likely to be perceived in separate streams, and the target melody was thus easily perceived, whereas at the start of INC blocks, both target melody and distractor were likely to fuse into a single stream and the target melody was not easily perceived. After 10 presentations of the target melody (16 s), the parameter level of the distractor was either increased (INC block) or decreased (DEC block) to the next level. The block ended when the parameter level reached either level 19 (in INC blocks) or level 0 (in DEC blocks). A paradigm where the parameter level was gradually changed was preferred over a completely randomized design in order to avoid resetting the "buildup effect" randomly, which would occur if the parameter level were varied significantly from one trial to the next.

A DEC block was always run first as a practice session, with the data from this block discarded. Following the practice session, INC and DEC blocks were run

[1] This experiment was designed to be replicated by listeners with impaired hearing.

twice each, with A-B-B-A/B-A-A-B order counterbalanced across participants.

During presentation of INC and DEC blocks, the participants continuously rated the difficulty of perceiving the target melody using a variable slider on a midi controller (EDIROL U33, Roland Systems Group, Dee Why, Australia). The slider was labeled from 0 (no difficulty hearing the melody) to 10 (impossible to hear the melody). Participants were instructed to move the slider to the "10" position if the target melody was impossible to perceive, and to the "0" position if the target melody could be easily perceived. Every time a note of the target melody was played, the slider position was encoded in 127 steps on a personal computer running MAX/MSP 5. In order to reduce possible pitch memory effects between sessions, a pitch increment, randomly chosen between 0 and 4 semitones, was added to all notes of the same session. The rating procedure used in this experiment, which involves a subjective measure of streaming, has been previously validated using a control experiment based on an objective detection task. The results showed that listeners who judged the melody to be difficult to perceive overall had the lowest accuracy scores in the detection task. Conversely, those who judged the melody as easy to perceive overall had the highest accuracy scores. Details and results of the control experiment can be found in Marozeau et al (2010).

Results and Discussion

OVERALL ANALYSIS

For each listener and each condition, the continuous ratings were divided by 10 and then averaged within each of the 20 levels. Therefore, the response took the form of 4 vectors of 20 elements: 20 levels, 2 directions (INC and DEC), and 2 repetitions. For each of the three acoustic parameter conditions, a mixed, repeated-measures general linear model (GLM) was conducted on the difficulty ratings. Level was considered as a continuous variable ($df = 1$), direction ($df = 1$) and repetition ($df = 1$) were considered as categorical within-subjects factors. Group ($df = 1$) was considered as a categorical between-subjects factor. The difficulty rating data were expected to follow a psychometric function with level that could be modeled as: rating = 1/(1 + exp(a*level + b)). Therefore, the difficulty ratings were transformed to: rating' = log(1/rating -1) in order to produce a linear relationship between the transformed rating' and the level. Ratings were bounded to .01 and .99 prior to transformation in order to avoid infinite values. Effects were considered significant for p values $< .05$. The effect size was evaluated with R^2, estimated as the F ratio between the sum of squares of a factor and the overall sum of squares. Statistical analyses were performed using the software R 2.11.0 (R Foundation for Statistical Computing, Vienna, Austria). The results from each condition will be discussed in turn, with a short commentary on how they relate to the results of Hartmann and Johnson (1991).

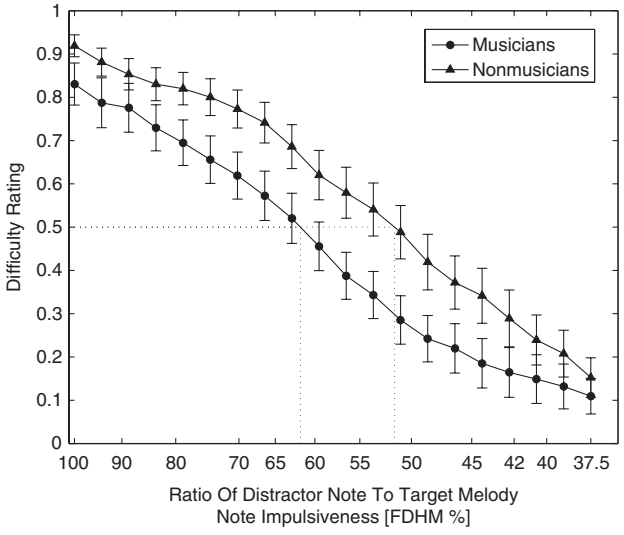

FIGURE 4. Results of Experiment 1 for the temporal envelope conditions. The abscissa represents the ratio of impulsiveness between the target melody and the distractor notes. Dotted lines indicate the impulsiveness ratio required for listeners to report a difficulty level of .5. The error bars show the standard error of the mean. The impulsiveness was defined as the full duration of the sound at half of the maximum amplitude (FDHM).

Temporal Envelope. The difficulty ratings for the temporal envelope condition were averaged across direction, repetition, and group, and are shown in Figure 4 (circles represent the average results for musicians and triangles for nonmusicians). The abscissa shows the ratio of impulsiveness between the target melody and the distractor. A ratio of 100% indicates that both the target melody notes and the distractor notes share the same temporal envelope. A ratio of 50% indicates that the distractor notes had an impulsiveness half as long as the target melody notes (i.e., 80ms FDHM). The ordinate represents the averaged difficulty rating. It was assumed that when listeners reported that the target melody was difficult to perceive, it had fused into a single stream with the distractor, and when listeners reported low difficulty in perceiving the melody, it was

perceived as segregated from the distractor. A difficulty rating of .5 was used as an estimation of the TCB[2] - the boundary where listeners did not always hear the melody as segregated (Bregman, 1990).

