Temporal Separation in Duo Rhythmic Performance

Chris Chafe[1], Juan-Pablo Cáceres[1] and Michael Gurevich[2]

[1]Center for Computer Research in Music and Acoustics (CCRMA),

Stanford University, Stanford, CA 94305 USA

[2]Sonic Arts Research Centre (SARC),

Queen's University Belfast, N. Ireland, UK

July 30, 2009

Abstract

Pairs of subjects were placed in separate sound-isolated rooms and asked to clap a rhythm together via headphones and without visual contact. Electronic time delay between the pair was manipulated. The study is motivated by an interest in quantifying the envelope of time delay within which two individuals can produce synchronous performances. The results indicate that "natural" time delay, i.e., delay within the narrow range associated with travel times across room-sized arrangements of groups and ensembles, supports the most stable performance. Conditions outside of this envelope, with time delays both above and below it, create characteristic dynamics in the coupled actions of the duo. Longer delays (corresponding to larger physical distances) produce the expected increasingly severe deceleration and then deterioration of performed rhythms. Looking at extremely short delays (i.e., associated with unnaturally close proximities), indicate that a small amount of delay is in fact required, without

which there is a tendency to accelerate. The envelope has implications for music collaboration over Internet, where with present-day technology stability can be achieved in split ensembles separated by distances of thousands of kilometers.

1 Introduction

Music ensembles produce synchronous performances inside an implicit envelope of time delay. The degree of "temporal separation" refers to the transmission time it takes for the sound from one performer to reach the ears of another performer. As has been shown, increased inter-musician delay introduces a lag which slows down the tempo (**?**). By way of comparison, the tolerance for temporal separation in speaking together over phone or Internet connection is much greater. Natural turn-taking is possible if one-way delays are below $500ms$ (**?**). Coupled rhythmic performance has a different requirement in which it's not alternation but the ability to simultaneously share, hear, and "feel" the beat that counts. In this study of temporal separation, we attempt to quantify the "sweet spot" for duos clapping a rhythm by manipulating the time delays so that we identify the bounds within which satisfactory rhythmic accuracy is possible. Our results have found that "natural" delay, i.e., delay within a narrow range associated with travel times across the usual spatial arrangements of clapping groups, ensembles, etc. supports the most stable performance. We show that, surprisingly, a small amount of delay is in fact required, without which there is a tendency to accelerate. Conditions outside of this envelope, with time delays both below and above it, create characteristic dynamics in the coupled actions of a duo. These are structured according to the degree of temporal separation.

Musicians collaborating via the Internet may find time delay a problem when trying to play together rhythmically (**?**). The first dramatic decrease in telecommunication delays happened in the early 2000's, when various research groups including Stanford University

and McGill University began testing IP network protocols for professional audio use, seeking methods for bi-directional WAN music collaboration . Long distance acoustic delays were now closer to room-sized acoustic delays and ensemble performances began to feel acceptable. The computer systems sent uncompressed audio through high-speed links like Internet2, Canarie, and Geant2 (with higher resolution and faster transmission than standard digital voice communication media like telephone, VoIP, Skype, etc.).

[2] A. X. Xu, W. Woszczyk, Z. Settel, B. Pennycook, R. Rowe, P. Galanter, J. Bary, G. Martin, J. Corey, J. R. Cooperstock, "Real-time streaming of multichannel audio data over Internet," J. Aud. Eng. Soc. 48, pp. 627-641, 2000.

With this capability has come a need to understand what the effects of temporal separation might be. The speed of sound in air causes a delay of approximately $9ms$ in one direction across a string quartet (by measuring the physical distance between the outside players arranged in the typical semi-circle, e.g., approximately $3m$). So, imagine the scenario encountered by two musicians separated by $45ms$ delay trying to play synchronously (they would be approximately 15m apart). In the simplest sense, player A is waiting for the sound of player B, who is waiting for the sound of player A and the tempo slows down from this recursion. Network delay of approximately $45ms$ is also what we encounter between San Francisco and New York using Internet2.

