WHERE DO YOU WANT YOUR EARS? COMPARING PERFORMANCE QUALITY AS A FUNCTION OF LISTENING POSITION IN A VIRTUAL JAZZ BAND

Adriana Olmos
Centre for Intelligent Machines
McGill University
aolmos@cim.mcgill.ca

Paul Rushka
Schulich School of Music
McGill University
paul.rushka@mail.mcgill.ca

Doyuen Ko
Schulich School of Music
McGill University
doyuen.ko@mail.mcgil.ca

Gordon Foote
Schulich School of Music
McGill University
gordon.foote@mcgill.ca

Wieslaw Woszczyk
Schulich School of Music
McGill University
wieslaw@music.mcgill.ca

Jeremy R. Cooperstock
Centre for Intelligent Machines
McGill University
jer@cim.mcgill.ca

ABSTRACT

This study explores the benefits of providing musicians with alternative audio rendering experiences while they perform with a virtual orchestra. Data collection methods included a field study with a large jazz band and a pilot study in which musicians rehearsed using a prototype that presented two different audio rendering perspectives: one from the musician's perspective, and a second from the audience perspective. The results showed that the choice of audio perspective makes a significant difference in some musicians' performance. Specifically, for some musicians, e.g., lead trumpet players, an acoustically natural mix results in improved performance, for others, e.g., drummers, it was easier to play along with the artificial "audience" perspective. These results motivate the inclusion of a music mixer capability in such a virtual rehearsal scenario.

1. INTRODUCTION

Ensemble rehearsal is a demanding activity for musicians, one in which they must deal not only with the complexities of their own part, but also, coordinate with the performance of other musicians. It is challenging for many musical groups to find sufficient opportunities for the entire ensemble to practice together. This challenge led to our work on the Open Orchestra Project, which simulates the ensemble rehearsal experience, using *both* high-definition video and high-resolution audio, rendered from the perspectives of individual instrumentalists. In other words, the musician sees the conductor and relevant part of the orchestra on a panoramic video display, and hears the rest of the orchestra, with his or her own part removed. The result combines the experience of ensemble rehearsal with the convenience and flexibility of solo study.

In normal ensemble rehearsal and performance, musicians see and hear the other instruments depending on their physical location within the orchestra. For example, in a large jazz band, a lead trumpet, surrounded by other trumpet players and positioned directly behind the trombones, primarily hears the brass section. The result, both in terms of relative loudness and arrival times of the sounds from the various instruments, is very different from the experience of a lead alto saxophone player, or for that matter, of an audience member. In the context of ensemble training with the Open Orchestra system, we consider whether it is better for the musicians to practice with the sounds of the other instruments reproduced in this natural manner, from their intended position, or as a more balanced mix, along the lines of that produced for a commercial recording. Specifically, we are interested in determining which option is preferred by the musician, which option is considered more realistic, and how this choice impacts the quality of the musician's performance.

Our initial hypothesis was that although lacking in the aesthetics of the audience experience, an audio image of the orchestra, rendered from the musician's individual perspective, would be the most desirable, since this provides the necessary audio cues to interact with one's closest orchestral neighbours, critical to an effective ensemble performance.

2. RELATED WORK

A variety of previous systems have been developed to present musicians with an experience of performing with an orchestra. One of the best known examples is perhaps "Music Minus One", [1] which consists of prepared recordings of a musical program from which an instrument or voice is missing. A musician may practice and learn the omitted performance, accompanied by the recording of the ensemble, much like karaoke systems. However, systems like this suffer from an absence of visual cues, and a limited control of the ensemble sound, providing only an audio image from a predetermined audience perspective. Another group of such systems supports real-time accompaniment

[1] http://musicminusone.com

by synthesis [1, 2, 3], anticipating the performer's trajectory through a non-improvised musical piece.

Audio spatialization and sound image rendering have long been the subjects of intensive study [4]. Considerable effort has been devoted to software systems that simulate acoustic environments, based on psychoacoustic models related to the perception of sound sources by the human ear [5]. These models led to techniques for sound localization using headphones or a limited number of loudspeakers, typically exploiting interaural level differences (ILD), interaural time differences (ITD), and sound filtering techniques such as reverberation, to recreate the impression of distance and direction.