Figure 4 shows that as the ratio between the impulsiveness values of the target and distractor melodies decreased (i.e., the distractor became less similar to the target melody), the listeners reported less difficulty in perceiving the target. When the distractor and the target melody shared the same impulsiveness (FDHM ratio of 100%), both groups reported that the melody was very difficult to perceive (.91 for nonmusicians and .83 for musicians), although at high difficulty levels, boundary effects may have influenced the result, leading to an underestimation of the true difficulty (the slider was limited to a maximum difficulty of 1). Overall, the musician group reported lower difficulty ratings at every level. The .5 difficulty rating point corresponded to an FDHM ratio of 62% for musicians. In other words, for a target melody with an impulsiveness of 160 ms, when the distractor notes shared the same spectral envelope, loudness level, overall pitch range, and had an impulsiveness longer than 62% of the melody (100 ms FDHM), the distractor notes were rated as difficult to distinguish from the target melody. For nonmusicians an FDHM ratio of 52% (distractor notes of about 80 ms FDHM) corresponded to the .5 difficulty rating.

In order to assess the significance of these effects, a mixed GLM with repeated measures was performed on the data from this condition. A main effect of group, $F(1, 35) = 4.41$, $p = .04$, $R^2 = .036$, was found, confirming that the difference observed between musicians and nonmusicians was statistically significant. As expected, a main effect of impulsiveness level was also found, $F(1, 35) = 182.14$, $p < .001$, $R^2 = .41$. A small but significant interaction was found between the direction and the level, $F(1, 35) = 5.23$, $p = .03$, $R^2 = .003$, with difficulty ratings generally lower in the DEC than INC blocks.

These results are in excellent agreement with Hartmann and Johnson (1991), who showed that listeners were able to correctly identify the names of two interleaved melodies with an average accuracy of 75% when there was an impulsiveness ratio difference between each melody of 30%. With an impulsiveness ratio of 37.5% (the lowest tested in the current study), listeners judged the melody as fairly easy to hear (average difficulty rating of .1 to .13).

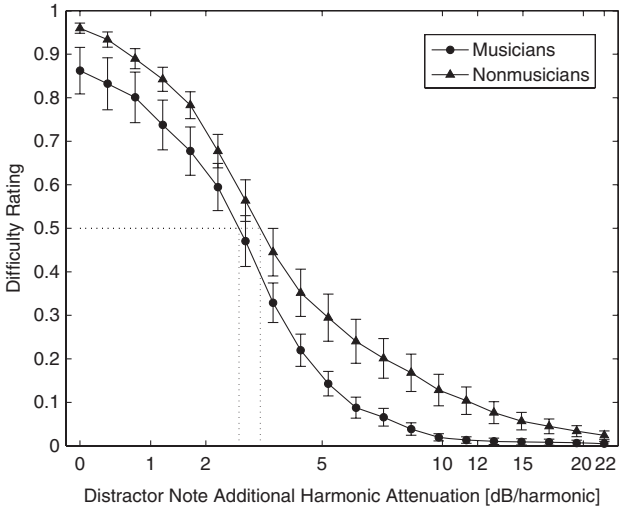

FIGURE 5. Difficulty ratings from the spectral envelope condition in Experiment 1. The abscissa represents the amount of additional attenuation between successive harmonics that were present in the distractor notes. Dotted lines indicate the amount of additional harmonic attenuation applied to the distractor notes for listeners to report a difficulty of .5. The error bars show the standard error of the mean.

Spectral Envelope. Average difficulty ratings from the spectral envelope condition are shown in Figure 5. The abscissa represents the amount of *additional* attenuation between successive harmonics that was present in the distractor notes. An additional attenuation of 0 dB thus indicates that both distractor and target melody notes had the same amount of attenuation per harmonic (3 dB per harmonic). An additional attenuation of 2 dB indicates that the distractor was composed of successive harmonics attenuated by 5 dB. As in the temporal envelope condition, the musician group showed lower average difficulty ratings at every level. However, the .5 difficulty rating point was fairly similar for the musicians and nonmusicians (2.7 and 3.2 dB additional attenuation, respectively) which corresponds to a difference in spectral centroid of 16% and 19% respectively. The GLM showed significant effects of group, $F(1, 35) = 6.48$, $p = .01$, $R^2 = .02$ and level, $F(1, 35) = 623.77$, $p < .001$, $R^2 = .66$ as well as a very small but significant repetition x level interaction, $F(1, 35) = 4.33$, $p = .04$, $R^2 < .0005$.

Although Hartmann and Johnson (1991) did not vary the amount of harmonic attenuation in their interleaved melodies, it is possible to calculate the difference in spectral centroid between the two melodies that were

[2] It is acknowledged that this measure of the acoustic difference between melody and distractor notes at the .5 difficulty rating level is not a true measure of the TCB, which would require a task involving the interruption of obligatory streaming. The point estimated by this procedure, however, is *equal to or below* the TCB.

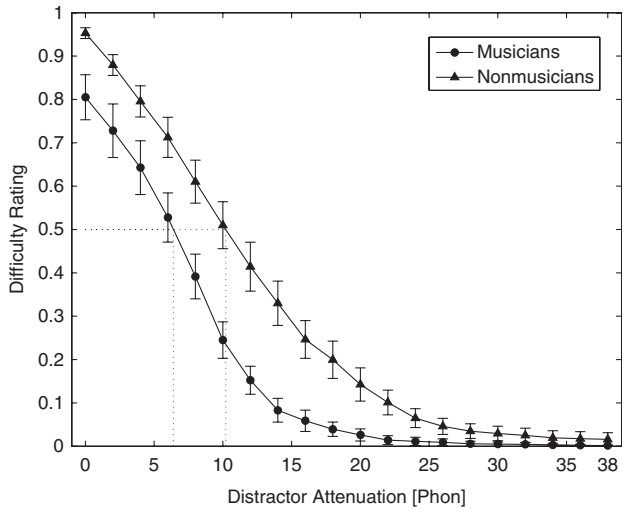

FIGURE 6. Difficulty ratings from the intensity condition in Experiment 1. The abscissa represents the difference in phons between the target melody and the distractor notes. The error bars show the standard error of the mean.

used in order to obtain an objective measure of the physical difference induced. Based on an implementation that considered the perceptual features of sounds (Marozeau & de Cheveigné, 2007), a difference in spectral centroid of 86% was found. In the results from the current study, this difference was tested when the distractor had an additional attenuation of approximately 22 dB per harmonic (the maximum level difference tested). In agreement with Hartmann and Johnson (1991), the participants in the current study reported no difficulty in perceiving the melody (almost 0 difficulty rating) at that level.