Our present study extends previous work in the field, some of it by our group. A pilot experiment was conducted with the same approach in 2002. The surprising findng of low-latency acceleration prompted us to remake the experiment with greater attention to low delays. This second experiment was then analyzed and confirmed the low-delay effect. However, our analysis failed to identify a "sweet spot" in the higher-delay region. Increasing delay simply correlated with worsening deceleration. Revisiting the data with the analyses presented here, we are now able to quantify a region below worsening deceleration.

[6] C. Chafe and M. Gurevich, "Network time delay and ensemble accuracy: effects of

latency, asymmetry," in Preprint no. 6208, 117th AES convention, San Francisco, CA, USA, Oct. 2004, pp. 2–7.

[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] Gurevich, M. et al., "Simulation of Networked Ensemble Performance with Varying Time Delays: Characterization of Ensemble Accuracy," Proc. of the Intl. Computer Music Conf., 2004.

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

[1] W. Woszczyk, J. Cooperstock, J. Roston, W. Martens, "Shake, rattle and roll: Getting immersed in multisensory, interactive music via broadband networks," J. Aud. Eng. Soc. 53, pp. 336-344, 2005.

"Ecological" test of music collaboration under manipulated delays have been conducted. Two pianists evaluated their ability to perform at delays of $50ms$and greater with the addition of self-delay.

[5] E. Chew, A. Sawchuk, C. Tanoue, and R. Zimmermann, "Segmental tempo analysis of performances in user-centered experiments in the distributed immersive performance project," in Proceedings of the Sound and Music Computing Conference (SMC), Salerno, Italy, Nov. 2005.

Pairs of wind players and string players performed Mozart and rated their ability to perform satistfactorily.

(**?**).

Groups of jazz players.

Carot, thesis, 2009.

The setting we studied consists of a duo of clappers without conductor or other artificial means of coping with delay. The setting is also "ecological," if not musical, in that nearly everyone has the ability to clap rhythmically. The task was simplified as a means of eliminating effects of phrasing and expressive nuance (Bartlette, et al's "internal" effects). If the temporal separation and the ability to hear are satisfactory, then a rhythm will be easy to sustain. In measuring rhythms performed under these "natural" conditions, the tempo should vary slightly around its mean and there should only be a slight amount of onset asynchrony between events intended to be simultaneous (**?**).

2 Experiment

2.1 Overview

We examined performances by pairs of clappers under different delay conditions. A simple interlocking rhythmic pattern was chosen as the clapping task (Fig. 1). The unadorned musical context was chosen so that conclusions about ensemble accuracy might be drawn directly from an analysis of tempo consistency. The rhythm was easily mastered by a pool of subjects formed without regard to musical ability. Subjects were seated apart in separate studios and monitored each other's sound with headphones (and with no visual contact). Delays in the range from $d = 0$ to $77ms$ (one-way) were introduced in the electronic sound path and were randomly varied per trial. The shortest $d = 0ms$ is equivalent to having a subject clapping right next to the other's ears. The longest $d = 77ms$ equates with a separation of approximately $26m$, a distance wider than many concert stages, and longer than the San Francisco–New York Internet path mentioned above. Recordings were processed

automatically with an event detection algorithm, ahead of further processing to extract tempo and synchronization information.

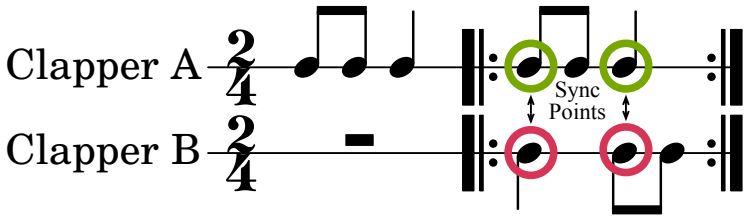

Figure 1: **Duo clapping rhythm used to test the effect of temporal separation.** Subjects in separate rooms were asked to clap the rhythm together while hearing each other's sound delayed by a slight amount. Common beats in the duo clapping rhythm provide reference points for analysis of ensemble synchronization. Circles represent synchronization points.

