Audio Engineering Society

Convention Paper

Presented at the 117th Convention
2004 October 28–31 San Francisco, CA, USA

Network Time Delay and Ensemble Accuracy: Effects of Latency, Asymmetry

Chris Chafe[1], Michael Gurevich[1]

[1] *Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, 94305, USA*

Correspondence should be addressed to Chris Chafe (cc@ccrma.stanford.edu)

ABSTRACT

Pairs of musicians were placed apart in isolated rooms and asked to clap a rhythm together. Each person monitored the other's sound via headphones and microphone pickup was as close as possible. Time delay from source to listener was manipulated across trials. Trials were recorded and clap onset times were measured with an event detection algorithm. Longer delays produced increasingly severe tempo deceleration and shorter delays (¡ 11.5 ms) produced a modest, but surprising acceleration. The study's goal is to characterize effects of delay on rhythmic accuracy and identify the region most conducive to ensemble playing. The results have implication for networked musical performance. Network delay is a function of transmission distance and / or internetworking (routing) delays. The findings suggest that sensitive ensemble performance can be supported over rather long paths (e.g., San Francisco to Denver at about 20 ms, one-way). The finding that moderate amounts of delay are beneficial to tempo stability seems, at first glance, counterintuitive. We discuss the observed effect.

1. INTRODUCTION

Music ensembles can be traced through artifacts as far back as Paleolithic cultures (bone pipes have been found in proximity dating more than 30,000 years old)[**?**]. Making music (with instruments, singing or clapping) is a fundamental group activity which takes place outdoors, indoors, and increasingly electronically. Meaningful rhythmic inflections at the ¡ 10 ms level [**?**] make ensemble timing one of the most discerning and one of the most demanding applications in telecommunications. Its communicative power rests on exquisite temporal precision. Preserving ensemble presence is a challenge in the design of "network pipes" for collaboration and requires a better understanding of constraints imposed by delay and bandwidth. Placing performers so that they can hear one another and interact naturally while apart means building "next-generation Internet music rooms" with acoustics as conducive as those developed through centuries of practice in

venues ranging from stadiums to studio booths.

Rhythmic accuracy deteriorates as delay increases to a point beyond which performing together becomes impossible. Our study of this effect has been carried out in a restricted context – a simple groove, with simple sounds and controlled acoustics. The experiment recreates a pilot study (by earlier members of our group) which confirmed greater tempo deceleration resulting from increased time delay. Very low delay, on the other hand, produced tempo acceleration. The unexpected discovery of "beneficial" delay, in which short delay amounts had the best tempo stability, suggested the importance of nailing down a precise role for delay, prompting our present work which has been designed using different controls, techniques and more closely-spaced trials in the near-zero delay region.

2. METHOD

Recent Internet "telemusic" tests have spanned continents and oceans, proving that ensembles can perform convincingly despite signal latencies in a range from 20 - 60 ms (one-way). The following experiment describes a test of ensemble accuracy in the range of 0 - 77 ms designed to study end-to-end latency (signal delay) as an independent variable. The longest equivalent path in air would be approximately 20 m.

An unadorned musical context with only a simple interlocking rhythm was chosen so that conclusions about ensemble accuracy might be drawn directly from an analysis of tempo consistency. 17 pairs of subjects (duos) were recorded while clapping the rhythm in Fig. 1 under varied time delays. Recordings were processed automatically with an event detection / tempo tracking algorithm and sessions from 15 duos were deemed viable for further analysis. Potential interaction of the results with absolute tempo was accounted for by performing trials at 3 nearby tempi (86, 90, 94 bpm).

2.1. Population and Task

Subjects were students and staff at Stanford University. A portion of the group was paid with gift certificates and others participated as part of a course in computer music. No qualification regarding musical performance ability was stipulated and no subjects were excluded in advance. Individuals in the pool were paired up randomly into duos.

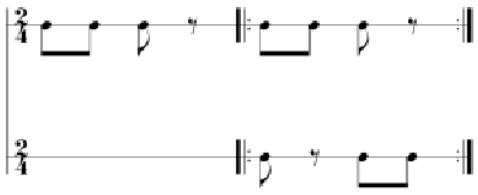

Fig. 1: Duo clapping rhythm used in experiment.

