A STUDY OF ENSEMBLE SYNCHRONISATION UNDER RESTRICTED LINE OF SIGHT

Bogdan Vera, Elaine Chew
Queen Mary University of London
Centre for Digital Music
{bogdan.vera,eniale}@eecs.qmul.ac.uk

Patrick G. T. Healey
Queen Mary University of London
Cognitive Science Research Group
ph@eecs.qmul.ac.uk

ABSTRACT

This paper presents a quantitative study of musician synchronisation in ensemble performance under restricted line of sight, an inherent condition in scenarios like distributed music performance. The study focuses on the relevance of gestural (e.g. visual, breath) cues in achieving note onset synchrony in a violin and cello duo, in which musicians must fulfill a mutual conducting role. The musicians performed two pieces – one with long notes separated by long pauses, another with long notes but no pauses – under direct, partial (silhouettes), and no line of sight. Analysis of the musicians' note synchrony shows that visual contact significantly impacts synchronization in the first piece, but not significantly in the second piece, leading to the hypothesis that opportunities to shape notes may provide further cues for synchronization. The results also show that breath cues are important, and that the relative positions of these cues impact note asynchrony at the ends of pauses; thus, the advance timing information provided by breath cues could form a basis for generating virtual cues in distributed performance, where network latency delays sonic and visual cues. This study demonstrates the need to account for structure (e.g. pauses, long notes) and prosodic gestures in ensemble synchronisation.

1. INTRODUCTION

In ensemble performance, musicians rely on a complex mixture of non-verbal communication via visual and auditory gestures (such as breathing) and the inherent timing information present within the acoustic signal of the performance. Together, these cues contribute to the musicians' common perception of musical time, and allows them to synchronize one with another. In certain cases, such as in distributed music performance, some of these cues are disrupted by factors such as network latency, which has been shown to affect synchronisation between musicians and delay video transmissions so much so as to make visual gestures ineffective. We are, therefore, interested in understanding the effects of disruptions of visual communication on ensemble performance, so as to advance research on assistive systems for distributed performance.

This paper presents a study on the effect of line-of-sight restriction between musicians working to synchronize onsets and interact in a performance. The remainder of the paper is organized as follows: Section 2 reviews related work in ensemble interaction and networked performance; Section 3 describes the experimental design; Section 4 presents the analysis; results and conclusions follow in Sections 5 and 6.

2. LITERATURE REVIEW

Musical gesture analysis has been an area of interest for researchers in music cognition, music performance, and human-computer interaction. McCaleb [1] compares ensemble interaction to the communication paradigm (likened to a telephone or postal service), acknowledging that this approach has not yet been critiqued from the perspective of a performing musician.

Leman and Godoy [2] performed a classification of musical gestures, distinguishing gestures that are part of sound production from those that are purely communicative and those which simply accompany music (such as dancing). Lim [3], as a step towards creating a robotic accompaniment system for flute, identified start and end cues which serve to visually mark the onsets and offsets of notes, and beat cues which are used to keep time during sustained notes, all of which were motion based. Eye contact has also been discussed as being important in ensemble synchronisation [4]. Breathing, as a musical gesture, has been touched upon by Vines et al. [5], and mentioned as an important cue in conversation, where it helps in coordinating turn taking.

In the context of network music performance, the effects of audio latency itself have been studied by researchers such as Chafe and Gurevich [6], Chew et al. [7] and Schuett [8]. This research shows that when latencies higher than around 25 ms are present, the tempo tends to decrease and synchronisation is adversely affected. It is not understood what effects visual isolation has on the performance in these cases, though the DIP project reports iniitial explorations of this area with attempts to provide distributed musicians with visual cues via video streaming [9]. They found that video latencies are to be too high for the trans-

mitted gestural cues to be of much use, especially over long distances. Solutions to the problem of latency based on prediction were theorized by Chafe [10], and Sarkar's TablaNet project [11] predicts tabla drum players' strokes ahead of time in order to create the appearance of zero transmission latency. Further work on predicting drum strokes has recently been done by Oda et al. focusing on estimating the velocity of drum mallets with high speed cameras and predicting their impact times [12]. Combining these ideas with Lim's approach of predicting gestures, we hypothesize that the issue of latency in video transmission could be ameliorated using predictive modeling of gestural cues.