Intensity. Average difficulty ratings for the intensity condition are shown in Figure 6. The abscissa represents the difference in loudness level (phons) between the target melody and the distractor. A difference of zero means that both the target melody and distractor notes were presented at the same loudness level (65 phons). As in the previous conditions, the musicians rated less difficulty overall in perceiving the melody. The average .5 difficulty point was 6.4 phons for the musicians, and 10.2 phons for the nonmusicians. It is worth noting that 10-phon attenuation in the distractor melody corresponds to approximately half the loudness of the target melody. This threshold can be compared with the .5 difficulty point of the temporal envelope condition, where nonmusicians also required distractor notes of approximately half the impulsiveness of the target melody in order to segregate the streams.

The GLM revealed a significant main effect of group, $F(1, 35) = 13.06$, $p = .001$, $R^2 = .035$, and level, $F(1, 35) = 570.59$, $p < .001$, $R^2 = .601$. A significant group X level interaction, $F(1, 35) = 7.19$, $p = .01$, $R^2 = .008$ was also found; however, this may reflect a floor effect for additional attenuations beyond about 20 phons. There were also small but significant direction X level, $F(1, 35) = 5.93$, $p = .02$, $R^2 = .001$, and group X direction X level, $F(1, 35) = 6.79$, $p = .01$, $R^2 = .002$, interactions.

It could be argued that the reduction of loudness of the distractor notes will also result in a reduction of perceptual salience of the distractor, which could facilitate stream segregation in addition to any perceptual differences. Nevertheless, results are in good agreement with Hartmann and Johnson (1991) who found that listeners named the interleaved melodies with an accuracy of only 70% with a difference of 8 dB in intensity between the two melodies. This score was slightly above the baseline result composed of two melodies with the same loudness. Results from Experiment 1 show that both musicians and nonmusicians found it moderately difficult to perceive the melody when the loudness level difference was 8 phons (about 40% and 60% for musicians and nonmusicians respectively).

Experiment 1 determined the difference between the target melody and distractor notes, in terms of three acoustic parameters, required for listeners to segregate the target melody from distractor notes with a difficulty rating of half the maximum difficulty. When averaged across listener groups, differences of 8.3 phons in loudness level, 57% FHDM ratio in impulsiveness, or 2.95 dB/harmonic attenuation were required. Listeners with music training were found to require significantly less difference on each parameter compared to those with no music training.

Overall, the results from the current experiment are in very good agreement with the previous experiment of Hartmann and Johnson (1991), suggesting that both types of task (naming interleaved melodies, and rating the difficulty of extracting a melody from interleaved random distractor notes) are representative of the same underlying ability to segregate sound sources.

Experiment 2

RATIONALE

In order to compare the streaming effectiveness of each parameter, each of which was measured using different physical units, it is necessary to establish a common *perceptual* scale. Experiment 2 therefore employed a dissimilarity rating paradigm to determine the degree of perceptual change induced by changing each of the

acoustic parameters. This allowed for the construction of a common perceptual scale. Once this common scale was established, the difficulty ratings from Experiment 1 were replotted using this new perceptual scale.

Method

PARTICIPANTS

Thirteen listeners with normal hearing (seven females, six males) were tested (average age: 28 years, standard deviation 4.8 years). Three of them also participated in Experiment 1. Listeners were divided into musicians and nonmusicians as before. The nonmusician group was composed of seven listeners not currently practicing any musical activity, and the musician group consisted of six listeners with 3 to 28 hours practice per week. The first two authors were included among the participants.

STIMULI

Stimuli were created to be similar to those in Experiment 1, except that the three acoustic parameters could be modified at the same time.

Fifteen 4-note melodies were used in this experiment. As in Experiment 1, the notes were complex 10-harmonic sounds. The F0 of each note in the melody was the same as the target melody in Experiment 1. Melodies were presented in pairs, and the first melody in the pair differed from the second by 1, 2, or 3 physical parameters simultaneously (the loudness level, and the spectral and temporal envelopes). As the similarity results were analyzed using MDS techniques, it was important to ensure that 1) the differences in each parameter applied induced perceptual changes that were on the same magnitude scale, and 2) all the stimuli were evenly distributed in a three dimensional space (as in the stimuli of Caclin, McAdams, Smith, & Winsberg, 2005). Therefore, for each parameter, five possible levels, spanning approximately the upper half of each psychometric function found in Experiment 1 were selected. The stimuli were presented at a loudness level of either 65, 63, 61, 59 or 57 phons, with an additional attenuation per harmonic of either 1.69, 1.19, 0.75, 0.35 or 0 dB and a FHDM of either 100, 112, 126, 142, or 160 ms. The first five stimuli were constructed by assigning to them a random permutation of the five levels, for each of the three parameters. The procedure was repeated two more times, while ensuring that none of the 15 stimuli were identical. Figure 7 shows the 3-dimensional physical space of the 15 stimuli. The stimuli were constructed using Matlab 7.5 and the experiment was implemented using MAX/MSP 5. The stimuli were played from a loudspeaker positioned 1 m from the listener's head. Listeners responded using a 7 inch external USB touchscreen.

