EFFECT OF TIMBRE ON MELODY RECOGNITION IN THREE-VOICE COUNTERPOINT MUSIC

Song Hui Chon, Kevin Schwartzbach, Bennett Smith, Stephen McAdams
CIRMMT (Centre for Interdisciplinary Research in Music Media and Technology)
Schulich School of Music
McGill University
songhui.chon@mail.mcgill.ca

ABSTRACT

Timbre saliency refers to the attention-capturing quality of timbre. Can we make one musical line stand out of multiple concurrent lines using a highly salient timbre on it? This is the question we ask in this paper using a melody recognition task in counterpoint music.

Three-voice stimuli were generated using instrument timbres that were chosen following specific conditions of timbre saliency and timbre dissimilarity. A listening experiment was carried out with 36 musicians without absolute pitch. No effect of gender was found in the recognition data. Although a strong difference was observed for the middle voice from mono-timbre to multi-timbre conditions, timbre saliency and timbre dissimilarity conditions did not appear to have systematic effects on the average recognition rate as we hypothesized. This could be due to the variability in the excerpts used for certain conditions, or more fundamentally, because the context effect of each voice position might have been much bigger than the effects of timbre conditions we were trying to measure. A further discussion is presented on possible context effects.

1. INTRODUCTION

1.1 Timbre Saliency

Timbre saliency is a new concept we proposed regarding the attention-capturing quality of timbre [1]. It was measured using tapping to perceptually isochronous ABAB sequences, the pitch (C4), loudness and effective duration of which were all equalized. The duration of each stimulus was controlled by imposing a raised cosine decay envelope at a point corresponding to the effective duration of 200 ms on a recorded sample from the Vienna Symphonic Library [2]. All sounds were selected from those playing mezzo-forte in the most basic manner (such as bowing on the cello rather than plucking). The hypothesis was that the more salient a timbre is, the more attention it will draw from the participants, and hence be tapped to more often. Figure 1 shows the one-dimensional saliency scale obtained from CLASCAL [3]. Although the saliency scale is one-dimensional, it is presented in two dimensions because of the seven instruments closely positioned around 0.

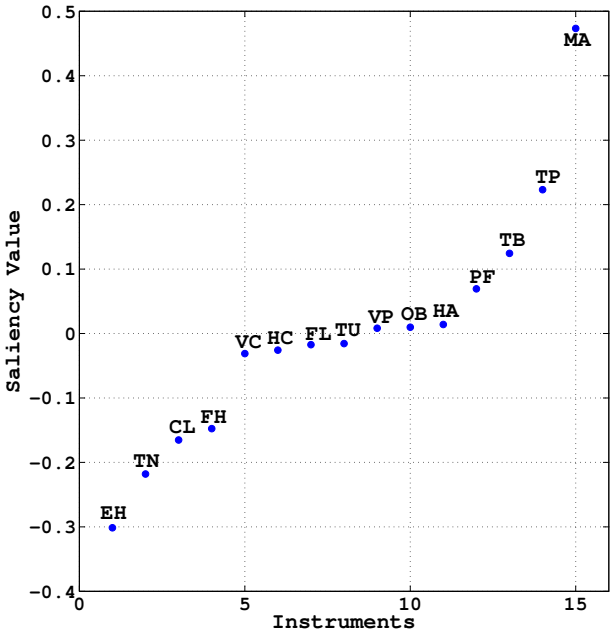

Figure 1. One-dimensional timbre saliency space of 15 timbres: Clarinet (CL), English Horn (EH), French Horn (FH), Flute (FL), Harp (HA), Harpsichord (HC), Marimba (MA), Oboe (OB), Piano (PF), Trombone (TN), Trumpet (TP), Tuba (TU), Tubular Bells (TB), Violoncello (VC), and Vibraphone (VP).

As saliency refers to the character of an object that makes it stand out from its surroundings, we next studied the effect of saliency on the perceived blending of concurrent unison dyads [4]. 105 composite sounds were created using pairs of non-identical timbres that were used in the tapping experiment [1]. Rating data from 60 people showed that, as we hypothesized, a highly salient timbre would not blend well with others, although the degree of correlation was mild at most. Attack time and spectral centroid were most efficient in describing the blend ratings, which are the two acoustic features that were reported in previous studies of the blend perception [5, 6], verifying that a sound will tend to blend better when it has more low-frequency energy and when it starts slowly.

After studying the effect of timbre saliency on the sim-

plest musical situation of unison concurrent dyads, the next step is an investigation in a more musically realistic scenario. For example, it has been known that the entries of inner voices are more difficult to detect than those of outer voices in polyphonic music [7]. Therefore, can we enhance the detection of an inner voice by applying a salient instrument timbre to it?