Numerous examples of such work can be found in interactive games, virtual reality, electroacoustic composition and audio conferencing [6, 7, 8]. Audio spatialization has also been employed in previous systems intended for musical practice and performance. For example, Schertenleib et al. [9] provided music amateurs the opportunity to conduct a group of musicians and produce a new kind of multimedia experience, rendering the orchestra from a central position, with attention to both visual realism and 3D sound rendering.

Other examples of spatialized audio rendering for immersive virtual environments include the work of Naef et al. [10], who optimized their system for rendering moving sound sources in multi-speaker environments using off-the-shelf audio hardware. Wozniewski et al. [11] proposed a framework for creation of perceptually immersive scenes, in which the entire environment functions as a rich musical instrument, with spatialization of sound sources an important element for musical applications. Somewhat closer to a real world reproduction, Martens and Woszczyk [12] accurately mapped and recreated nine virtual rooms in which Haydn's music would have been played. This work was carried out in the context of re-creating a small concert performance as it would have been experienced acoustically in the eighteenth century.

Musicians are influenced by both internal timing variances and external latencies while they try to coordinate their timing [13]. The former are attributed to performer anticipation and delays associated with expressive performance, errors in performance, and random timing variations due to physiological and biological constraints. External latencies are the result of audio propagation through the air as sound travels from the instrument to the performers' ears. Players in a small chamber ensemble typically experience such latency of 5-10 ms, whereas a double bass player could encounter latency of 30-80 ms in the sounds from the percussion section of a large orchestra (e.g., symphony). During their training, musicians develop various techniques that allow them to overcome these latencies, following or isolating certain sounds or instruments in order to coordinate the timing of the musical passage that they are performing. For a simple rhythmic clapping task, Chafe et al. [14] offer a detailed analysis of the effects of such latency. Given his central position with regard to the ensemble, the conductor's role in this process is thus to coordinate the musicians, adjusting not just the tempo, but also ensuring a suitable balance between the different instruments.

Despite the large body of work in the domain of spatialized audio and its importance to music, little attention has been given to alternative audio experiences, or audio rendering perspectives for a musician sitting in a particular orchestral position. Specifically, there does not appear to be any prior work investigating the effects of audio perspective on the musician's performance, that is, whether it makes a difference if the sound is rendered "naturally" from the position of that performer, or from some other perspective such as that of the audience.

3. METHODOLOGY

We investigated the above question through both a field study in the context of orchestral rehearsal, as well as via an experimental study. The former involved observations and recording of the behaviour and actions of the orchestral participants within their work context, without interfering with their activities. These observations were complemented by exposing two of the musicians to an audiovisual "music minus-one" type of system, aiming to elicit conversations between the musicians and the design team. The experimental study involved a pilot experiment in which musicians were asked to rehearse with a prototype built to test two different audio rendering perspectives or conditions, one rendered from the musician's perspective, the other from the audience perspective.

Following a discussion of the musicians' preferences and the quality of their performance, as assessed by a big band Jazz conductor, we discuss the implications of these results in relation to our ongoing work on the design of a virtual orchestral rehearsal system.

3.1 Observing real and virtual rehearsal sessions

To support the exploratory nature of our initial research, a quick ethnographic model was employed in the early stages. This involved fly-on-the-wall [15] observations within a real scenario, and also employed a mock-up prototype of the system that allowed the musicians to rehearse with a recording from a previous rehearsal session. These observations were complemented with conversations with the conductor and musicians. The field study was carried out over a full three-month academic term with the McGill Jazz Orchestra I, an ensemble of 18 students, the most experienced jazz band in the Schulich School of Music. After the fly-on-the-wall observation sessions, our written notes were integrated into a presentation [2] to a user group in order to solicit their feedback. This user group consisted of conductors and professional musicians involved in teaching and mentoring activities at the university level. Although informal, this stage was valuable since it provided a general understanding of the orchestral rehearsal process at the outset of the study. Observations clarified initial assumptions regarding the importance of audio and visual cues and helped inform the design of the pilot experiment,

[2] `http://tinyurl.com/5wgm42b`

described in Section 3.2, intended to explore various audio rendering conditions.