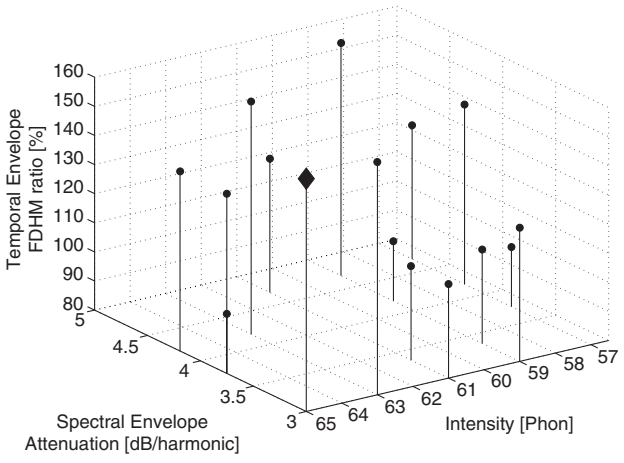

FIGURE 7. The three-dimensional physical space of the 15 stimuli in Experiment 2. Each circle represents one of the 15 stimuli used in Experiment 2. They are positioned in this 3D space according to their physical parameters. The diamond represents the physical attributes of the target melody of Experiment 1.

PROCEDURE

In the first part of the experiment, listeners were presented with each of the 15 stimuli in random order to acquaint them with the range of differences in the set of stimuli. They were allowed to hear them as many times as they wanted. In the second part, they were informed that the goal of the experiment was to estimate the similarity between pairs of melodies. Then they were presented with every possible pair of the 15 stimuli (in random order), totaling 105 pairs. A single pair was presented in each trial. In each trial, the participants were instructed to judge how similar the pairs were, and to respond by moving a cursor on a slider bar labeled from "most similar" to "least similar." Participants could listen to the pair as many times as they wanted, by pressing a "listen again" button. When they were satisfied with their judgment, they pressed a "validate" button, and the next trial began. The order within pairs and the order of pairs was random, and a different randomization was used for each session and subject. For each pair presented, the similarity response was stored in a matrix as a continuous value ranging from 0 (similar) to 1 (different).

RESULTS

In order to test whether music training had an effect two MDS bootstrap analyses (as developed in Bigand,

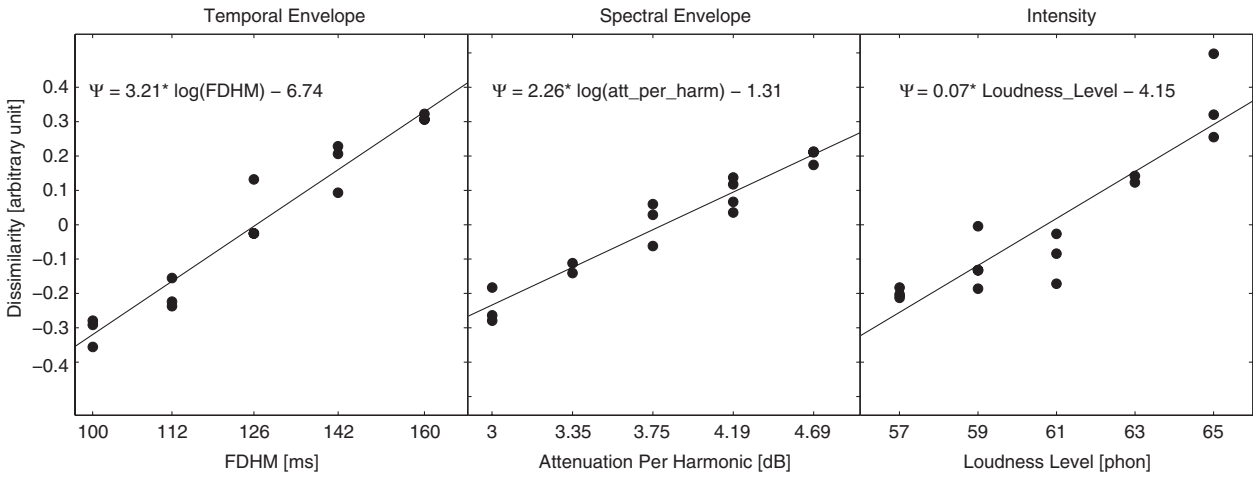

FIGURE 8. Scatterplot between the perceptual dimensions derived from the MDS analysis and physical parameters (left panel the temporal envelope, middle the spectral envelope, and right panel the loudness level). The equation in each panel describes the regression line.

Vieillard, Madurell, Marozeau, & Daquet, 2005) were performed on the musician and nonmusician groups separately. This technique allows statistical comparison between the two MDS solutions. First, a MDS space was created by randomly selecting participants, with replacement. Then another space was created with another random selection of participants. This technique was repeated 200 times. Each space was rotated towards the same target. The greater the variability between subjects, the greater the difference between the spaces. These 200 analyses combined defined a distribution of positions within the MDS space for each stimulus. It was therefore possible to define the 95% confidence volume for each stimulus.

When superimposing the bootstrap solution of the musician group with the nonmusician group, 95% confidence intervals for 14 out the 15 stimuli overlapped on the three dimensions. One stimulus overlapped only on the first two dimensions. This result indicates that the three acoustic cues had similar perceptual salience for musicians and nonmusicians. Their data were therefore combined for the rest of the experiment.

The similarity matrices were averaged across listeners. A three-dimensional space was then extracted using the MDSCAL procedure, implemented according to the SMACOFF algorithm (Borg & Groenen, 1997). As the MDSCAL solution is rotationally undetermined, the solution was rotated with a procrustean procedure that minimized the least-squares fit between the perceptual and physical spaces.

As expected, all three MDS dissimilarity dimensions were correlated with the physical dimensions (Figure 8): the first dimension correlated with the five levels of temporal envelope, $r(14) = .97$, $p < .001$; the second dimension with the five levels of spectral envelope, $r(14) = .97$, $p < .001$; and the third dimension with intensity, $r(14) = .91$, $p < .001$.

In order to determine the amount of *dissimilarity* difference between levels of the physical parameters, the slope of the regression line between each physical parameter and the MDS dimension was plotted (Figure 8). A slope of 28.93 (dissimilarity units per log10 of FDHM in ms) was found for the temporal envelope dimension, 20.39 (dissimilarity units per log10 of attenuation per harmonic) on the spectral envelope dimension and 0.62 (dissimilarity units per phon) on the loudness level dimension.

