Simulation of Networked Ensemble Performance with Varying Time Delays: Characterization of Ensemble Accuracy

Michael Gurevich, Chris Chafe, Grace Leslie, and Sean Tyan
Center for Computer Research in Music and Acoustics
Department of Music, Stanford University
gurevich@ccrma.stanford.edu

Abstract

The conditions of networked ensemble performance were simulated in an experiment. Pairs of musicians were placed apart in isolated rooms and given a simple rhythm to clap together. A microphone was placed as close as possible to each performer's hands. Each monitored the other's sound via headphones, and a delay was introduced between the source and listener. Starting tempo, given by a recorded count-in, and delay time were varied across trials. Recordings of the trials were analyzed with a precise event detection algorithm to locate clap onsets, from which the tempo was inferred. The rate of deceleration increased with longer delays, while shorter delays ($\leq$ 11.5 ms) produced a modest, but significant acceleration. The goal is to identify the region of delay time that is most conducive to maintaining a steady tempo. This will help to determine the necessary delay conditions to support networked musical performance (which may be over long distances or in adjoining rooms). Humans performed significantly better than a simple model of a memoryless instantaneous reaction.

1 Introduction

Increasingly, members of music ensembles are separated in space, either in isolated rooms in a studio complex, or in separate buildings or cities. Physical separation implies propagation delay, and music depends on temporal precision. Meaningful rhythmic inflections can occur on the order of 10 ms (Iyer 1998), which places strong demands on communication systems that support ensemble performance. In the case of long-distance network transmission, and now often in studios, audio signals are transmitted digitally between ensemble members, incurring some routing, processing and transmission delays. As an example Internet2 can deliver bidirectional stereo audio between San Francisco and New York in about 75 ms round trip (Chafe et al. 2000). In studio settings, digital mixing consoles and computers may have throughput latencies on the order of tens of microseconds to tens of milliseconds (MacMillan, Droettboom, and Fujinaga 2001). This poses the challenge of enabling ensemble presence in the design of networks to support music collaboration. A similar problem has existed for some time with respect to speech communication in telephony (Krauss and Bricker 1967), and has recently resurfaced with the advent of IP telephony.

Remote ensemble performance can be conceived in terms of a "virtual Internet music room", in which performers can hear one another and interact naturally while apart. In order to support musical interaction, the acoustics of this virtual room must reflect some of the same properties as real spaces. An important property of real rooms is that the propagation delay of one musician's sound to another ensemble member is normally low. As the propagation delay between performers increases, rhythmic accuracy and the ability to play together deteriorate, until it is impossible to maintain a performance. This can be demonstrated by trying to clap a steady rhythm over mobile phones. Our study attempts to quantify this effect in a restricted context that represents the most basic elements of a musical performance. Subjects were given a simple rhythm to clap in an acoustically damped, isolated environment. Real music in real rooms might indeed yield different results, but we chose to begin with a more general case in which we can be relatively confident that musical context is not a significant factor. A pilot study by earlier members of our group demonstrated that tempo deceleration increased with longer delay, and surprisingly found that tempo tended to accelerate with very short delay. This prompted the present study, which examines the problem in greater detail, using different techniques.

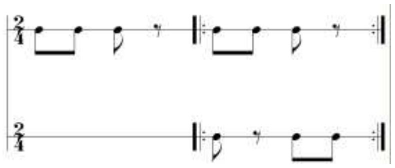

Figure 1: Duo clapping rhythm used in experiment.

2 Method

The experiment tested ensemble accuracy under controlled one-way delay times of 0 - 77 ms. Tempo consistency, which was known to suffer under long delays was used as a measure of ensemble accuracy. The simple, repeating, interlocking rhythm shown in Fig.1 was used in order to simplify the task, enable relatively easy tempo analysis and remove stylistic or contextual effects. 3 nearby starting tempi (86, 90, 94 bpm) were used to account for dependence of the results on absolute tempo. 17 pairs of subjects performed the task, which consisted of 18 trials in one session. Of the 18 trials, 12 are included in this analysis. The others represented more sparsely sampled conditions of asymmetric delays and different tempi. The experimental protocol is explained in greater detail in Chafe et al. (2004).