The main findings from the early observations are now summarized. Students in the jazz band rehearsed in three modalities: on their own at home, in groups (depending on the instrument or musical voice they played), and with the full orchestra and conductor.

Quoting a trumpet player from the Jazz ensemble:

> "[In the orchestra] I practice for listening... it is a team work. At home is more for learning the piece... If you focus and play while listening, all the music comes together..."

During the rehearsals, musicians add comments to their own music parts. These typically consist of descriptions or guidelines from the conductor's feedback regarding how to play a music section. When the orchestra is learning a new piece, the conductor might choose to play back a recording of the music piece that they are about to learn. Other times, they start with a session of music reading, and together decide how to play or interpret certain difficult parts of the piece. Some of these instructions make it to the music part in the form of annotations, while others are simply memorized and indicated by the conductor's gestures while performing.

There is no question that playing within the ensemble involves a team effort to interpret the piece as a whole. Audio and visual cues are important elements of the ensemble's rehearsal dynamics. As part of our early observations, we wanted to expose the musicians to the experience of rehearsing with a playback of their previous rehearsal session. This was done in part to prompt them to express their thoughts regarding the concept, as well as to elicit feedback about potential future improvements based on an early prototype.

Although initially skeptical, two of the musicians agreed to spend an hour of rehearsal time with a playback of their previous rehearsal session. To their surprise, the experience proved to be much better than they had expected; significantly, they were able to "pause" the conductor, assimilate his feedback, and repeat a section of music, incorporating the guidelines or instructions provided into their practice. Conversely, during actual "live" rehearsals, these capabilities are not possible.

Of direct relevance to our question of rendering perspective, the musicians immediately identified (by ear) the position from which the recording was made and were able to articulate the differences from what they would normally experience in real life, e.g., hearing more of the lead trumpet. The students also expressed an interest in hearing how they sound with the whole band from an audience perspective, which is not possible from their position in the ensemble.

3.2 Pilot study: comparing two audio image conditions

Based on the observations from the field study, above, we designed our pilot study to address the following questions:

1. While rehearsing with a high-fidelity simulator, which audio image is preferred by the musician, an egocentric perspective or one rendered from the audience position, and what accounts for this preference?
2. Would the preferred audio image be regarded as the most realistic, i.e., in relation to a real-world orchestral environment?
3. How does the choice of audio image rendering affect the performance of the musician?

The pilot study was conducted with eight jazz musicians, four from the McGill Jazz Orchestra I, who played in the recordings used in the prototype, and the remainder from other jazz bands. Both groups consisted of trumpet, trombone, sax and drums players. All the participants were enrolled in a university music program at either the Masters or undergraduate level.

The musicians were exposed to two conditions: an unmodified binaural recording, captured from the musician's perspective, and an "audience" mix, equivalent to that of a central audience member's perspective. The music piece used in this experiment was a recording of the McGill Jazz Orchestra I in Tana Schulich Hall, playing Nestico's "Basie Straight Ahead". Multiple binaural recordings were made using a Neumann KU 100 dummy head with binaural stereo microphones. Each of the musician's position (lead trumpet, lead trombone, lead alto (sax) and drums) was substituted, one at a time, during the band's performance. A Sony PMW-EX3 HD camcorder was located beside the dummy head, facing the conductor, to record the video from the same musician's perspective. For the generation of the audience mix, close microphone placement on all the main orchestra sections (trumpet, trombone, sax, bass, and main) allowed for independent control of the balance of various sections. The mix was created by a sound recording engineer, aiming to produce an audio image from a central audience position, similar to the conductor's perspective. The audio mixes and video were synchronized manually using Final Cut Pro by examining the onsets of the audio waveforms recorded both by the built-in camera microphone and separately by microphones covering the instrument sections. The synchronized audiovisual content was then validated by the conductor of the jazz band.