To answer this question, we decided to employ a melody recognition task. Iverson, and Bey & McAdams found that having two highly dissimilar timbres helped the recognition of the target melodies that were interleaved with distractors [8,9]. Using concurrent melodies, Huron observed that in general musicians were capable of correctly identifying the number of voices, although the performance degraded as the number of voices increased, especially beyond three [7]. Gregory found that concurrent melodies that had simultaneous note onsets in the same pitch range in a related key tended to be easier to perceive if they were distinguished by timbre differences [10]. Although this result suggests that listeners can attend to more than one musical line at a time, it might need to be interpreted with caution because the voices in musical excerpts in the study were not controlled carefully and some excerpts might have been too well-known (such as the one from Mozart's *Don Giovanni*).

As we aimed to expand the study of the effect of timbre saliency in a more musically realistic setting, the method of melody recognition in counterpoint music was deemed to be appropriate. There are two or more musical lines with virtually equal musical importance. Since the authors, who knew the melodies in the excerpts by heart, could not listen to all voices in an excerpt at once, it is practically impossible for listeners to attend to every note of every voice. Therefore they would tend to focus on whatever voice catches their attention. Hence, if we can control the timbre saliency of the voices in music, listeners' tendency to attend to a specific voice must reflect the voice's saliency. But since it is difficult for us to figure out which voice each listener is hearing out at a given moment, we decided to use a comparison task based on melody recognition. If, for example, a listener happened to focus more on the high voice melody and was tested with a high-voice comparison melody, he or she would be more likely to answer correctly than someone who happened to focus on the low voice. Therefore performance in this task should covary with voice prominence.

Since this is a very complex experiment, we had to run two experiments for preparation. One was to study the dissimilarity of the timbres that were used in our saliency experiment (Section 1.2). The other was a melody comparison experiment to make sure that the changes on a voice were easy enough to hear out in isolation (Section 3). The design of musical stimuli, which took place before the melody comparison experiment, is explained in detail in Section 2. Section 4 discusses the main experiment, then finally a general discussion and conclusions are presented in Section 5.

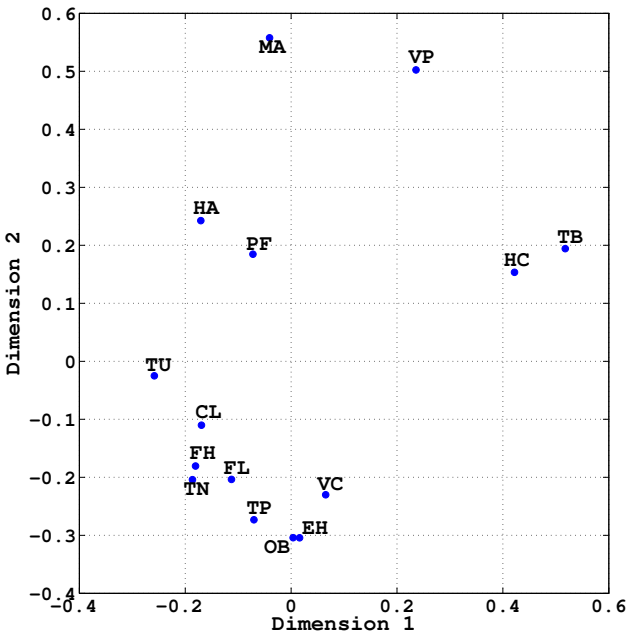

Figure 2. Two-dimensional timbre dissimilarity space. See Fig. 1 caption for abbreviations.

1.2 Timbre Dissimilarity

A classic timbre dissimilarity experiment was carried out using the same set of 15 isolated instrument sounds used in the timbre saliency experiment [1]. Twenty participants, balanced in gender and musicianship were recruited, with ages from 19 to 39 with a median age of 26.5 years.

Repeated-measures ANOVAs on dissimilarity ratings showed no effect of gender or musicianship. The dissimilarity judgments were formed into 20 individual lower triangular matrices, then analyzed by CLASCAL [3] to obtain the dissimilarity space. The best solution turned out to have two dimensions with specificities and five latent classes of participants (Figure 2).

Note that the percussive instruments are all located above the $y = 0$ line. This suggests that the second dimension may be related to attack time. Correlations were computed between each of the two dimensions and the acoustic features computed by the Timbre Toolbox [11]. The first dimension shows a high correlation with spectral centroid in the ERB-FFT spectrum, $r(13) = .845, p < .0001$, and the second dimension a moderate correlation with attack time, $r(13) = -.692, p = .004$. This is in agreement with previous studies in timbre dissimilarity showing that attack time and spectral centroid are two of the most important acoustic features [12–17].

This two-dimensional timbre dissimilarity space in Figure 2 will provide a basis for the selection of stimuli for Experiments 1 and 2. This is necessary because it is not feasible to study all 15 timbres' effect on melody recognition, and therefore we need to select timbres that best represent the experimental conditions. This timbre dissimilarity space will also be essential in data analysis as the dissimilarity distance is one of the main parameters for Experiment 2.