2.1 Population and task

Subjects were students and staff at Stanford University. No qualification regarding musical performance ability was stipulated and no subjects were excluded in advance. A questionnaire revealed a range of 0 to 40 years of musical training with a mean of 10.4 years. Subject pairs were formed randomly. At the beginning of the experiment, subjects were given the rhythm in music notation, listened to assistants perform it, and were allowed to practice face-to-face until they felt comfortable. Their task was to "keep the rhythm going evenly" during each trial, and they were not given a strategy or any hints about how to do that. After being placed in separate rooms, the computer-controlled experiment began. Trials were presented in random order, advanced manually by an assistant, and retakes were allowed if a trial was interrupted. For each trial, one randomly chosen subject was presented with a starting tempo (a calibrated series of recorded claps) and the other heard nothing until the initiator began to clap. Inter-room monitoring was automatically shut off after 36 sec to signal the end of a trial.

2.2 Experiment configuration

The experiment was carried out under controlled acoustical conditions. The two subjects were placed in acoustically isolated rooms, and could not see each other or the test administrators. Sound-isolating headphones were used, and identical microphones were placed at a fixed distance of 0.3 m from the chairs. See Fig. 2. A single Linux PC provided recording, playback, delay and the GUI-based experiment interface. Delay times were calibrated with analog oscilloscope measurement. 0 ms delay was achieved with an analog bypass. Each trial was recorded as a 16 bit, 96 kHz stereo sound file, with the two direct microphone signals synchronously captured on separate channels.

2.3 Analysis

Each session had 12 trials at different levels of delay, with a starting tempo randomly selected from 86, 90 or 94 bpm. The 12 delay levels followed the sequence:
d_n (ms) $= n + 1 + d_{n-1}$, where $d_0 = 0$. Initial analysis measured the dependence of tempo consistency on delay time and starting tempo. The automated analysis proceeded independently on each channel.

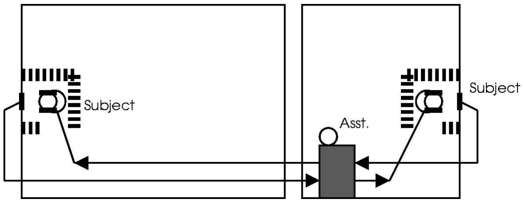

Figure 2: Subjects clapped to each other from separate rooms through computer-controlled delays.

Event detection The "amplitude surfboard" technique (Schloss 1985) was used because of its high time-resolution ability (accurate to within ± 0.25 ms in this case) and consistency on percussive attacks. The algorithm first found an amplitude envelope by recording the maximum dB amplitude in successive 50-sample windows, while preserving the sample-index of each envelope point. A 7-point linear regression (the "surfboard") estimated the slope at every envelope sample. Samples with high slope are likely to be event onsets. Candidate events are local maxima in the vicinity of samples with slopes that fall within some threshold of the maximum slope. In the event of several candidates in close proximity, the one with the highest amplitude was chosen. After an event was identified, there was a "dead period", in which another cannot occur. Inter-onset intervals (IOI's) were calculated from the event times.

Quarter note expansion and event pruning To convert IOI's to instantaneous tempo samples, each quarter note was expanded into two eighth notes. The position of the "new" event in the middle of a quarter note was linearly interpolated. At the same time, spurious events were pruned, defined by an uncharacteristically short or long IOI. Differentiating real from spurious events and eighth notes from quarter notes was not straightforward in the presence of deceleration. A one-pole filter adaptively tracked the tempo "inertia" in order to dynamically define the spurious event - eighth note- quarter note bounds. The output of this process was an instantaneous tempo time series, with two samples per quarter note, and $\mathbf{tempo} = (f_s * 60/\mathbf{IOI})/2$. The factor of 2 represents the fact that IOI's are between eighth notes. See Fig. 3.