Based on these comparisons, it is possible to redraw Figures 4, 5, and 6 with a new scale on the x-axis. Figure 9 shows the difficulty ratings of all three conditions of Experiment 1 for the musician group on the same plot according to this new dissimilarity scale.

Figure 9 shows that for musicians, difficulty ratings decreased more quickly when the dissimilarity induced by the loudness level of the distractor notes was increased compared to both the temporal envelope and spectral envelope conditions. This indicates that musicians needed less perceptual change on the dimension of loudness level in order to allow stream segregation, than the two other dimensions. To assess the statistical significance of these differences, the difficulty ratings were averaged across the two repetitions, and a GLM was performed on the difficulty ratings for musicians and nonmusicians separately, with a categorical factor for the condition (temporal envelope, spectral envelope, loudness level) and a continuous factor for the level

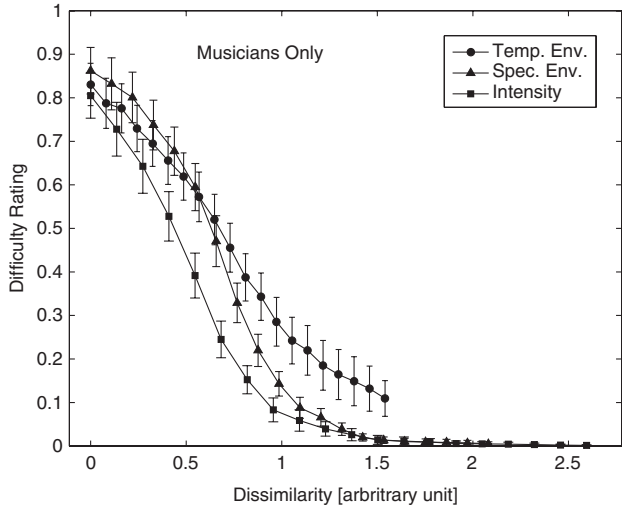

FIGURE 9. Difficulty ratings from all conditions in Experiment 1 for the musicians only. The figure replots the difficulty ratings from Experiment 1, but with a different abscissa based on a perceptual scale determined from dissimilarity ratings (Experiment 2). The error bars show the standard error of the mean.

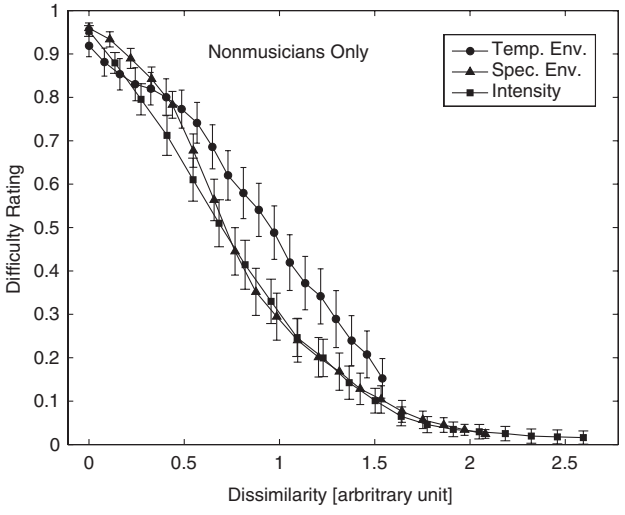

FIGURE 10. Difficulty ratings from all conditions in Experiment 1 (nonmusicians only), replotted on a perceptual scale calculated from dissimilarity ratings for each parameter in Experiment 2. The error bars show the standard error of the mean.

(level 1 to level 14). The analysis was performed across the first 14 levels of the perceptual scale, where data points were available from all three conditions. In order to perform this analysis, difficulty ratings were calculated for each condition by interpolation of the physical level. Significant main effects were found for condition, $F(2, 34) = 13.4$, $p < .001$, level, $F(1, 17) = 178.03$, $p < .001$, and a condition x level interaction, $F(2, 34) = 7.33$, $p = .048$.

A posthoc Tukey-HSD analysis was run on the condition and the level. The level was regrouped into two categories encompassing perceptual distance levels 0-6 and 7-14. The analysis on the condition factor across all perceptual distance levels showed that difficulty ratings in the loudness level condition were significantly lower than in the temporal and spectral envelope conditions, $p < .001$ for both. Furthermore, difficulty ratings were significantly lower in the spectral envelope compared to temporal envelope conditions, $p = .042$.

For perceptual distance levels 0-6, difficulty ratings in the intensity condition were significantly lower than those in both the temporal and spectral envelope conditions, $p < .001$. For perceptual distance levels 7-14, difficulty ratings were significantly lower in the temporal envelope condition compared to the intensity and spectral envelope conditions, $p < .001$.

Combined, the two analyses revealed a different effect of the perceptual distance on the difficulty rating judgments for each condition (i.e., the slope of the difficulty ratings as a function of perceptual distance depended on the perceptual dimension manipulated, and the magnitude of the perceptual difference).

Figure 10 shows the result of the three conditions of Experiment 1 for the nonmusician group. Difficulty ratings were generally lower and decreased more rapidly as the perceptual differences induced by the spectral envelope and intensity parameters were increased compared to the temporal envelope. This indicates that nonmusicians needed more perceptual change of the temporal envelope to improve the perception of a melody interleaved with distractor notes than for the other two parameters. An analogous statistical analysis was repeated. Significant main effects were found for condition, $F(2, 36) = 5.97$, $p < .006$, level, $F(1, 18) = 414.18$, $p < .001$, and a condition x level interaction, $F(2, 36) = 8.32$, $p = .001$. Across all perceptual distance levels, Tukey-HSD posthoc tests showed significantly higher difficulty ratings in the temporal envelope condition compared to the spectral envelope and the intensity conditions ($p < .001$ for both). No significant difference was found between the spectral envelope and intensity conditions. When considering only perceptual distance levels 0-6, no significant differences between the three conditions were found. However, for perceptual distance levels 7-14, difficulty ratings were significantly lower in the spectral envelope compared to temporal envelope conditions, $p < .001$, and in the loudness condition compared to temporal envelope condition, $p < .001$.