3.2.1 *Experimental design*

The musicians were asked to rehearse with the orchestral simulator, which was similar to the one presented in Figure 1. The experimental sessions lasted approximately 80 minutes, including a break of 10 minutes at the half-way point. Rehearsals were carried out in four blocks of two trials (eight trials in total). For each block, musicians played the entire song twice, once with the binaural recording and once with the audience mix, with the order of presentations balanced across blocks, conditions, and musicians. The musicians were not informed as to which recording was presented at each trial. At the end of each block, the musicians were given the following questionnaire:

1. Which audio track allowed you to perform to the best of your ability?

Figure 1. Trumpet player rehearsing with an early prototype of the Open Orchestra Project

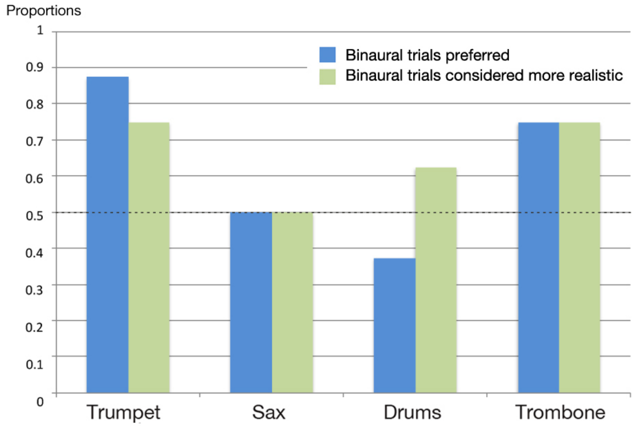

Figure 2. Responses across musicians for preference and perceived realism of the binaural condition.

2. Which track felt more realistic, as related to your experience with a real orchestra?

3. Rate the selected track in terms of realism on a scale of 1 (unrealistic) to 5 (identical to a real orchestra).

4. ANALYSIS AND FINDINGS

The responses to the questions above, along with unsolicited comments, were then analyzed. [3] In addition, audio recordings of the musicians' rehearsals were evaluated by a conductor who was not involved in the original performance.

4.1 Musicians' perceptions

Overall, musicians preferred to rehearse with the binaural recording in 20 of the 32 blocks across all the subjects, but this trend was not statistically significant ($\chi^2(1)$= 2.000, p = 0.15). Independently of which audio condition was chosen, musicians considered their preferred choice to be the most realistic ($\chi^2(1)$= 21.125, $p<0.0001$). This can be observed in Figure 2, where the proportions of binaural preference are presented in response to the first two questions from the questionnaire above. Accordingly, musicians rated the chosen audio recording similarly in terms of realism, regardless of whether it was the binaural (Mean = 3.5, SD=0.49) or the audience mix (Mean=3.6, SD=0.43). This finding could be explained through the concept of processing fluency, which relates to the ease with which information is processed in the mind. Research in psychology has shown that processing fluency influences different kinds of judgments. For instance, perceptual fluency contributes to the experience of familiarity and positive affect [16].

4.2 Expert review

All of the trial recordings were evaluated by a conductor naive to the experimental set up and audiovisual conditions. The conductor was asked to assign a rating from 1 to 5 to each of the parameters of articulation, time feel, pitch (intonation), note shape (inflection), sound quality, dynamic contrast, and rhythm. These parameters were selected based on various conversations with the conductor of the McGill Jazz Orchestra I. To avoid bias, the recordings were rated in a randomized order.

[3] The data from these sessions is available from `http://tinyurl.com/2fnp9ob`.