Linear regression The slope, $b_{\hat{t}}$ of a linear regression through the merged tempo time series was used as a measure of acceleration over the course of the trial. The variance of the residual of the regression, s^2 was used as a measure of tempo jitter.

$$s^2 = \frac{\sum(\mathbf{t} - \hat{\mathbf{t}})^2}{n-1} \tag{1}$$

where $\mathbf{t}$is a tempo time series and $\hat{\mathbf{t}}$is the linear regression.

3 Results

Out of the 17 sessions, 2 were discarded due to an inability to consistently perform the rhythm. ANOVA indicated that for these 2 trials the tempo jitter $\bar{s}_i^2$ $(i = 1, 2, ...17)$ was

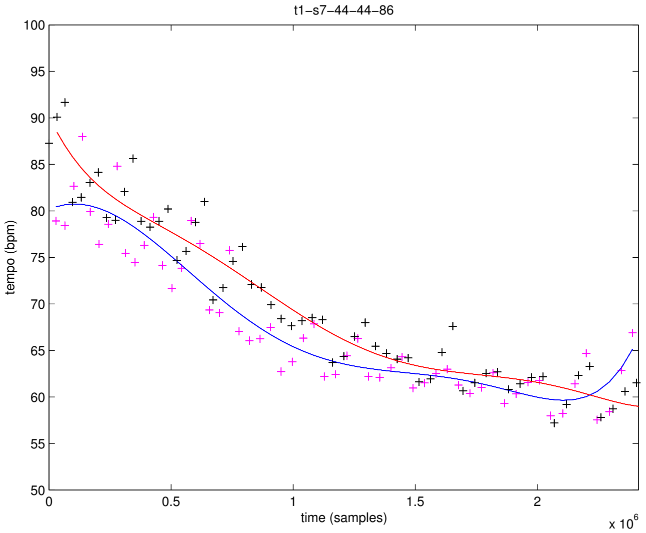

Figure 3: Tempo curves of subjects clapping together for one recorded trial, delay = 44 ms, starting tempo = 86 bpm (6th-order regression through each subject's tempo time series).

significantly different from all others ($p = 1.0 \times 10^{-8}$). A total of 173 trials are included in the analysis: 15 valid sessions, 7 individual trials discarded because the task was incomplete.

3.1 Dependence of tempo on delay time

After $b_{\hat{t}}$ and s^2 were calculated for each trial, the mean, variance and 95% confidence intervals were calculated for each delay-tempo combination, as well as for each delay at all tempi combined. The mean of the tempo slopes m_{acc}, for each delay at all 3 tempi, shown in Fig. 6a, shows a strongly linear relationship between m_{acc} and delay, with $r^2 = 0.98$. The negative slope of the linear model in Eq. 2 confirms that tempo tends to decelerate with increasing delay.

$$\hat{y} = 0.58 - 0.05x + \epsilon \tag{2}$$

The positive y-intercept reproduces the earlier finding of acceleration at very low delay. Indeed, at 0, 2 and 5 ms delay, the mean and entire 95% confidence intervals were positive, indicating significant acceleration. At 20 ms and above, the mean and entire confidence intervals were negative, indicating significant deceleration. The theoretical best delay, where $\hat{y} = 0$ is 11.5 ms. For delays shorter than this, 74% of the performances sped up. At delays of 14 ms and above, 85% slowed down.

3.2 Significance testing

ANOVA and multiple comparisons of the mean tempo slope m_{acc} at each of the 3 starting tempi revealed no significant difference between 3 cases ($p = 0.25$), ruling out a dependence on absolute tempo. ANOVA of m_{acc} grouped by delay time showed that the 3 significantly accelerating cases of 0, 2, and 5 ms all have statistically different means from the first significantly decelerating case of 20 ms. A full two-way ANOVA testing all 36 delay-tempo combinations yielded no significant interaction between tempo and delay. There were no cases where the marginal m_{acc} at a given delay and starting tempo was significantly different from that at the same delay and a different tempo. Grouping pairs according to mean or minimum musical experience yielded 2 pairs which had significantly aberrant tempo jitter. These were the only pairs in which one member had no experience; the same two that were exluded from the tempo slope analysis above.