3. EXPERIMENTAL SET-UP

Two simple violin-cello duet pieces [1] were composed by Vera for the experiment, and were played by a violin and cello duo under three different line of sight conditions. The participating musicians were both classically trained, and active in chamber orchestras, but had never played together before. The first condition, S1, involved normal performance with no line of sight obstruction, with the musicians located in their preferred positions. In the second scenario, S2, the musicians were made to face in opposite directions, removing line of sight, but allowing auditory gestures such as breathing. In the third scenario, S3, a translucent curtain was placed between the musicians, and their shadows were cast onto it by two bright studio lamps, allowing them to view only each other's silhouettes, with no fine details such as facial expression (see Figure 1). The musicians were asked to play the chosen pieces for the first time, with little rehearsal time.

The pieces were specially designed with a focus on key experimental features. The first piece is relatively easy to play (i.e. the musicians were not expected to greatly improve over time and they required almost no rehearsal to play it). It consists of a sequence of very long notes (lasting two bars at a moderate tempo) followed by equally long pauses. In the middle of the piece, the pauses are replaced by faster rhythmical dialogues between the performers, before returning to the long separations. The aspect explored in this piece is the timing between the musicians at the beginning of each note, where they have to cue each other into a new section without relying on rhythmic information, the only exception being the middle section where the fast paced rhythms are expected to improve synchronisation. The main hypothesis in this case is that lack of visual contact would result in greater asynchrony between the onsets of simultaneously sounded notes, and that asynchrony would be reduced where timing information is carried by rhythmic patterns in the music itself.

The second piece is a similarly slow paced composition, but without pauses. After four bars of solo cello, the two instruments play simultaneous notes, until later in the piece where some counterpoint is introduced between the parts. In this case, our hypothesis was that the presence of a stronger rhythm and lack of pauses will result in less asynchrony, compared to the performance of the first piece, when line of sight is affected.

The musicians were recorded playing 3 takes of each piece (four in the case of the 'no line of sight' recordings, due to extra time at the end of the recording sessions), in each scenario, over two days. Due to time constraints, the musicians played through each scenarios in sequence, and thus some improvement over time is expected as the musicians became accustomed to playing the pieces. Both instruments were recorded with attached pickups, in an attempt to isolate the two instruments as much as possible.

[1] Scores available at http://tinyurl.com/nqhq2pp

4. ANALYSIS

A quantitative analysis of the difference in time between the note onsets of the performers' simultaneously sounded notes was performed, comparing their performances in the three scenarios. As obtaining reliable note onsets from bowed instruments is difficult with automatic methods, the onsets were hand annotated using Sonic Visualiser [13]. Even when annotating onsets by hand, it can be difficult to determine the exact onset time of soft notes. Notes played by bowed instruments have varying attack times, and one can, for example, choose either the start of the unpitched bowing sound or the moment when a fundamental frequency becomes audible. In this case, the annotation focused on the latter feature, using Sonic Visualiser's adaptive spectrogram to inspect the notes. The resulting set of onset time differences was then analyzed in Matlab, simply by subtracting the onset times of the violinist from those of the cellist for simultaneous score notes.

4.1 First Piece: Long Notes, Long Pauses

For this first piece, the onset annotations for an example recording are shown in Figure 2. In this case the onset annotations separate the piece into four-bar segments. No fast section onsets were considered for the initial analyses. Fast section annotations marking two bar long sections were later added to inspect the effects of rhythmic vs. non rhythmic patterns on synchronization, and they were treated as secondary to the longer notes, allowing a more focused comparison between the onset times of the long notes separated by pauses, and those linked by rhythmic sections.