Assistants provided an instruction sheet and read it aloud. Subjects could read the rhythm from the handout and listen to the assistants demonstrating it. Initially, duos practiced face-to-face. They were told their task was to "keep the rhythm going evenly" once it started, and they were not given a strategy or any hints about how to do that. After they felt comfortable clapping the rhythm together, they were assigned to adjacent rooms designated "San Francisco" and "New York."

Each duo performed 18 trials. For the initial and final trials of each session, the subjects clapped against a recording of metronome-guided clapping rather than each other. These trials provide calibration and a measure of subject consistency over a session.

For all other trials, one subject was randomly chosen to initiate the rhythm and their partner heard nothing until the initiator began to clap. Trials followed a computer-controlled sequence: **1)** room-to-room audio monitoring switches on; **2)** a voice recording (saying "San Francisco" or "New York") plays only to the respective initiator; **3)** an isolated metronome (5 sec recording of clapped beats at the new tempo) plays to the initiator; **4)** initiator starts rhythm at will; **5)** partner joins in at will; **6)** after a total of 36 secs room-to-room monitoring shuts off, signaling the trial's end. Assistants advanced the sequence of trials manually after each take was completed. Short breaks were allowed and a retake was made if a trial was interrupted.

2.2. Acoustical and Electronic Configuration

Acoustical conditions minimized room effects and extraneous sounds (jewelry, chair noise, etc.). Subjects were located in two acoustically-isolated rooms

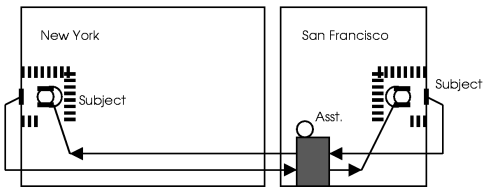

Fig. 2: Subjects clapped to each other from separate rooms through computer-controlled delays.

(CCRMA's high-quality recording and control room pair). Seated in opposite positions and facing apart, they were surrounded by sound absorbing partitions, Fig. 2. One microphone (Schoeps BLM3) was located 0.3 m in front of each chair. Its monaural signal fed both sides of the opposite subject's headphone (isolating headphones, Sennheiser HD280 pro, reduced headphone leakage to microphones and glasses wearers were required to remove their frames to enhance the seal).

A single computer provided recording, playback, adjustable delays and the automated experimental protocol with GUI-based operation. The setup comprised a Linux PC with 96kHz audio interface (M-Audio PCI Delta 66, Omni I/O). Custom software was written in C++ using the STK set of open-source audio processing classes which interface to a real-time audio subsystem. All delays were confirmed with analog oscilloscope measurement. Absolute 0 ms delay through the system was obtained via an analog bypass around the audio interface.

Each trial was recorded as a stereo, 16bit, 96kHz sound file. The direct microphone signals from both rooms were synchronously captured to the two channels. A database of the recordings is being maintained on a networked server for continuing analysis. Sessions are indexed by a code system to preserve subject anonymity.

2.3. Trials

Three experiments were embedded in the trials and were randomly interspersed. Experiment 1 investigates the effect of symmetrical time delay within a narrow range of tempi. Delays were varied in 12 steps according to the sequence $d_n = n + 1 + d_{n-1}$ and were presented in random order. Each duo performed each condition once. Starting tempo per trial was randomly selected from 3 prerecorded "metronome" tracks of clapped beats at 86, 90, and 94 bpm.

Experiment 2 tests the effects of delay asymmetry (2 trials) and Experiment 3 the influence of starting tempo across a wide range (2 trials), 60 to 120 bpm. Overall, one session took about 25 minutes to complete.

2.4. Measurement of Tempo Consistency

Sound files in the database were analyzed to measure tempo consistency as a function of delay and as a function of starting tempo. An automated procedure detected and time stamped true claps, and stored inter-onset intervals (IOI's) as an instantaneous tempo time series, Fig. 3. Detection proceeded per subject (one audio channel at a time). These individual series were merged to track a duo's tempo change.

Candidate events were detected using the "amplitude surfboard" technique[?], tuned to measure onsets to an accuracy of ±0.25 ms. The extremely clean clapping recordings allowed false events (usually spurious subject noises) to be rejected using simple amplitude thresholding. A single threshold coefficient proved suitable for the entire group of sessions.

Conversion from IOI to tempo in bpm (by combining two eighth-notes into one quarter-note beat) was ambiguous in the presence of severe deceleration and required that very slow eighth-notes be distinguished from quarter notes by adaptively tracking tempo "inertia."