3.3 Modeling

A simple model of the human performer in this situation is as a memoryless system that perfectly detects tempo, and therefore incorporates the perceived delay as a decrease in tempo. This model performer has no knowledge of the tempo before the most recent beat. It can react instantaneously and clap at a perfect tempo. Tempo, M would therefore decrease in bpm according to Eq. 3, as depicted in Fig. 4

$$M(n) = 60/(T_0 + nd) \tag{3}$$

where $T_0 = 60/M_0$ is the starting period in seconds, n is the quarter-note number and d is the delay time. A delay of $d = 0$ would produce a perfectly steady tempo, and any delay $d > 0$ would cause deceleration according the curves in Fig. 5a.

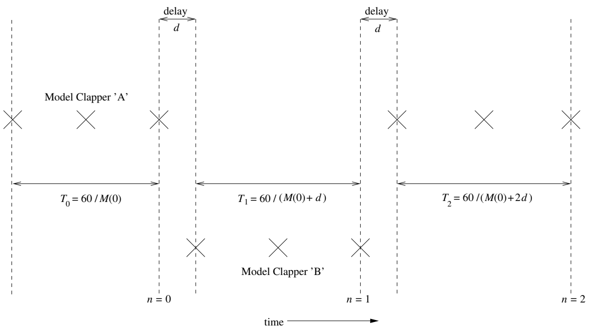

Figure 4: Theoretical sequence of claps according to the model. 'B' waits for 'A' who waits for 'B', ...

Subjects performed much better than this model. A nonlinear least-squares fit of each trial's merged tempo time series to the model in Eq.3 found the equivalent delay d_{eq} that would make the model most closely resemble the performance. If humans performed as the model predicted, we would get $d_{eq} = d$, the actual delay, for all cases. The number of quarter notes was estimated based on the number of events in the tempo time series vector. As expected, given the previous analysis of tempo slope, Fig. 6b shows that $\bar{d}_{eq}$, the mean of d_{eq} across all trials at a given delay d begins negative and increases almost monotonically with d. The extent to which humans outperformed the model, however, is surprising. $\bar{d}_{eq}$ ranged from -0.6 ms to 4.6 ms, meaning that under the worst delay condition of 77 ms, humans performed as well as the memoryless, instantaneous-reaction model at just 4.6 ms delay. The relationship between $\bar{d}_{eq}$ and delay time is also strongly linear ($r^2 = 0.96$).

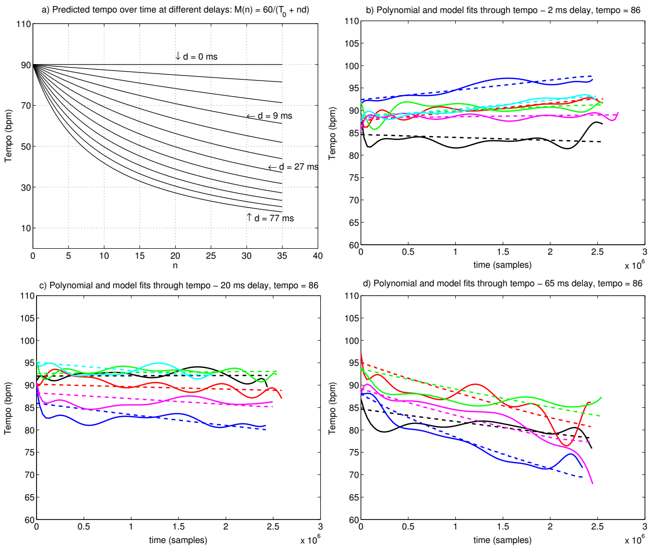

Figure 5: a) Predicted tempo over time at each tested delay time according to the model in Eq. 3. b-d) Best fits to polynomial(solid) and model(dashed) for all trials at progressively longer delays (2, 20, 65 ms) at starting tempo 86 bpm.