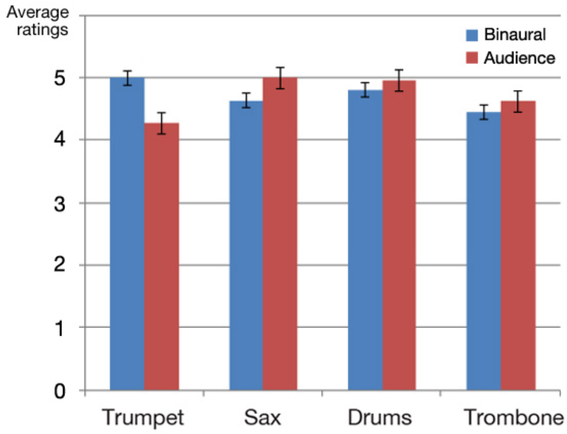

Figure 3. Average ratings per instrument for the binaural and audience audio mix. Error bars indicate the standard error.

A total of 17 out of the 32 blocks were assigned higher average ratings when performing with the binaural recording, but the difference was not statistically significant. Refining the analysis by instrument can be more instructive. As seen in Figure 3, the average performance of the trumpet players proved to be significantly better when rehearsing with the binaural mix (t(7) = 2.938, p = 0.013). The Bonferroni corrected p-value was marginally non-significant. However, the result was consistent with the general preference of the trumpet players for the binaural recording, as indicated in Figure 2.

4.3 Discussion

While the limited number of participants in this pilot study precludes us from stating any strong conclusions, some interesting trends are suggested by the observations. For the lead trumpet in a jazz band, the binaural audio image seems to make a significant difference and helps improve the performance. However, for the lead alto sax players, the lack of clear preference between the binaural and audience mix might be explained by the fact that their sitting position in the band is central, adjacent to the conductor. From that position, they typically receive a more balanced sound, similar to what the audience would hear. The drummers sit to the side of the brass section and mostly hear their own sound, relying on the conductor's gestures for guidance. As one drummer player mentioned "as long as I can hear the bass, I am happy." This might also explain the lack of a clear preference between the binaural and audience mix, given that in both, the bass can be heard clearly. Nevertheless, the drum players were able to identify the audience recording, labelling it a bit less realistic from the experience encountered in a typical orchestral situation because they could hear the band more clearly. Despite its artificiality, these musicians noted that it was easier to play with the audience mix. On the other hand, trombone players were comfortable with either recording but indicated a preference to hear a bit more of the brass section, commensurate with their natural experience.

It is important to mention that the audio conditions were independent of the fixed video display perspective for each instrument. In other words, the video content rendered to the musicians was always acquired from the perspective of that instrumentalist, regardless of whether the binaural or audience audio mix was used. One could argue that this performer-centric video perspective might have influenced the results in favour of selecting the binaural mix as the most realistic, since this is the one with which it was congruent. The motivation for using this same video perspective regardless of the audio environment was to ensure that the musician had visual access to the gestures from the conductor, which were unavailable from the audience perspective. Without such a view of the conductor, the experience would almost certainly have felt less natural. It would be interesting to consider a further audio-only experiment of the two conditions to remove the potential confound of audiovisual congruency from these results.

In any case, the results from this study suggest that the value of providing a dedicated audio image to the musicians, rendered from their own placement within the orchestra, is dependent on the individual instrument. Perhaps even more importantly, providing a mixer capability appears to be desirable. This mixer would provide, as a starting point, a default audio setup that resembles the rendering from the position of the given instrumentalist, but allowing fine tuning of what the musician hears in an ensemble.

5. SUMMARY AND FUTURE WORK

The body of work presented here investigated the idea of providing musicians with different audio experiences while performing with a virtual orchestra. Two audio rendering perspectives were presented to a group of eight jazz musicians, one at the time, sitting in a particular orchestral position (lead trumpet, lead alto sax, lead trombone, drums). We investigated the effects on the musician's performance while rehearsing with both an audio perspective provided "naturally" from the position of that performer and that from the audio perspective of an audience member.

While there was a slight preference towards the audio experience rendered from the musician's perspective, this was not significant across all instruments tested. However, there was a significant difference in the performance by the lead trumpet players while rehearsing with the audio rendered from their perspective. These results seem to suggest that the value of providing a dedicated audio image to the musicians is dependent on the individual instrument position. Our ongoing work is examining the customization of these audio parameters for a given musician based on recommendations from a mentor or conductor. We expect that this approach will foster an interesting learning environment in which the musician could practice and improve his skills while performing within an orchestra context.