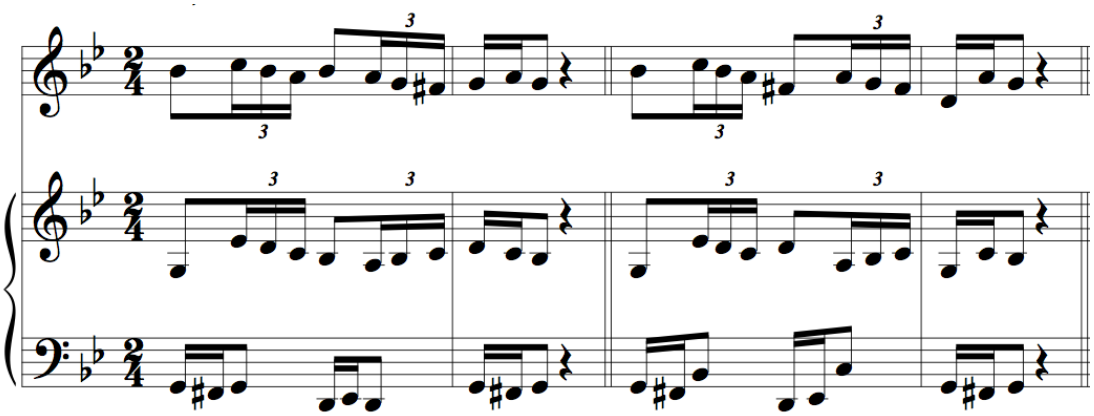

Figure 3. An example excerpt and corresponding comparison melodies

2. MUSICAL STIMULUS DESIGN

A number of excerpts and their comparison melodies are needed to avoid any unexpected training effect from participants. We selected nine excerpts from J.S. Bach's *Trio Sonatas for Organ*, BWV 525 – 530, because the music was already clearly written for three-voices (right hand, left hand and pedal) and relatively unknown in comparison with other three-voice pieces (such as the *Sinfonias*).

We looked for the parts with all three voices clearly in action with about equal note onset density. Any excerpts with voice crossings were avoided. We also did some editing of the excerpts such as transposing the melodies to a new key (often to accommodate the playing ranges of selected instruments), changing the pitch of a note (often by an octave) to avoid voice crossing, or breaking a longer note into two shorter notes to maintain the note onset density.

For each voice in each excerpt, a comparison melody was composed by changing the pitches of two notes, which resulted in a different pitch contour, following the approach in auditory streaming studies using interleaved melodies [9]. An example is shown in Figure 3. The first two measures show the three-voice excerpt by Bach and the last two measures the corresponding comparison melodies. In the actual experiment all three voices in an excerpt will play together, whereas the three comparison melodies will never be heard together.

The excerpts were first encoded in Finale [18], the MIDI timings of which were exported to Logic [19]. The stimuli were created using the recorded samples in the Vienna Symphonic Library [2] based on the MIDI timing information. The specific timbre combinations used for stimulus generation are presented in the next section.

2.1 Timbre Combinations

A subset of instruments was chosen that would best represent the timbre saliency and timbre dissimilarity conditions from the two spaces in Figures 1 and 2, respectively. We decided to focus on a subset of timbre combinations in which two timbres are similar and the other one is different (i.e., two are close to each other and the third one is far from these two in timbre *dissimilarity* space), and one is a highly salient timbre and two others are of lower saliency. Three timbre dissimilarity conditions combined with three timbre saliency conditions resulted in nine conditions (Table 1).

D1, D2 and D3 represent the three *dissimilarity* conditions according to the assignments of three timbres to three voices. Among the three timbres, T1, T2 and T3, *T3* is always the "far" timbre and is highlighted in blue italics. Similarly, S1, S2 and S3 represent the three *saliency* conditions. The "**H**igh" saliency timbre of the three timbres is highlighted in a bold red font. For example, the D1S1 column in Table 1 shows that in this condition high and middle voices have timbres that are of low saliency and close in dissimilarity space. This factorial combination of saliency and dissimilarity allows us to test their separate contributions to melody recognition, as well as their potential interaction.

Even though there are nine conditions, it turns out that only four sets of timbre assignments are required – {D1S1, D2S2, D3S3}, {D1S2}, {D1S3, D2S3, D3S2}, and {D2S1, D3S1}, as specified with four types of fonts in Table 2. These combinations were chosen considering not only the relative positions in timbre dissimilarity and timbre saliency spaces, but also the instrument ranges, because some instruments cannot play higher notes in the top voice and others cannot play the lower notes in the bottom voice.

In addition, we need to test the same-timbre version of all stimuli, to determine baseline performance in the absence of timbre differences. We decided to use the piano (PF) for this, not only because it has a sufficient range for all excerpts, but because its timbre is quite homogeneous over the middle range, which is used primarily in the current study.

In searching for the right timbre combinations for the conditions specified in Table 2, we had to make some compromises by using some instruments with medium saliency. More specifically, Harpsichord (HC) was used in place of some lower saliency instruments. This was the best we could do with the two given spaces (Figures 1 and 2), especially because nine out of fifteen timbres were located together in the lower left corner of the timbre dissimilarity space (Fig. 2).