4 Discussion

The results in Fig. 6 demonstrate that human performers cope with delay remarkably well. In spite of outperforming the model, the degree of deceleration in the high-delay cases ($\geq$ 35 ms) was still beyond the limits of what could be considered musically acceptable. The linear relationship of both d_{eq} and m_{acc} to delay time implies that there is some specific "best" delay time, rather than a wide range of conducive conditions as we expected. This concept might coincide with tuned-oscillation theories of rhythmic perception(McAuley 1995). Comparison of these experimental data to a more thorough neurological or perception-reaction model will hopefully shed more light on the processes at work, and in turn inform the design of networking applications for collaborative audio. On a higher level, a more systematic study of the effectiveness of different coping mechanisms might indicate some of the specialized cognitive adaptations of musicians.

Continuing with the analogy of a "virtual room" for network delay, we know that orchestral performers separated by distances of 10m (comparable to ˜35 ms propagation delay) can maintain a steady tempo (with a conductor). Further study is needed to quantify these phenomena in real rooms. Two obvious differences between real and virtual spaces are the availability of nearly instantaneous visual communication in real rooms, and the acoustics of a real space. While the first is an interesting area of research, it is not likely a problem that can be addressed in networking, where the delay times for video will be at least as long as those for audio. The second, however, suggests the possibility of implementing characteristic "virtual acoustics" that relate to network delays to facilitate ensemble performance.

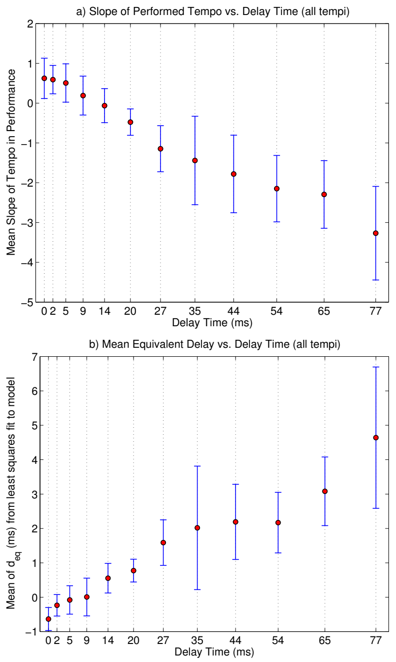

Figure 6: a) Mean tempo slope m_{acc} as a function of delay. b) Mean equivalent delay d_{eq} as a function of delay.

5 Acknowledgments

Thanks to Nathan Schuett, the 2001 pilot study team, and Jay Kadis for his help in the studio. This research was supported in part by a grant from Stanford's Media-X program.

References

Chafe, C., M. Gurevich, G. Leslie, and S. Tyan (2004). Effect of time delay on ensemble accuracy. In *Proceedings of the International Symposium on Musical Acoustics,*, Nara, Japan.

Chafe, C., S. Wilson, R. Leistikow, D. Chisholm, and G. Scavone (2000). A simplified approach to high quality music and sound over IP. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFX-00)*, Verona, Italy.

Iyer, V. S. (1998). *Microstructures of Feel, Macrostructures of Sound: Embodied Cognition in West African and African-American Musics*. Ph.D. Thesis, Univ. of Cal. Berkeley.

Krauss, R. M. and P. Bricker (1967). Effects of transmission delay and access delay on the efficiency of verbal communication. *Journal of the Acoustical Society of America 41*(2), 286–292.

MacMillan, K., M. Droettboom, and I. Fujinaga (2001). Audio latency measurements of desktop operating systems. In *Proceedings of the International Computer Music Conference*, pp. 259–262. International Computer Music Association.

McAuley, J. D. (1995). *Perception of time as phase: Toward an adaptive-oscillator model of rhythmic pattern processing*. Ph.D. Thesis, Indiana Univ.

Schloss, W. A. (1985). *On the automatic transcription of percussive music from acoustic signal to high level analysis*. Ph.D. Thesis, STAN-M-27, CCRMA, Stanford Univ.