Asynchrony analysis for the three scenarios is visualized in the box plots in Figure 4, showing the extent of asynchrony, which we define as the unsigned time difference between the onsets of ideally simultaneously sounded notes, from all the recordings in each scenario. The results show a median asynchrony of 52.1 ms in the normal line of sight scenario. This increases to 104.8 ms in the no line of sight scenario, showing a worsening of synchrony. In the partial line of sight scenario, the median asynchrony was 46.3 ms, which is slightly lower than in the normal line of sight scenario. Table 1 contains the p-values of pairwise Kolmogorov-Smirnov tests between the scenarios, showing that the scenario with no line of sight was

Figure 1. The two musicians on either side of the shadow curtain in the partial LoS scenario

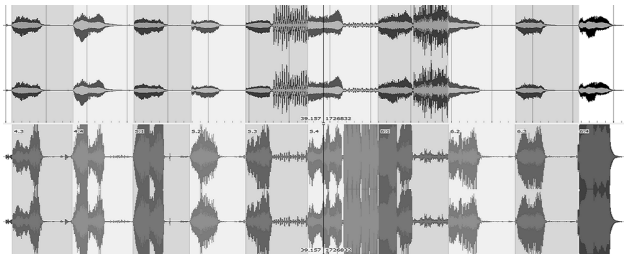

Figure 2. Example segmentations for one take of the first piece (top - cello, bottom - violin)

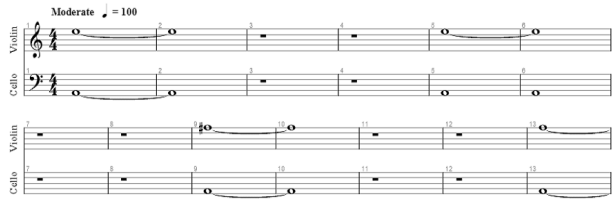

Figure 3. Excerpt of the first piece showing the long notes separated by pauses before the rhythmical section

significantly different from the others, and that the normal and partial line of sight conditions were not significantly different. Figure 5 shows the box plots of onset time differences, without taking the absolute value. The analysis showed that the violinist tended to play ahead of the cellist (i.e. most values are positive). Figure 6 shows median absolute onset time difference per segment, for each scenario. From this graph it is notable that for segments 6, 7 and 8 – the segments linked by rhythmic patterns – the musicians seem to have achieved better synchrony than in the rest of the piece.

Scenario Pair	S1 vs S2	S1 vs S3	S2 vs S3
P-Value	0.0222	0.9360	0.0205

Table 1. Pairwise Kolmogorov-Smirnov p-values between asynchronies in each scenario for the first piece

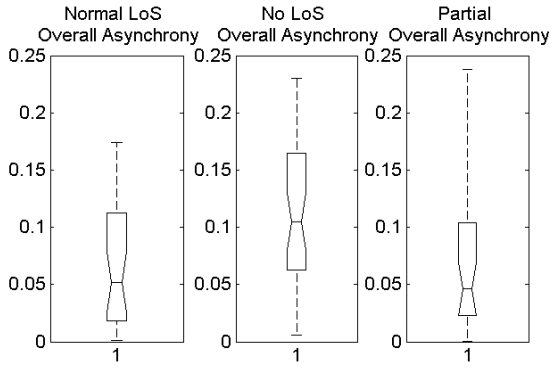

Figure 4. Asynchrony boxplots for each scenario for the first piece (y-axis is in seconds)

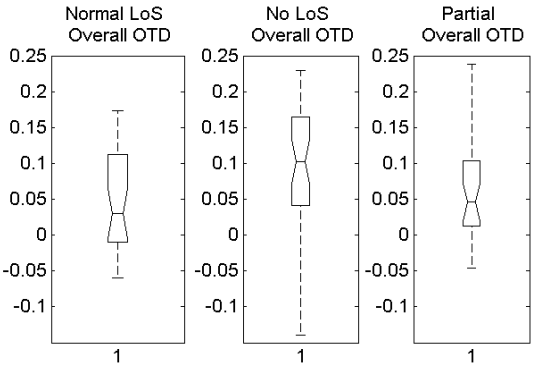

Figure 5. Onset time difference boxplots for each scenario for the first piece (y-axis is in seconds)