3. EXPERIMENT 1: MELODY DISCRIMINATION

The goal of this experiment was to verify that the changes in pairs of melodies were easy enough to detect in isolation

Table 1. Timbre conditions for three-voice excerpts

	D1S1	D1S2	D1S3	D2S1	D2S2	D2S3	D3S1	D3S2	D3S3
High	T1L	T1L	T1**H**	T2L	T2L	T2**H**	*T3*L	*T3*L	*T3***H**
Middle	T2L	T2**H**	T2L	*T3*L	*T3***H**	*T3*L	T1L	T1**H**	T1L
Low	*T3***H**	*T3*L	*T3*L	T1**H**	T1L	T1L	T2**H**	T2L	T2L

Table 2. Timbre assignments for three-voice excerpts

	D1S1	D1S2	**D1S3**	*D2S1*	D2S2	**D2S3**	*D3S1*	**D3S2**	D3S3
High	T1L	T1L	**T1H**	*T2L*	T2L	**T2H**	*T3L*	**T3L**	T3H
Middle	T2L	T2H	**T2L**	*T3L*	T3H	**T3L**	*T1L*	**T1H**	T1L
Low	T3H	T3L	**T3L**	*T1H*	T1L	**T1L**	*T2H*	**T2L**	T2L
T1	CL	EH	**TP**	*MA*	TN	**TN**	*VP*	**TP**	CL
T2	TN	TP	**TN**	*VP*	CL	**TP**	*MA*	**TN**	TN
T3	MA	HC	**HC**	*CL*	MA	**HC**	*CL*	**HC**	MA

at least 75% of the time, because if participants cannot hear changes in corresponding melodies in isolation, they will not be able to hear out changes on one voice in a mixture with other voice(s). The stimuli were 108 ordered pairs of "original" and "comparison" multi-timbre melodies from all three voices in nine excerpts: original-original, original-comparison, comparison-original, and comparison-comparison. These were presented to the participants in a random order without an option to repeat. Participants were required to indicate whether a given pair of melodies was identical or not on the graphic user interface, which then automatically proceeded to the next trial.

Twenty musicians (10 males) without absolute pitch were recruited, aged from 18 to 37 with a median of 24 years. There was quite a large variability in the participants' average performances, ranging from 69% to 92% correct, with a median of 84%. All melody pairs showed correct discrimination above 75% with the exception of one pair at 72.5%. As the 75% threshold was somewhat arbitrary and 72.5% is not too far from 75%, we decided to proceed to the main experiment using the current modified melodies without any further adjustments.

4. EXPERIMENT 2: MELODY RECOGNITION IN THREE-VOICE COUNTERPOINT MUSIC

4.1 Methods

This experiment studied the role of timbre dissimilarity and saliency in melody recognition in counterpoint music. Stimuli were the three-voice Bach excerpts, as well as the individual monophonic melodies. For each trial, a multi-voice excerpt would play first, followed by a monophonic melody. The monophonic melody could be the original or comparison melody corresponding to one of the voices in the preceding excerpt. Participants were required to indicate whether the monophonic melody was the same as or different from a voice in the excerpt by pressing on the appropriate button on the graphic user interface. There was no option to listen to the stimuli again to prevent participants from strategically learning all voices by attending to one voice each time over repeats. Once an answer was submitted, the next trial would start automatically, playing a new multi-voice excerpt.

Thirty-six musicians without absolute pitch took part in the experiment. Their ages ranged from 18 to 37, with a median of 24 years. There were equal numbers of males and females. Nineteen of them identified themselves as "professional" musicians and the rest as "amateurs". In terms of their listening habits, 15 claimed to be "harmony-listeners" and 21 to be "melody-listeners." Although we have not come across any literature on the effect of this listening habit on the listeners perception of voices in counterpoint music, we thought the melody-listeners might focus on one prominent voice whereas the harmony-listeners would focus on emergent properties of all voices.

4.2 Results

4.2.1 Average Performance Per Condition

The main goal of this experiment was to examine the melody recognition performance in terms of timbre conditions based on timbre saliency and timbre dissimilarity. For this purpose, we computed the average recognition rate over all melodies used per voice per condition and compared those average values (Figure 4). The horizontal axis shows the saliency conditions and each line represents the dissimilarity conditions.

Considering only the mean values (blue dots, black stars and red triangles), we see they loosely follow a v-shape, although sometimes flipped upside down or almost flattened. The three v-shaped lines in the middle voice appear to maintain the same direction, which suggests that the timbre saliency condition may play an important role in the recognition of the middle voices. The fact that the lines keep a similar shape in the middle voice graph but not in other two voices implies a possible main effect of voice position or an interaction between timbre saliency and voice position.