2.2 Method

2.2.1 Number of subject pairs and trials

17 pairs of subjects (duos) were recorded under various temporal separation conditions. 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, all gave their informed consent according to Stanford University IRB policy. 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.

Each duo performed 18 trials, 12 of which constitute the present experimental data. Sessions from 16 duos were deemed viable for further analysis (see below). Some specific trials where discarded and others with repairable problems were adjusted with manual intervention where possible[1][2], resulting in a total of 163 individual trials available for analysis.

[1]The experiment's database and the source code in MATLAB for this analysis is available at `http://ccrma.stanford.edu/groups/soundwire/research/clapping-experiment/`. The `README.txt` file contains description of files and naming convention.

[2]MATLAB's source file `select_valid_trials.m` contains a list of valid, discarded and repaired trials.

2.2.2 Protocol

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" and they were not given a strategy or any hints to help make that happen. After they felt comfortable clapping the rhythm together, they were assigned to adjacent rooms designated "San Francisco" and "New York."

Starting tempo was established by playing a short clip of recorded clapping at the target tempo. In order to avoid any effects of over-training to one absolute tempo, 3 starting tempi were used in random order (86, 90, 94bpm).

Trials were computer-controlled. Each time a new trial began, one subject was randomly chosen by the trial control process to be the rhythm *initiator*. Trials proceeded according to the following steps:

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 seconds recording of clapped beats at the new tempo) plays to the *initiator*.

4. *Initiator* starts rhythm at will. The other has heard nothing until this point where the initiator begins to clap.

5. The other joins in at will.

6. After a total of 36 seconds, the room-to-room monitoring shuts off i.e., communication is cut, 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.

2.2.3 Acoustical and electronic conditions

Acoustical conditions minimized room reverberation effects and extraneous sounds (jewelry, chair noise, etc.). Subjects were located in two acoustically-isolated rooms (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 meters 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). The distance from clapping hands to microphone introduces a time delay of about $1ms$ and is not added into our reported delays.

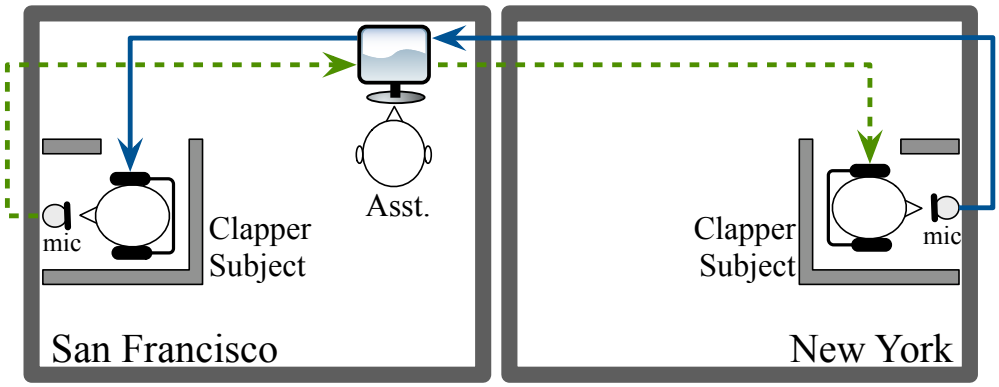

Figure 2: **Experiment setup.** Subjects clapped to each other from separate rooms through computer-controlled delays.

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[3] 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 millisecond 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.