4.2 Second Piece: Long Notes, No Pauses, Counterpoint

The same analysis was performed for the second piece. In this case the segments examined were chosen to correspond with all note onsets. Because the violin and cello parts contain many notes that do not have simultaneous onsets, segmentation points from each part were replicated in the other part by automatically choosing time points at appropriate note subdivisions between adjacent segmentations. This provided a set of estimated segmentations based on available data ensuring that both parts have com-

Scenario Pair	S1 vs S2	S1 vs S3	S2 vs S3
P-Value	0.6055	0.0234	0.1072

Table 2. Pairwise Kolmogorov-Smirnov p-values between asynchronies in each ccenario for the second piece

parable segmentation points.

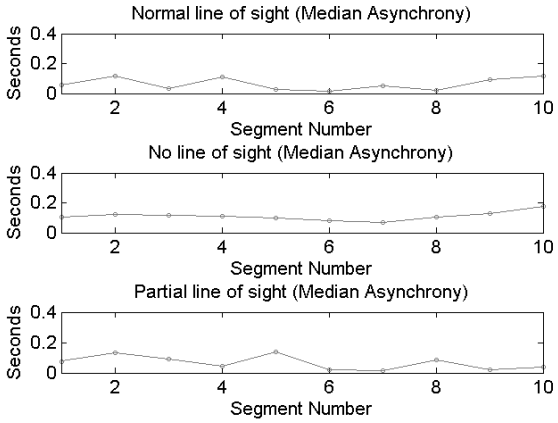

Figure 6. Asynchrony against segment number for the first piece

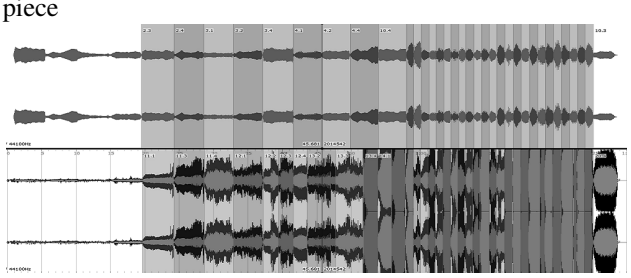

Figure 7. Segmentations for the second piece (top - cello, bottom - violin)

Although comparing a real onset with an estimated one does not give a precise value for note onset synchrony, it does however give an indication of the performers' degree of synchronisation to each other's timing, i.e. a note played by the cellist that starts in the middle of the violinists' held note would have its onset time compared to the calculated middle of the violin's two closest adjacent note onsets. The first four notes, which are played only by the cellist are not annotated. The segmentation points (before the addition of estimated segmentations) are shown in Figure 7.

Unlike for the previous piece, the median asynchrony decreases with each scenario, indicating that the effect of the musicians getting better at playing the pieces was more significant than that of reduced line of sight. The median asynchrony was 80 ms for the baseline scenario, 74.7 ms for the second, and 59.6 ms for the third. The paired Kolmogorov-Smirnov test results, presented in Table 2, show that there was no significant worsening caused by reduced line of sight. We instead see a significant improvement of synchrony between the first and last scenarios.

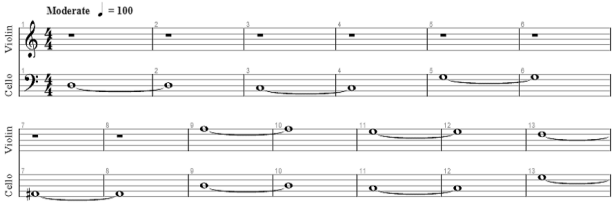

Figure 8. Excerpt of the second piece showing long notes (w/o pauses) in the cello joined by long notes in the violin