A three-way repeated measures ANOVA was performed on the average recognition rate per condition as the dependent variable. The voice position (high, middle or low), dissimilarity and saliency conditions in Table 1 were within-subjects factors. The only significant effects were interactions between voice position and saliency, $F(4, 140) =$

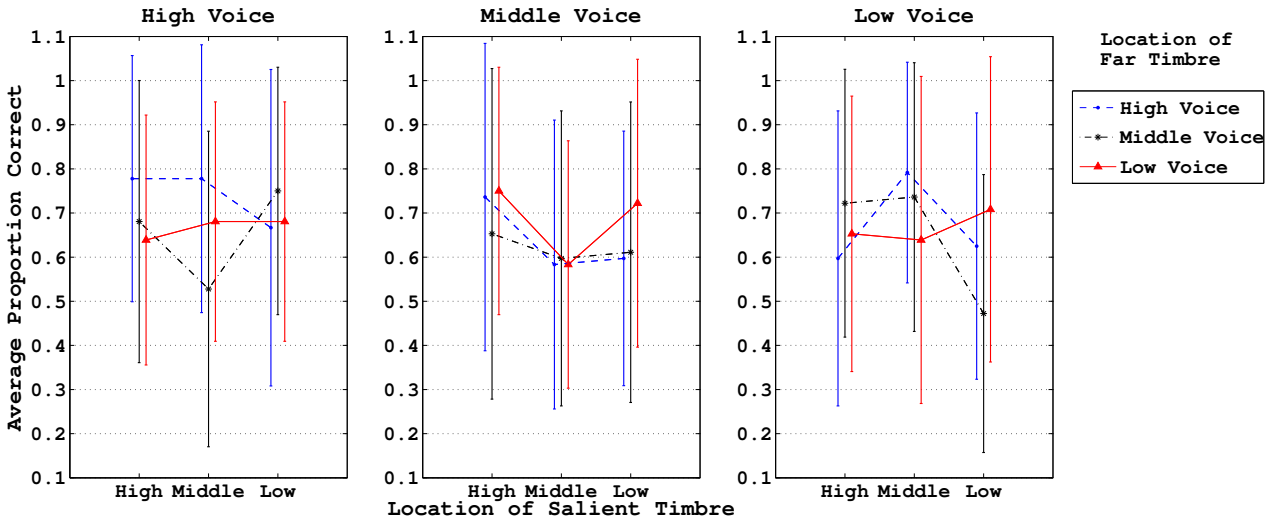

Figure 4. Results for three-voice excerpts. The error bars show ± one standard deviation.

$3.86, p = .005$, and between voice type, saliency, and dissimilarity, $F(8, 280) = 3.14, p = .002$. None of the other effects was significant.

The significant voice position × saliency interaction means that our hypothesis that the effect of saliency condition differs across voices was correct. This may imply that the innate 'voice prominence' from this musical structure may have a bigger impact on melody recognition than the controlled timbre conditions. The significant three-way interaction of voice position, dissimilarity and saliency indicates that the two-way interaction effect between dissimilarity and saliency differs depending on the voice type. This is in agreement with the fact that in Figure 4, the dissimilarity × saliency interaction (i.e., the angles of v-shape lines) seems to be higher for high and low voices, but negligible for the middle voice.

Two-way ANOVAs were performed to study the effect of timbre dissimilarity and timbre saliency for each voice type. On the high voice, the interaction effect was significant, $F(4, 140) = 3.12, p = .017$, but not the main effects of timbre dissimilarity, $F(2, 70) = 2.28, p = .11$, or of timbre saliency, $F(2, 70) = 0.91, p = .41$. In the high-voice graph of Figure 4, the locations of the nine points, corresponding to average performance across participants in nine timbral conditions, are quite different according to timbre conditions, although their vertical or horizontal (per line) averages do not show significant differences (hence non-significant main effects).

On the middle voice, the main effect of timbre saliency turned out to be significant, $F(2, 70) = 4.69, p = .012$, but not timbre dissimilarity, $F(2, 70) = 1.04, p = .36$, nor their interaction $F(4, 140) = 0.71, p = .59$. The three lines in the middle voice graphs of Figure 4 have similar shapes (hence no significant interaction effect) and locations (hence no significant main effect of dissimilarity). The nine points representing the nine conditions have very different vertical means (therefore a significant main effect of saliency), but not so different horizontal means (hence a non-significant main effect of dissimilarity). What is strange is that the performance on the middle voice was at its worst when the salient timbre was on the middle voice. This can be observed in all three dissimilarity conditions, probably suggesting that the effect of a salient timbre was minimal on the middle voice. It is also hard to understand why the recognition performance on the middle voice (black dash-dotted line connecting stars) was the worst when the far timbre was assigned to the middle voice. In summary, this graph seems to suggest the absence of our hypothesized effects of dissimilarity or saliency on the middle voice.