Two parameters are of interest: tempo slope $b_{\hat{t}}$, the slope of a linear regression through the merged time series (a measurement of acceleration), and tempo jitter s^2, defined as variance of the residuals of the linear regression.

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

where $\mathbf{t}$ is a vector of IOI's and $\hat{\mathbf{t}}$ is the linear regression.

3. ANALYSIS

3.1. Qualification of Trials

Of the 17 sessions, 2 were discarded because of an

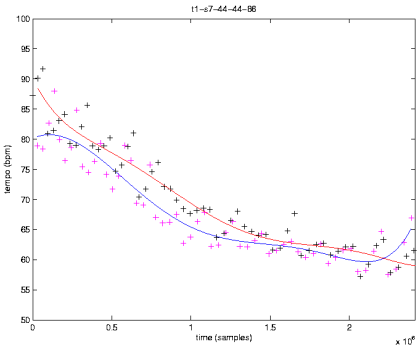

Fig. 3: Tempo curves of subjects clapping together for one recorded trial, delay = 44 ms, starting tempo = 86 bpm (6th-order regression through each subject's IOI time series).

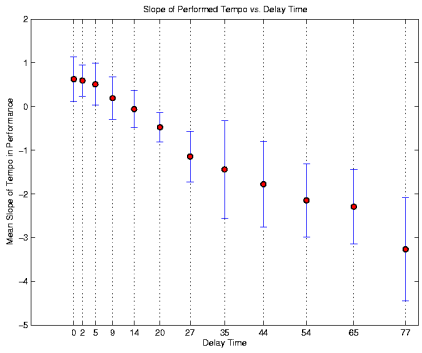

Fig. 4: Mean tempo slope as a function of delay (all trials).

inability to perform the clapping rhythm. Lack of competence was judged subjectively and was confirmed by high tempo jitter. ANOVA and multiple comparisons of the mean tempo jitter of each session $\bar{s_i^2}$ $(i = 1, 2, ...17)$ revealed a significant difference between the 2 discarded sessions and 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).

3.2. Tempo consistency as a Function of Delay and Starting Tempo

Fig. 4 reveals an orderly relationship of m_{acc}, the mean of tempo slopes $b_{\hat{\mathbf{t}}}$, across the range of sampled delays. Expressed as a linear model

$$\hat{y} = 0.58 - 0.05x + \epsilon \qquad (2)$$

it has the expected negative relationship and confirms the non-zero y-intercept found in our previous study. The model fits extremely well, $r^2 = 0.98$.

An ANOVA of the tempo slope at each of the 3 starting tempi confirmed that there was no significant difference between performances at the tempi presented $(p = 0.25)$. A full two-way ANOVA of tempo slope grouped by delay time and starting tempo (36 combinations - 12 delays and 3 tempi) revealed no significant interaction between starting tempo and delay. That is, 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.

3.3. Trial Characterization

Even given the metronome which played at the start of each trial, starting tempi varied noticeably as can be seen in Fig. 5 which shows tempo curves for 4 conditions (using 12th-order regressions through each merged series). Some performances departed from the rest (Fig. 5c) and the longest delays usually yielded high variance (Fig. 5d).

4. DISCUSSION

Model (2) is in agreement with a negative linear relationship between tempo acceleration and delay. And it indicates the existence of an ensemble accuracy sweet spot, d_{best} (where $m_{acc} = 0$) at 11.5 ms. For delays shorter than this, 74% of the performances sped up. At delays of 14 ms and above, 85% slowed down. No correlation with starting tempo was found in the range sampled.

Long delays create an impossible situation in which two performers must stay globally-synchronized while operating from two different "time zones." Something has to give, and that discrepancy goes into slowing down the beat. One way to think of the problem is "player A, waiting for player B, waiting for player A..." in a recursive drag on tempo. Theoretically, that produces a decreasing sequence in bpm of $60/(T_0 + n(d - d_{best}))$, starting from an initial period T_0, and delay d with a leading slope which is more severe than we found even in the worst