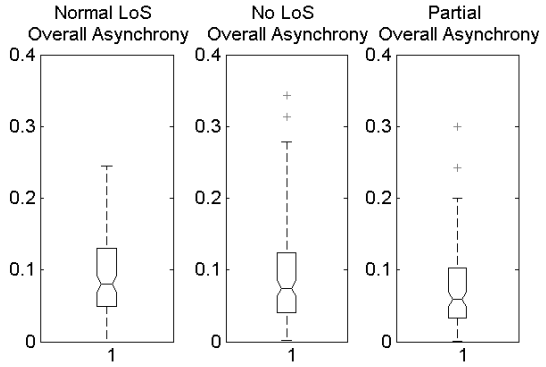

Figure 9. Asynchrony boxplots for each scenario for the second piece (y-axis is in seconds)

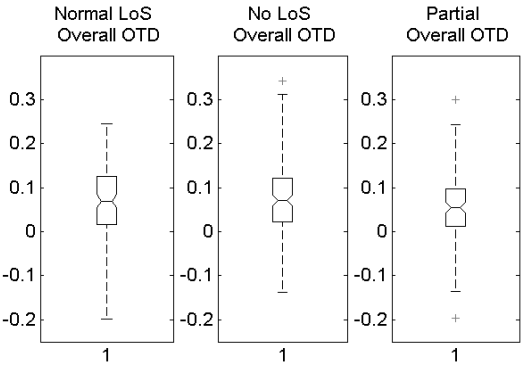

Figure 10. Onset time difference boxplots for each scenario for the second piece (y-axis is in seconds)

4.3 Use of Breath for Cueing

In the recordings of the first piece, it was notable that the violinist took highly regular and audible breaths before each note following a pause, which the musicians identified as important cues. To investigate the use of breath, an extra set of annotations was created, marking the start and end times of each breath, picked by investigating the spectrograms of the recordings. However, as the breaths do not have clear onsets or offsets, this data may be noisy and usable only at a fairly coarse level. An example annotated breath sound is shown in Figure 12. This is a task that could possibly be automated, for example using an algorithm like the one presented by Ruinskiy and Lavner [14].

The first feature of interest was the set of breath start times as a ratio of pause length, or how far into the pauses did the breaths tend to start. Another problem in this case

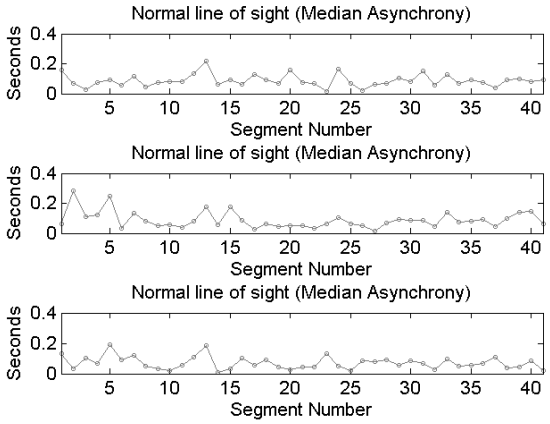

Figure 11. Asynchrony against segment number for the second piece

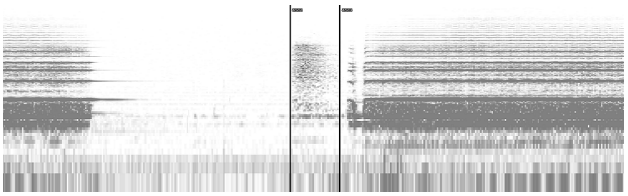

Figure 12. Example of annotated breath sound on spectrogram

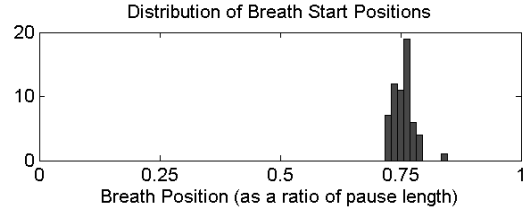

Figure 13. Histogram of breath position.