A two-way ANOVA on the low voice showed two significant effects: the main effect of timbre saliency, $F(2, 70) = 3.66, p = .031$, and its interaction with timbre dissimilarity, $F(4, 140) = 4.56, p = .002$. The main effect of timbre dissimilarity was not significant, $F(2, 70) = 0.28, p = .75$. The v-shapes face different directions, reflecting the significant interaction effect. Although the per-dissimilarity condition (i.e., per-line) averages are all located in a similar area (hence no main effect of dissimilarity), the vertical means are at different locations, confirming the significant main effect of saliency. However, it is strange to see that the vertical mean was at its lowest when the salient timbre was on the low voice. Having the salient timbre on the low voice was expected to help the recognition performance, but apparently it did not. A close look reveals that the performance was not too bad when the salient timbre was on the low voice and the far timbre was on the high or low voice. But somehow having a far timbre on the middle voice hindered the recognition of the low voice melody so much that the performance actually fell below 50%. This might result from the saliency differences inherent in the stimuli: somehow the low voice melodies were not salient at all and participants' attention was drawn to the salient high-voice melodies in the given condition.

Overall, it is quite disappointing to see that recognition was not highest (with an exception of the high voice) when a voice had both the salient and the far timbre, which had been hypothesized to have the maximum effect on the recognition task. For example, the high voice graph on the left of Figure 4 reaches the maximum performance at the left blue dot, when the salient and far timbre happened to be on the high voice, but this is not the case in the other two

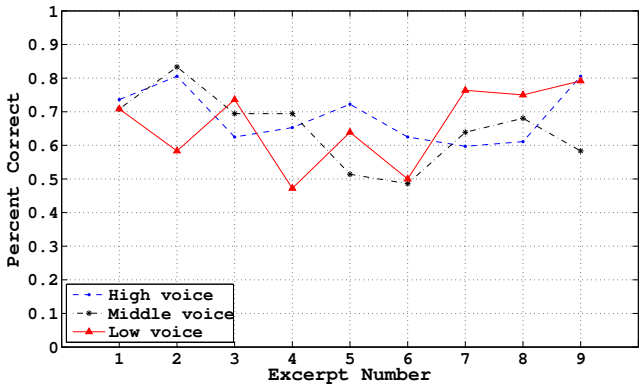

Figure 5. Average recognition per excerpt in the multi-timbre conditions

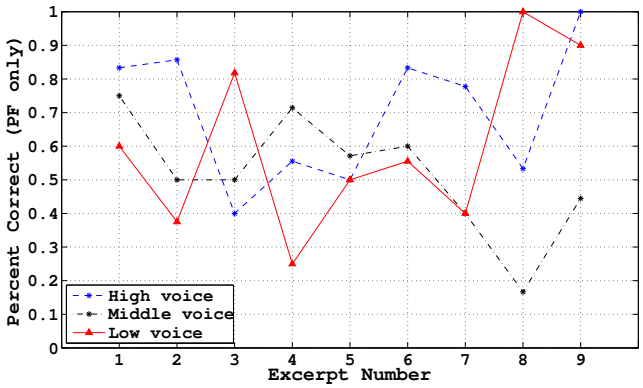

Figure 6. Average recognition per excerpt in the mono-timbre condition

graphs. The black star in the middle of the dash-dotted line of the middle voice graph, which was hypothesized to be the highest point, is located much lower than the actual highest point (a red triangle). In fact, it is puzzling to see the low performance on the middle voice when it was played with the salient timbre. We began to wonder if the middle voice melodies used for this condition happened to be too difficult. To study this, we decided to analyze the average recognition performance for each stimulus, which is presented in the next section.

4.2.2 Average Performance Per Excerpt

The average recognition rates of the nine excerpts across all participants are shown in Figures 5 and 6. There is quite a bit of variability across the excerpts used. This might be due to the fact that some excerpts are more difficult to remember than others. At first glance, the multi-timbre average curves look a bit different from the mono-timbre ones, but paired-sample t tests show that these seeming differences are mostly non-significant. One marginally significant difference was found on the middle voice, $t(8) = 1.89, p = .096$, where the average recognition rate of the middle voice in multi-timbre condition was 0.65 ($STD = 0.11$), whereas that in the mono-timbre condition was 0.52 ($STD = 0.18$). This may suggest that having a distinctive timbre on the middle voice, which is usually the most difficult to listen to in the given musical structure, helps its recognition slightly.

Since the average performance per excerpt varied quite a bit, we came to wonder if this is related to how easily the changes in corresponding voices could be heard out in Experiment 1. Hence, the average recognition rate per excerpt was analyzed in terms of the average percent correct values from Experiment 1. Spearman's rank correlation showed that no correlation was significant. This lack of correlation could reflect the fact that the current experimental task is too complex to be successfully predicted by the control experiment result.

4.3 Discussion

In this experiment, we studied the effects of timbre saliency and timbre dissimilarity on the melody recognition in counterpoint music with nine three-voice excerpts in nine timbre conditions. Considering previous work in auditory streaming that has shown that greater timbre dissimilarity leads to better recognition of interleaved melodies [8, 9], as well as our measurement of timbre saliency [1], we hypothesized that a highly dissimilar or a highly salient timbre would enhance a voice's prominence in a multi-voice texture. We were also confident of our choice of counterpoint music excerpts, where each voice had about equal musical importance.