Discussion

THE EFFECT OF PERCEPTUAL CUES ON MELODY SEPARATION

The current study aimed to determine the relative effects of loudness and temporal and spectral aspects of timbre on melody separation. By using a combination of difficulty of melody perception and dissimilarity rating tasks, the most efficient acoustic cue (that which required the least perceptual change while still allowing perception of the melody) was determined.

Several recent studies on auditory streaming have compared the role of peripheral and central processes in stream segregation. Some studies (Beauvois & Meddis, 1996; Hartmann & Johnson, 1991) have argued that streaming processes are mainly influenced by the amount of separation on the basilar membrane (the peripheral channeling theory). Others (Carlyon et al., 2001; Moore & Gockel, 2002) have shown the influence of attention in stream segregation, thus demonstrating an active role of high level processes. The results of Experiment 1 were dependent on music training. Musicians showed lower streaming difficulty ratings for all of the acoustic cues tested. This could be because the musicians were able to perceive smaller acoustic differences between the streams, or because high-level streaming processes were more effective in musicians. In Experiment 2, it was shown that the MDS space for musicians and nonmusicians was the same. However, when the results from Experiment 1 were reanalyzed based on the perceptual scale determined in Experiment 2, it was found that when the perceptual difference between streams was small, the loudness level was the most effective acoustic cue, particularly for musicians.

Therefore, a given acoustic difference between sounds induced the same dissimilarity for both musicians and nonmusicians, but the effect on streaming difficulty ratings differed depending on the specific perceptual dimension. The same dissimilarity between streams allowed a greater reduction in streaming difficulty ratings for musicians compared to nonmusicians, suggesting that for a given dissimilarity between streams, streaming difficulty ratings were affected by an interaction between the specific perceptual dimension and musical expertise.

THE INTERACTION OF TRAINING AND THE TYPE OF PERCEPTUAL DIMENSION ON MELODY SEPARATION

The main finding was that for those with music training, the loudness level was the most efficient cue, whereas for nonmusicians, the most efficient cues were both the loudness level and the spectral envelope. The results show a clear difference based on music training. In Experiment 1, it was shown that musicians required less of a difference than nonmusicians in loudness level, impulsiveness, and the spectral aspect of timbre to better perceive a 4-note melody within random interleaved distractor notes. Experiment 2 showed that for musicians, the loudness level was the most efficient perceptual cue, particularly when the perceptual differences between streams were small. In other words, less perceptual change in loudness was required to improve the perception of a melody interleaved with distractor notes than change in timbre (influenced by the temporal and spectral envelopes). For musicians, this result may be related to the acoustics of musical instruments. In most instruments, loudness is the parameter, after pitch of course, that musicians can control most effectively. It may therefore be used predominantly to promote melody segregation. For example, on the snare drums, different streams of rhythmic pattern are created predominantly by applying different loudness to specific hits. The importance of loudness cues for musicians also increases as the perceptual differences between streams become smaller – as the task of segregating becomes more difficult, the role of loudness cues appears to become more important compared to timbre cues. When the perceptual differences between streams are larger, however, both spectral envelope and loudness cues are significantly more effective than temporal envelope cues.

For the nonmusician group, the timbre dimension that correlated with the temporal envelope was the least efficient perceptual cue compared to both loudness and the spectral envelope when the differences between streams were large. Even with a relatively large perceptual change on the temporal envelope dimension, listeners had difficulty hearing the melody among the interleaved distractor notes. As the perceptual difference between streams became smaller, however, the difference between the perceptual dimensions disappeared.

Thus, when the differences between streams were small, and the task became more difficult, musicians were able to utilize the difference in loudness cues between streams in order to perform the melody segregation task, whereas nonmusicians were not, despite the fact that the MDS task indicated that both groups reported the same perceptual difference induced by all acoustic cues, including the loudness level. The findings from the current study indicate that musicians generally found the target melody easier to perceive, compared to nonmusicians, and were able to take advantage of loudness cues when the task became difficult. Training may therefore have an effect on the ability to achieve voluntary stream segregation.

It is not clear why individuals with music training judge the melody as easier to separate from the distractor than nonmusicians. However, it is well known that music training has effects on brain structure and function at a variety of levels. For example, a combined magnetoencephalography and structural MRI study (Schneider et al., 2002) has shown that musical aptitude is correlated with both the gray matter volume of Heschl's gyrus (a structure containing the primary auditory cortex) as well as tone-evoked neural activity in this gyrus. At lower levels in the brain, musicians also show faster responses and enhanced representation of pitch and timbre in the brainstem to music and speech stimuli (Musacchia, Strait, & Kraus, 2008). While the cortical training-related changes may reflect schema-driven effects on streaming, the brainstem changes could combine to produce an enhanced perception of the acoustic streaming cues, and therefore improved auditory streaming. The results from the current study are consistent with top-down explanations, although they cannot rule out the involvement of mechanisms that may provide enhanced perception of the acoustic cues that allow stream segregation.

A SPECULATION ON THE AUDITORY PROCESSING OF LANGUAGE AND MUSIC

It is unclear why different perceptual cues are more or less efficient depending on music training. We speculate that this difference may be due to the different ways in which musicians and nonmusicians process different types of auditory stimuli. By showing that the overlap in fMRI activation between speech-related and music-related areas of the brain decreases with increasing musical expertise, Wilson, Abbott, Lusher, Gentle, and Jackson (2011) have suggested that nonexpert singers are more dependent than expert singers on language networks while singing. It is therefore reasonable to speculate that the nonmusicians in the current study may have performed the task as if the signal was a speech-like signal. In that context, it is reasonable to suppose that the dimension that correlated with the temporal envelope requires the largest perceptual change before streaming starts to occur. As in speech-like signals, sequential phonemes must remain grouped together despite large temporal differences (impulsive consonants against vowels).