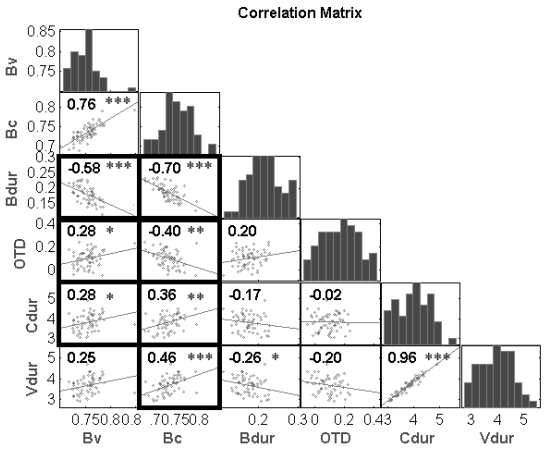

Figure 14. Correlation table of breath and pause variables (stars indicate significance levels: $^{*}0.05, ^{**}0.01, ^{***}0.001$)

was that finding the start times of the pause segments is difficult as string instruments do not always have distinct note offsets. Annotation of these offsets was based on finding the point where the higher harmonics start to decay, as the fundamental often had a much longer decay, and the musicians often left one string resonating through the pause itself. Pause start times were then taken to be the average offsets of the cellist and violinist's previous notes. Breath position with respect to the violinist's pause offset (i.e. note onset) was then expressed as a value between 0 and 1 representing how far along a pause a breath occurs, as shown in Equation 1:

$$B_{violin,i} = \frac{2b_{i,0} - c_{i,0} - v_{i,0}}{2v_{i,1} - c_{i,0} - v_{i,0}} \quad (1)$$

where i is the pause index, $b_{i,0}$ is the breath onset time, $c_{i,0}$ is the cello's pause onset time, and $v_{i,0}$ and $v_{i,1}$ are the violin's pause onset and offset times, respectively. We correspondingly define B_{cello} as the breath position with respect to the cello's pause offset:

$$B_{cello,i} = \frac{2b_{i,0} - c_{i,0} - v_{i,0}}{2c_{i,1} - c_{i,0} - v_{i,0}}. \quad (2)$$

Figure 13 shows the histogram of breath start positions for all pauses from all recordings. The violinist mostly used breath gestures at the 0.76 point, which closely corresponds to the last half note in the 2-bar pause. To better understand the effect of the breath cues, regression and correlation analysis was performed. This is shown in Figure 14, where Bv and Bc represent B_{violin} and B_{cello}, and Bdur is the duration of the breath as a ratio of the violinist's pause length; OTD is the difference between the violin and the cello onset times, and Cdur and Vdur are the durations of the violinist's and cellist's pauses (a value greater than zero means that the violinist played before the cellist).

From this analysis it is notable that the breath position, despite its small range of variation, had a significant effect on the onset time difference. Later breath positions with respect to the violin's pause had a slight tendency to correspond with positive differences (meaning that the violinist started first), possibly by indicating a later start time to the cellist and making her start later. The correlation between the violinist's breath position with respect to the cello pause time and the onset time difference was much more significant, and the two variables are inversely related, higher values (nearer 0.75) essentially lowering the gap between the musicians' onsets. We also see a very significant inverse correlation between the breath's start position and the length of the breath sound. We also see a positive correlation between the length of the pause (from both musician's perspectives) and the position of the breath along the pause, meaning that in longer pauses the breath started later. The correlations of interest are emphasized in Figure 14. These characteristics identify breath as a type of cue that could be used in a networked scenario to predict a performer's intent to begin a note, serving as the basis for synthesis of virtual cues that can be sent ahead of time to bypass latency, in a manner similar to the rhythm prediction in Sarkar's TablaNet project [9].

5. RESULTS

The results of this study suggest that line of sight is important in achieving good synchronization in a string duo, especially when the music being played contains pauses during which the musicians cannot easily track time. As the partial line of sight scenario did not cause a significant decrease in synchrony, it appears that very simple body motion was sufficient for effective gestural cueing. In scenarios with restricted line of sight, performers can rely on non-visual and extra-musical cues such as breath for synchronization. In this study, the leading musician issued breath cues that were synchronized to their own perception of musical time, and served as advance warnings of note onset intent. The following musician then used this cue to estimate the beginning of the next note. Small variations in breath onset within pauses were correlated with variations in note onset time delay, suggesting that musicians pay close attention to these cues, and that mis-communication of timing by breathing too soon or too late can have direct consequences on synchronization.