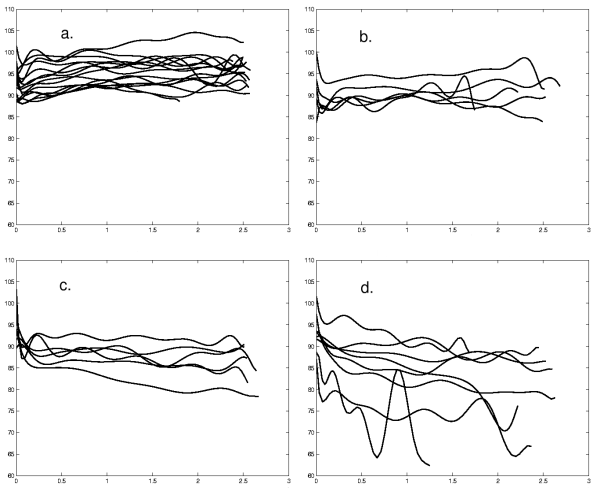

Fig. 5: All trials for progressively longer delays (0, 2, 27, 77 ms) at starting tempo 90 bpm.

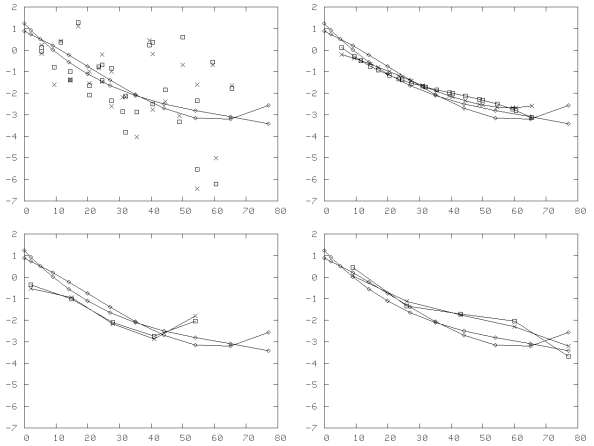

Fig. 6: Asymmetry, a) asymmetric acceleration slopes projected at mean of delays and 3rd-order fit of all symmetric slopes (NY and SF) b) fit of asymmetric slopes at mean of delays c) projected at shorter delay d) at longer delay

case trials. A close look at the performances reveals why this isn't the whole story. The interplay seen in our recordings suggests that players are often anticipating and pushing back on the drag, or intermittently ignoring one another. We would expect structures resulting from musical "messaging" or expressiveness, and mechanisms of attending and production to also contribute to observable flux. With regard to the perceptual "engine" involved in this interplay, Large and Palmer describe human tempo tracking as an attentional function with its own dynamics and uneven temporal profile [**?**].

We searched for a possible bias favoring tempo evenness since it was an instruction we had charged the subjects with. This would manifest itself as a wider sweet spot or tendency pulling more data into the x-intercept $m_{acc} = 0$. The strongly linear model suggests that no bias exists, and we can only speculate at the reasons. Bias toward tempo stability would have implied that the duos were able to perform like a self-correcting system (e.g., coupled in the form of a negative-resistance oscillator). The model instead crosses a sharply-tuned resonance at d_{best}, suggesting some as yet unidentified (human) time constant is dominating the experiment.

5. FURTHER EXPERIMENTS

5.1. Delay Asymmetry

Asymmetrical delays were included in two trials of each session and a total of 29 valid recordings were collected. Asymmetric ratios in a range of 1.2 - 4.5 were created by choosing from the above sequence of delays, with short side delays in the range of 2 - 54 ms and long side in the range of 9 - 77 ms.

Total sampling yielded one or two trials per condition across subjects and is thick enough for observation but too sparse for predictive analysis.

Fig. 6 compares symmetric and asymmetric conditions. All four graphs, a) - d), show the same 3rd-order fit of all previous symmetric data divided as a pair of piece-wise curves showing all "NY" clappers and all "SF" clappers. The contrasting asymmetric data was also divided accordingly. Fig. 6 a) shows all asymmetric acceleration slopes projected at the mean values of their two delays and b) shows a 3rd-order fit of the same data divided into curves for NY and SF. Fig. 6 c) shows the two curves obtained by instead projecting the data onto the short-side delay, and d) onto the long-side delay.

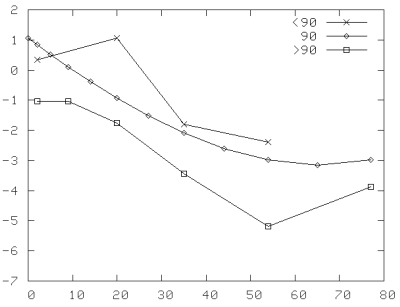

Fig. 7: Mean tempo slope for all nominal-90 bpm trials and tempo trials above and below 90 bpm.