In a study comparing language and music composition, Patel and Daniele (2003) computed the duration variability of adjacent phonemes in the English language and rhythmic patterns in English music. They calculated the "normalized Pairwise Variability Index" nPVI:

$$nPVI = \frac{100}{m-1} \times \sum_{k=1}^{m-1} \left| \frac{2(d_k - d_{k+1})}{(d_k + d_{k+1})} \right| \qquad (1)$$

In Equation 1, m is the number of vocalic intervals or musical notes in an utterance, and d_k is the duration of the kth interval. The nPVI provides an overall measure of the variability of adjacent elements in both music and speech. An nPVI of 66.99 was found for English language and 46.91 for English music. Using the nPVI, it is possible to compare the result of the impulsiveness condition in Experiment 1 with the data collected by Patel and Daniele (2003). Interestingly, the impulsiveness ratio where nonmusicians show an average difficulty rating of .5 (which can be considered as the fission boundary) is 51.70%, which is equivalent to a nPVI of 63.7, and the corresponding ratio for the musician is 61.75% which is equivalent to a nPVI of 47.3. This result shows that the nonmusicians in the current study reached the .5 difficulty rating when the ratio of impulsiveness between the target and the distractor was about the same as the average variability in speech, and for musicians, when the ratio was about the same as the average variability in music.

A similar acoustical analysis for loudness level was performed by comparing the variability within speech sentences and 5-second monophonic musical phrases. The speech sentences were extracted from the DARPA TIMIT database (Fisher, Doddington, & Goudie-Marshall, 1986). It is composed of five sentences recorded by 11 American-English speakers (including four females). The music samples were composed of nine instruments (cello, clarinet, flute, female singer, male singer, trumpet, and violin). For each instrument, four different 5-s musical samples were extracted from commercial recordings. Each sample was monophonic and was typical of the instrument played (for example, a Bach cello suite and a trumpet solo by Miles Davis). The loudness level variability was extracted according to a time-variable loudness model (Glasberg & Moore, 2002). This model describes how the perceived loudness of a sound varies with time. The loudness level variability was evaluated by calculating the standard deviation of the short-term instantaneous loudness for all the speech and musical segments. The average standard deviation for the musical samples was 8.7 phons, and for the speech samples was 15.8 phons. These values are broadly within the same range as the loudness level difference between target melody and distractor notes at the .5 difficulty point for the musicians (6.4 phons) and nonmusicians (10.2 phons).

Similarly, the spectral centroid was derived using a running Hanning window of 200-ms for each sentence.

As suggested by Marozeau and de Cheveigné (2007), the F0 was extracted using the YIN algorithm (de Cheveigné & Kawahara, 2002) and subtracted from the centroid in order to control for its effect. The standard deviation of the percentage variation from the mean estimated spectral centroid over the 5-s sample was then calculated. The average centroid variation was 26% for musical samples, and 34% for speech. These results are somewhat larger compared to the 16% change in spectral centroid for musicians and 19% for nonmusicians found at the .5 difficulty point in Experiment 1. However, results from the spectral centroid depend on many external parameters such as the type of data base, the size of the window, or the accuracy of the F0 extractor, and so must be interpreted with care.

TASK CONSIDERATIONS

It could be argued that the task employed in the current study is susceptible to variations in response bias between participants. This potential drawback was addressed in a previous study (Marozeau et al., 2010), where results from the difficulty rating task used in the current study were compared with an objective detection task. In the detection task, participants were asked to detect occasional variations in the target melody as the acoustic parameters of the interleaved distractor notes were slowly varied. The results showed that the listeners who judged the melody to be difficult to perceive overall had the lowest accuracy scores in the detection task. Conversely, those who judged the melody as easy to perceive overall had the highest accuracy scores. Since stream segregation is necessary to detect the occasional melody variations, this comparison showed that ratings in the difficulty rating task were a reliable indicator of stream segregation. Furthermore, the pattern of responses in Experiment 1 of the current study varied from one acoustic cue to another, but not from one repetition to another.

Conclusion

This paper presented two experiments that aimed to estimate segregation boundaries based on three physical cues, study the effect of music training on segregation, and study the relative efficiency of three perceptual cues on stream segregation.

The first experiment showed that participants with music training needed a difference of 61.75% in impulsiveness ratio, 16% in spectral centroid, or 6.4 phons in loudness level to reach a .5 subjective difficulty rating threshold for segregating a simple melody from interleaved distractor notes. Participants without a musical background needed a difference of 51.7% in impulsiveness ratio, 19% in spectral centroid, or 10.2 phons in loudness level to reach the .5 difficulty rating. Statistical analyses revealed a significant effect of music training on segregation difficulty ratings, with musicians requiring significantly less acoustic difference between streams in order to segregate them.

The second experiment showed that when the perceptual difference between the target melody and distractor notes was small, participants with music training required smaller perceptual differences in loudness to allow segregation of the melody from interleaved distractor notes, compared to the perceptual differences required based on spectral and temporal envelopes. For participants with less music training, smaller perceptual changes in spectral envelope and loudness level were required compared to changes in temporal envelope, although these effects vanished when the perceptual difference between target melody and distractor notes was small. It is speculated that these differences between musicians and nonmusicians may be due to differences in the processing of the stimuli. While the musicians processed the stimuli using specialized music networks that develop during training, nonmusicians may rely on networks also used for the analysis of speech signals.