Future work will involve a larger study including other genres of music and expanding the number of participants, as well as the number of expert reviewers or conductors assessing performance of the musicians. As noted above, we are also interested in the experimental outcome of an audio-only presentation of the two renderings, without the potential confound of a video perspective that is congruent with only one of the conditions.

6. ACKNOWLEDGEMENTS

The research described here was funded under a Network Enabled Platforms (NEP-2) program research contract from Canada's Advanced Research and Innovation Network (CANARIE). The project is being developed in collaboration with colleagues at the Centre for Interdisciplinary Research in Music, Media and Technology (CIRMMT) at McGill University. In particular, the authors would like to thank Antoine Rotondo and Nicolas Bouillot for their help in the set up of the recording sessions, Mick Wu for valuable discussions on the experimental design and analysis, the CIRMMT technical staff for their continual assistance, and John Roston, for his important role in this work.

7. REFERENCES

[1] R. Christopher, "Demonstration of music plus one: a real-time system for automatic orchestral accompaniment," in *Proceedings of the 21st National Conference on Artificial Intelligence*. AAAI Press, 2006, pp. 1951–952.

[2] R. B. Dannenberg, "An on-line algorithm for real-time accompaniment," in *Proceedings of the International Computer Music Conference*, 1984, pp. 193–198.

[3] B. Vercoe, “The synthetic performer in the context of live performance,” in *Proceedings of the International Computer Music Conference*, 1984, pp. 199–200.

[4] J. Blauert, *Spatial Hearing: The Psychophysics of human sound localization*. The MIT Press, 1997.

[5] J. Chowning, “The simulation of moving sound sources,” *JAES*, vol. 19, no. 1, pp. 2–6, 1971.

[6] G. Walker, J. Bowskill, M. Hollier, and A. McGrath, “Telepresence: Understanding people as content,” *Presence: Teleoperators and Virtual Environments*, vol. 9, no. 2, pp. 119–136, 2000.

[7] R. Kilgore and M. Chignell, “The vocal village: Enhancing collaboration with spatialized audio,” in *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*. Washington, DC, USA: AACE, 2004, pp. 2731–2736.

[8] W.-G. Chen and Z. Zhang, “Highly realistic audio spatialization for multiparty conferencing using headphones,” in *IEEE International Workshop on Multimedia Signal Processing*, 2009.

[9] S. Schertenleib, M. Gutierrez, F. Vexo, and D. Thalmann, “Conducting a virtual orchestra,” *IEEE Multimedia*, vol. 11, no. 3, pp. 40 – 49, 2004.

[10] M. Naef, O. Staadt, and M. Gross, “Spatialized audio rendering for immersive virtual environments,” in *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*. ACM, 2002, pp. 65–72.

[11] M. Wozniewski, Z. Settel, and J. R. Cooperstock, “A framework for immersive spatial audio performance,” in *New Interfaces for Musical Expression*. IRCAM - Centre Pompidou, 2006, pp. 144–149.

[12] W. Martens and W. Woszczyk, “Virtual acoustic reproduction of historical spaces for interactive music performance and recording,” *Acoustical Society of America*, vol. 116, no. 4, pp. 2484–2485, 2004. [Online]. Available: http://link.aip.org/link/?JAS/116/2484/4

[13] C. Bartlette, D. Headlam, M. Bocko, and G. Velickic, “Effect of network latency on interactive musical performance,” *Music Perception*, vol. 24, pp. 49–59, 2006.

[14] C. Chafe, J.-P. Céceres, and M. Gurevich, “Effect of temporal separation on synchronization in rhythmic performance,” *Perception*, vol. 39, no. 7, pp. 982–92, 2010.

[15] IDEO, *IDEO Method Cards: 51 Ways to inspire Design*. William Stout Architectural Books, 2003.

[16] R. Reber, P. Winkielman, and N. Schwarz, “Effects of perceptual fluency on affective judgments,” *Psychological Science*, vol. 9, pp. 45–48, 1998.