Two observations can be made. First, that side-to-side difference is less influential for either symmetric and asymmetric conditions than is the difference between the conditions. Since asymmetric configurations were organized with the shorter of the two delays always on the same side, the two curves represent relative length. However, this difference does not appear influential and leads to the hypothesis that clapping ensembles stayed locked despite asymmetry.

Secondly, the portions of projections which most closely resemble the symmetric condition are the low-valued delays of d) and the higher values of a). The hypothesis which this suggests is that the longer delay of an asymmetric pair dominates in the shorter region (up to about 25 ms) and in the longer region the effective delay most closely matches the median of the two delays. Both hypothesis require further experimental sampling to test their significance.

5.2. Wider Tempo Range

Two trials per session were included covering a wider range of starting tempi (60,80,100,120 bpm) at a sparser set of symmetrical delays (2, 9, 20, 35, 54, 77 ms). Valid results were obtained from 22 recordings, 8 of them below 90 bpm and 14 above. Fig. 7 summarizes the result: the nominal-90 bpm data (86 - 94) from above is shown as one curve and the wide tempo data as two curves, one with tempi ¡ 90 bpm (60, 80) and the other with tempi ¿ 90 bpm (100, 120). Comparison in this fashion shows that slower tempos sped up relative to the rest and faster ones slowed down, but that overall acceleration vs. delay follows a similar inverse trend.

6. RELEVANCE TO NETWORKS AND COLLABORATIVE AUDIO

The observed optimal one-way delay $d_{best} = 11.5$ ms equates with a physical radius of 2,400 km (assuming signals traveling at approximately 70% the speed of light and no routing delays). An ideal audio network would have similarly symmetrical, constant delays through which audio data is always delivered intact. Closed networks do meet such criteria, but long-distance networks with Internet Protocol (IP) routing often result in asymmetry, jitter and packet loss, and further work is required to understand these effects.

The headphone configuration ensured that our performers did not hear their own echoes as recirculating sound and thus had no sense of the delay as it changed from trial to trial (in fact, some were prone to blame tempo deviations on their partner's shortcomings rather than on the effect). In contrast, CCRMA's real-world tests have avoided headphones and allowed audio feedback. Closed-loop monitoring allows the parties to "hear the delay" and can even use short echo paths to advantage by creating composite or shared "rooms" directly [**?**].

Reverberation almost certainly masks the effect in more natural situations by cushioning sharp-edged signal arrivals. The latencies manipulated in our experiment would in real-life be smeared by multiple acoustical reflections, an effect we intend to address in the future. Apart from reverberation, a time smear "window" on the same order (10's and 100's of milliseconds) is significant in the perception of event order, and thought to be at work on the listener side as a necessity of cognitive function (asynchronous "buffering") [**?**]. The sound of who just clapped that beat first, me or you, may not be as simple a process as it seems. Such an ambiguity would also contribute to flux and complex interplay.

7. ACKNOWLEDGMENTS

Many thanks to our 2001 pilot study team and Nathan Schuett's analysis of the pilot data for his undergraduate Honors Thesis. We gratefully acknowledge grant support from Stanford's Media-X program and the assistance of Jay Kadis.

8. REFERENCES

[1] d'Errico, F. et al., "Archaeological evidence for the emergence of language, symbolism, and music," J. of World Prehistory, 17(1): 1–68, 2003.

[2] Vijay S. Iyer, Microstructures of Feel, Macrostructures of Sound: Embodied Cognition in West African and African-American Musics. PhD Thesis, Univ. of Cal. Berkeley, 1998.

[3] Schloss, W.A., On the automatic transcription of percussive music from acoustic signal to high level analysis. PhD Thesis, STAN-M-27, CCRMA, Stanford Univ., 1985.

[4] Large, E. W. and Palmer, C., "Perceiving temporal regularity in music," Cognitive Science 26: 1 – 37, 2002.

[5] Dennett, D. and Kinsbourne, M., "Time and the observer," Behavioral and Brain Sciences, 15: 183 – 247, 1992.

[6] Chafe, C., "Distributed internet reverberation for audio collaboration", Proc. of the AES 24th Int. Conf., 2003.