However, the results from 36 musicians did not confirm our hypothesis. Analysis of per-condition performance of middle and low voices showed a significant effect of saliency, although not in the direction we expected: the average performance was poorer when the salient timbre was located on the target voice. This is completely against our hypothesis, and essentially nullifies the conjecture of the timbre saliency's effect on melody recognition in multipart music.

In searching for an answer to this unexpected pattern, we looked at the average recognition performance for each of the excerpts used. It turned out that there was a large difference in per-excerpt performance, which could have come from various degrees of memorability that affected the recognition performance. This variance in per-excerpt performance could also have contributed to differences in per-condition performance.

As there were no significant differences in average recognition of each excerpt-voice according to the timbre conditions (multi-timbre vs. mono-timbre), with an exception only for the middle voice, the lack of effect of timbre saliency may actually indicate a greater 'voice prominence' in the given musical structure than whatever timbral effects we expected. After all, we had not studied the intrinsic saliency of each voice in the three-voice counterpoint structure. This could be a case of the experimental context affecting the measurement of saliency differences of the objects in the experiment.

However, the fact that the average recognition of the middle voice was marginally higher in the multi-timbre condition in comparison with the mono-timbre condition does speak for the case of timbral effects. The middle voice, which is the most difficult to listen to in three-voice music, became easier to recognize with the use of a timbre different from those on the other voices. Unfortunately, this effect seems too weak to be reflected and measured prop-

erly in the current experimental setup.

The large variance in recognition performance also makes us hesitate in drawing firm conclusions based on the analysis. In the per-condition performance, all the means and the respective confidence intervals overlapped without exception. Hence, the analyses based on mean values lose their effectiveness when we consider the large variance.

It was disappointing not to see the expected effects of timbre saliency and timbre dissimilarity. What we saw instead was another incidence of a context effect, which was possibly a lot stronger than our planned timbral effects in this experiment. To clarify unanswered questions, another experiment using the untested portion of the current stimuli seems to be in order, which will provide data that can complement this experiment so that we can apprehend the big picture.

5. SUMMARY & GENERAL DISCUSSION

To examine the effect of timbre saliency and timbre dissimilarity in a more realistic music listening setting, a melody recognition experiment was carried out as a natural extension of the previous study of the perception of blend in concurrent unison dyads [4]. As a mild negative relationship between timbre saliency and the perceived blend was observed in the concurrent unison dyads, we hypothesized that a highly salient timbre would show little blend with other voices in the musical texture and therefore be heard out more easily. Also considering the effect of timbre dissimilarity, we expected to confirm previous findings in the auditory streaming literature [8,9] that a highly dissimilar timbre on a voice would help detect changes in that voice more easily in the presence of other voices in multipart music.

The high voice did not show any main effects of timbre saliency and timbre dissimilarity conditions; it is already the most prominent voice in the chosen musical structure. This 'voice prominence' was probably a lot more salient than any possible additional benefits from timbre saliency and dissimilarity conditions. There was a significant interaction effect observed though, suggesting that the effect of timbre dissimilarity varied with timbre saliency (and vice versa). Middle and low voices showed a significant effect of timbre saliency condition, but this effect did not go in the same direction as our hypothesis. In fact, the average recognition performance was lowest when the salient timbre was located on the target voice. This was completely unexpected, and we are still puzzled by it.

So we decided to look into the per-excerpt average performance, hoping that it would shed light that could explain the aforementioned observations on middle and high voices. When each excerpt's average recognition performance in multi-timbre condition was contrasted with that in mono-timbre condition, the only marginally significant difference was observed on the middle voice. The recognition performance was much higher on average (by 13%) in the multi-timbre condition. This suggests that the middle voice, which has the least 'voice prominence' in the chosen musical structure, benefited from having a different timbre from the other voices, which agrees with previous literature on timbral effects on auditory streaming.

However, the fact that this additional benefit did not make any significant differences in average performances per timbre condition led us to think about the context effects again. As we hypothesized, there exists an intrinsic saliency for each object and an extrinsic saliency for each context in which the object's saliency is measured. Considering this, the limit in our experiment might have been that we did not consider the inherent prominence of each voice position in the musical form that was selected for the experiment. Even the strong recognition improvement on the middle voice in the condition with multiple timbres may not have covaried systematically with the hypothesized timbre conditions, which could be why there is lack of effect of the timbre conditions.

Reflecting on the complexity of Experiment 2, we wonder if we should have started with a simpler experiment. Perhaps it would help to carry out a new experiment with simplified conditions to verify the effect of timbre saliency and timbre dissimilarity, where the stimuli have only two conditions – a "high" condition with a highly salient and dissimilar (i.e., far in dissimilarity space) timbre and a "low" condition with a not-so-salient and similar timbre. This should be able to clearly contrast the performance in each condition to examine the effect of timbre saliency and timbre dissimilarity. We can also conduct Experiment 2 again with the set of stimuli that were not tested currently. Because each three-voice excerpt in a particular timbre combination was tested with only one voice, we can make use of the untested voices and run the same analysis on the combined data.