When the music had no pauses and contained counterpoint and rhythm, the musicians did not exhibit worse synchronization in the absence of visual contact, suggesting that auditory cues embedded in the music itself were sufficient for synchronization.

6. CONCLUSIONS

The findings of this study indicate a need for further research into the fine dynamics of cues that are transmitted both visually and sonically. Auditory features of interest may be variations in dynamics or pitch in relation to critical synchronization points. Due to the improvement seen in the partial line of sight scenario, we propose that visual cues are likely more dependent on general motion than on eye contact, facial expression, or other such fine details.

Further work should focus on obtaining a larger dataset for study, although the main difficulty is obtaining multitrack, acoustically isolated recordings done in controlled conditions. We intend to continue the study of ensemble synchronization by including visual and motion tracking data in our analysis, in order to discover the most important types of visual gestures and their relationship with the music being performed.

7. ACKNOWLEDGEMENTS

The authors thank Laurel Pardue and Dr. Kat Agres for participating in the experiment. This project was funded in part by the Engineering and Physical Sciences Research Council (EPSRC).

8. REFERENCES

[1] M. McCaleb: "Communication or Interaction? Applied environmental knowledge in ensemble performance," *Proceedings of the CMPCP Performance Studies Network International Conference*, 2011.

[2] R. I. Godoy, M. Leman: *Musical Gestures Sound, Movement and Meaning*, Routledge, 2010.

[3] A. Lim: "Robot Musical Accompaniment: Real-time Synchronization using Visual Cue Recognition," *Proceedings of the IEEE/RSJ International Conference on Intelligent RObots and Systems (IROS)*, 2010.

[4] A. Williamon: "Coordinating Duo Piano Performance," *Proceedings of the Sixth International Conference on Music Perception and Cognition*, 2000.

[5] B. W. Vines, M. M. Wanderley, C. L. Krumhansl, R L. Nuzzo, D. J. Levitin: "Performance Gestures of Musicians: What structural and emotional information do they convey?" In A. Camurri, G. Volpe (eds.): *Gesture-Based Communication in Human-Computer Interaction*, LNCS 2915, Springer, 2004.

[6] C. Chafe, M. Gurevich "Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry," *Proceedings of the 117th AES Convention*, San Francisco, 2010.

[7] E. Chew, A. Sawchuk, C. Tonoue, R. Zimmerman "Segmental Tempo Analysis of Performances in User-Centered Experiments in the Distributed Immersive Performance Project," *Proceedings of the Sound and Music Computing Conference*, 2005.

[8] N. Schuett "The Effects of Latency on Ensemble Performance," Undegraduate Honors Thesis, Stanford University, 2009.

[9] A.A. Sawchuk, E. Chew, R. Zimmermann, C. Papadopoulos and C. Kyriakakis "From Remote Media Immersion to Distributed Immersive Performance," *Proceedings of the ACM SIGMM 2003 Workshop on Experiential Telepresence*, 2003.

[10] C. Chafe "Tapping into the Internet as a Musical/Acoustical Medium," *Contemporary Music Review*, 2009.

[11] M. Sarkar "TablaNet: a real-time online musical collaboration system for Indian percussion," S.M. Thesis, MIT, 2007.

[12] R. Oda, A. Finkelstein and R. Fiebrink "Towards Note-Level Prediction for Networked Music Performance," *Proceedings of the 13th International Conference on New Interfaces for Musical Expression*, 2013.

[13] C. Cannam, C. Landone, M. Sandler "Sonic Visualiser: An Open Source Application for Viewing, Analysing, and Annotating Music Audio Files," *Proceedings of the ACM Multimedia 2010 International Conference*, 2010.

[14] D. Ruinskiy, Y. Lavner "An Effective Algorithm for Automatic Detection and Exact Demarcation of Breath Sounds in Speech and Song Signals," *IEEE Transactions On Audio, Speech, and Language Processing, Vol. 15*, 2007.