Moore and Gockel (2002) have suggested that, "the extent to which sequential stream segregation occurs is directly related to the degree of perceptual difference between successive sounds" (p. 331). This hypothesis can be interpreted to suggest that the same amount of perceptual change should allow the same amount of stream segregation, independently of the acoustic cues that relate to the perceptual differences. If this were the case, the results in Figures 9 and 10 should show no differences between the three cues tested. However, this was not the case – the results showed that the type of acoustic cue had a different effect on the difficulty of segregating the melody, even though the perceptual differences between streams were the same. Therefore we suggest a new form of Moore and Gockel's hypothesis: "The extent to which sequential stream segregation occurs is directly related to the degree of perceptual difference between successive sounds, the type of perceptual dimension, and the listener's musical background." Further research will be required to test this new form of the hypothesis, especially in the case of obligatory streaming.

Author Note

The authors are grateful to the volunteers who participated in this research. Financial support was provided

by the Jack Brockhoff Foundation, Goldman Sachs JB Were Foundation, Soma Health Pty Ltd, Mr Robert Albert AO RFDRD, Miss Betty Amsden OAM, Bruce Parncutt and Robin Campbell, The Frederick and Winnifred Grassick Memorial Fund, NHMRC Project Grant 1008882, and the Victorian Lions Foundation. The Bionics Institute acknowledges the support it receives from the Victorian Government through its Operational Infrastructure Support Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors thank Professor Stephen McAdams and three anonymous reviews for their helpful suggestions on an earlier version of this manuscript.

Correspondence concerning this article should be addressed to Jeremy Marozeau, The Bionics Institute, 384-388 Albert Street, East Melbourne, VIC 3002. E-mail: jmarozeau@bionicsinstitute.org

References

American National Standard Institute. (1996). *Specification for audiometers*. S3.6-1996 C.F.R.

American National Standard Institute. (2007). *Procedure for the computation of loudness of steady sounds*. S3.4-2007 C.F.R.

Beauvois, M. W., & McAdams, S. (1996). Stimulus intensity and auditory stream formation. *ACUSTICA, 82*, S85.

Beauvois, M. W., & Meddis, R. (1996). Computer simulation of auditory stream segregation in alternating-tone sequences. *Journal of the Acoustical Society of America, 99*, 2270-2280.

Beauvois, M. W., & Meddis, R. (1997). Time decay of auditory stream biasing. *Perception and Psychophysics, 59*, 81-86.

Bendixen, A., Denham, S. L., Gyimesi, K., & Winkler, I. (2010). Regular patterns stabilize auditory streams. *Journal of the Acoustical Society of America, 128*, 3658-3666.

Bey, C., & McAdams, S. (2003). Postrecognition of interleaved melodies as an indirect measure of auditory stream formation. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 267-279.

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Daquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion, 19*, 1113-1139.

Borg, I., & Groenen, P. J. F. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America, 118*, 471-482.

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences, 8*, 465-471.

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 115-127.

Cusack, R., & Roberts, B. (1999). Effects of similarity in bandwidth on the auditory sequential streaming of two-tone complexes. *Perception, 28*, 1281-1289.

Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception and Psychophysics, 62*, 1112-1120.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111*, 1917-1930.

Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive Psychology, 5*, 322-337.

Fisher, W. M., Doddington, G. R., & Goudie-Marshall, K. M. (1986, February). *The DARPA Speech Recognition Research Database: Specifications and Status.* Paper presented at the Proceedings of DARPA Workshop on Speech Recognition.

Glasberg, B. R., & Moore, B. C. J. (2002). A model of loudness applicable to time-varying sounds. *Journal of the Audio Engineering Society, 50*, 331-342.

Hartmann, W. M., & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception, 9*, 155-184.

Iverson, P. (1995). Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 751-763.

Marozeau, J., & de Cheveigné, A. (2007). The effect of fundamental frequency on the brightness dimension of timbre. *Journal of the Acoustical Society of America, 121*, 383-387.

Marozeau, J., de Cheveigné, A., McAdams, S., & Winsberg, S. (2003). The dependency of timbre on fundamental frequency. *Journal of the Acoustical Society of America, 114*, 2946-2957.

Marozeau, J., Innes-Brown, H., Grayden, D. B., Burkitt, A. N., & Blamey, P. (2010). The effect of visual cues on auditory stream segregation in musicians and non-musicians. *PLoS ONE, 5*, e11297.

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research, 58*, 177-192.

Moore, B. C. J., & Gockel, H. (2002). Factors influencing sequential stream segregation. *Acta Acustica, 88*, 320-332.

Musacchia, G., Strait, D., & Kraus, N. (2008). Relationships between behavior, brainstem and cortical encoding of seen and

heard speech in musicians and non-musicians. *Hearing Research, 241*, 34-42.

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition, 87*, B35-B45.

Pressnitzer, D., & Hupe, J. M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology, 16*, 1351-1357.

Schneider, P., Scherg, M., Dosch, H. G., Specht, H. J., Gutschalk, A., & Rupp, A. (2002). Morphology of Heschl's gyrus reflects enhanced activation in the auditory cortex of musicians. *Nature Neuroscience, 5*, 688-694.

Singh, P. G., & Bregman, A. S. (1997). The influence of different timbre attributes on the perceptual segregation of complex-tone sequences. *Journal of the Acoustical Society of America, 102*, 1943-1952.

van Noorden, L. (1975). *Temporal coherence in the perception of tone sequences.* Unpublished doctoral dissertation, University of Eindhoven, Eindhoven, Netherlands.

Vliegen, J., & Oxenham, A. J. (1999). Sequential stream segregation in the absence of spectral cues. *Journal of the Acoustical Society of America, 105*, 339-346.

Wessel, D. (1979). Timbre space as a musical control structure. *Computer Music Journal, 3*, 45-52.

Wilson, S. J., Abbott, D. F., Lusher, D., Gentle, E. C., & Jackson, G. D. (2011). Finding your voice: A singing lesson from functional imaging. *Human Brain Mapping, 12, 2115-2130*.

Zendel, B. R., & Alain, C. (2009). Concurrent sound segregation is enhanced in musicians. *Journal of Cognitive Neuroscience, 21*, 1488-1498.