Another idea is to conduct an experiment utilizing top-down attention instead of the current melody recognition paradigm, which depends on bottom-up attention and short-term memory. Imagine that a short cue, an isolated note at a certain pitch and timbre, is played right before a polyphonic excerpt is played. What happens to the recognition rate? Do listeners tend to get drawn more towards the voice close to the pitch of the cue? Or to the voice that has the same timbre? This may bring us to an interesting interaction of top-down and bottom-up attention together.

Also, more fundamentally, the relationship between timbre saliency and timbre dissimilarity needs to be examined. In the design of experiments in this paper, we proceeded from assumptions that timbre saliency and timbre dissimilarity would be at least somewhat related to each other and that there would not be any negative interaction between them. Do our assumptions still hold? What is the difference between saliency and dissimilarity? Can one explain the other? After studying their relationship, we might have a new insight to bring to understanding the current results.

One thing that we learned from carrying out this complex experiment is that counterpoint music is such a sophisticated art that it could not be sufficiently analyzed with our model. Saliency is a function of context, and our measure of timbre saliency might not have been effective in the context of melody recognition in counterpoint music, especially when each voice position's prominence is un-

known. As this was our first attempt to explain the perception of multipart music in terms of timbre saliency, any findings are important. However disappointing or puzzling the findings were, these will lead to a new journey with more questions to answer, which will eventually help us understand what catches our attention in music, which was the starting point of timbre saliency.

Acknowledgments

This work was funded by grants from the Canadian Natural Sciences and Engineering Research Council (NSERC) and the Canada Research Chairs (CRC) program to Stephen McAdams and by the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) to Song Hui Chon. A special thanks to the members of the Music Perception and Cognition Lab for discussions in the design of the experiment.

6. REFERENCES

[1] S. H. Chon and S. McAdams, "Investigation of timbre saliency, the attention-capturing quality of timbre," *Proceedings of the Acoustics 2012 Hong Kong (Invited Paper)*, 2012.

[2] (2011) Vienna symphonic library. Vienna Symphonic Library GmbH. [Online]. Available: http://vsl.co.at

[3] S. Winsberg and G. De Soete, "A latent-class approach to fitting the weighted euclidean model, clascal," *Psychometrika*, vol. 58, pp. 315–330, 1993.

[4] S. H. Chon and S. McAdams, "Exploring blending as a function of timbre saliency," *Proceedings of the 12th International Conference of Music Perception and Cognition*, 2012.

[5] G. J. Sandell, "Roles for spectral centroid and other factors in determining "blended" instrument pairings in orchestration," *Music Perception*, vol. 13, pp. 209–246, 1995.

[6] D. Tardieu and S. McAdams, "Perception of dyads of impulsive and sustained sounds," *Music Perception*, vol. 30, no. 2, pp. 117–128, 2012.

[7] D. Huron, "Voice denumerability in polyphonic music of homogeneous timbres," *Music Perception*, vol. 6, no. 4, pp. 361–382, 1989.

[8] P. Iverson, "Auditory stream segregation by musical timbre: Effects of static and dynamic acoustic attributes," *Journal of Experimental Psychology*, vol. 21, no. 4, pp. 751–763, 1995.

[9] C. Bey and S. McAdams, "Postrecognition of interleaved melodies as an indirect measure of auditory stream formation," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, no. 2, pp. 267–279, 2003.

[10] A. H. Gregory, "Listening to polyphonic music," *Psychology of Music*, vol. 18, pp. 163–170, 1990.

[11] G. Peeters, P. Susini, N. Misdariis, B. L. Giordano, and S. McAdams, "The timbre toolbox: Extracting audio descriptors from musical signals," *Journal of the Acoustical Society of America*, vol. 130, no. 5, pp. 2902–2916, 2011.

[12] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *Journal of the Acoustical Society of America*, vol. 61, pp. 1270–1277, 1977.

[13] J. M. Grey and J. W. Gordon, "Perceptual effects of spectral modifications on musical timbres," *Journal of the Acoustical Society of America*, vol. 63, pp. 1493–1500, 1978.

[14] C. L. Krumhansl, "Why is musical timbre so hard to understand?" in *Structure and Perception of Electroacoustic Sound and Music*, S. Nielzen and O. Olsson, Eds. Amsterdam: Excerpta Medica, 1989, pp. 44–53.

[15] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, and J. Krimphoff, "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," *Psychol Res*, vol. 58, pp. 177–192, 1995.

[16] S. Lakatos, "A common perceptual space for harmonic and percussive timbres," *Perception & Psychophysics*, vol. 62, no. 7, pp. 1426–1439, 2000.

[17] A. Caclin, S. McAdams, B. K. Smith, and S. Winsberg, "Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones," *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 471–482, 2005.

[18] (2012) Finale. MakeMusic, Inc. Eden Prairie, MN.

[19] (2012) Logic. Apple Computer. Cupertino, CA.