2.2.4 Trials

Delays were varied in 12 steps according to the sequence $d_n(\text{ms}) = n + 1 + d_{n-1}$ which produces the set:

$$d_n = \{0, 2, 5, 9, 14, 20, 27, 35, 44, 54, 65, 77\}(\text{ms})$$

The sequence was chosen in order to weight the distribution towards the low-delay region (it bears no special significance, otherwise). Delays were chosen from the set in random order and each duo performed each condition once. Starting tempo per trial was also randomly selected from one of three pre-recorded "metronome" tracks of clapped beats at 86, 90 and 94bpm. Other trials were inserted randomly in the sequence and are not analyzed as part of the present experiment (2 for diverse tempi, 2 for asymmetric delays, with also 2 subject-against-recorded-track runs at the beginning and end). Overall, one session took about 25 minutes to complete.

2.2.5 Recorded segments of interest

We are interested only in the sections of the recordings where both clappers are performing together. Since the protocol allowed the initiating clapper to clap solo for a variable length of time before the second one joined, we first identified the region in which both clappers are involved, Fig. 1. Clapper A (green circles) starts the experiment and is followed by clapper

[3] `http://ccrma.stanford.edu/software/stk/`

B (red circles). Circled notes correspond to the common beats which are automatically identified in a first pass on the raw data (Fig. 3).

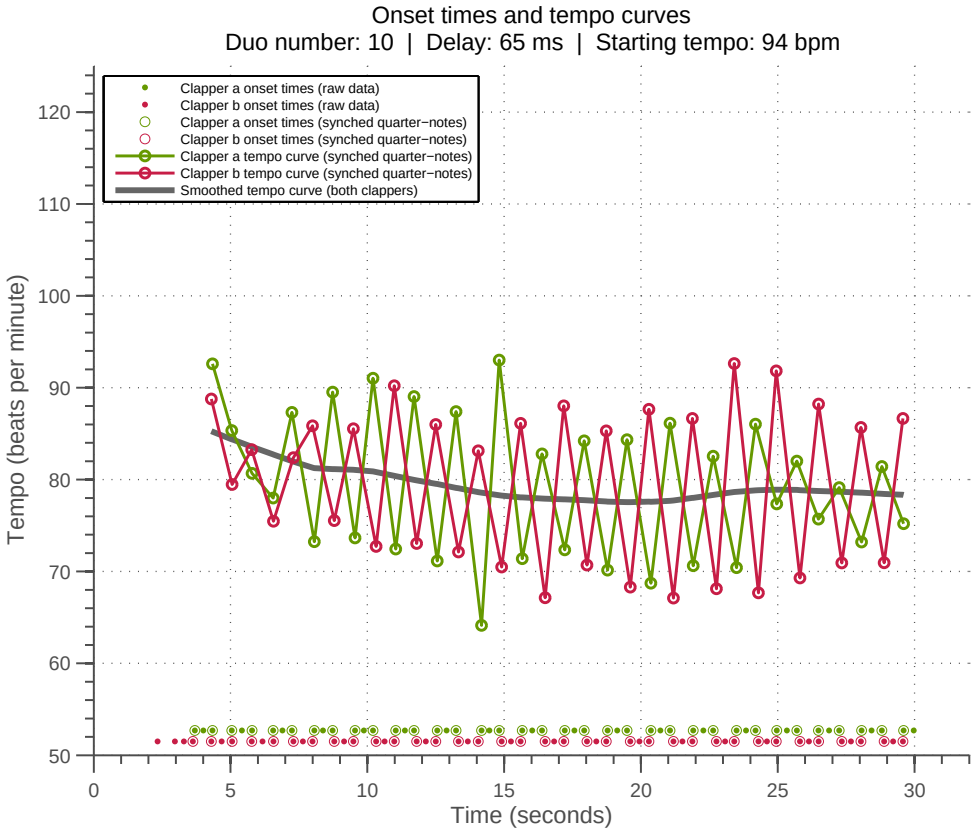

Figure 3: **Onset times, synchronization points and tempo curves for one trial.** A smoothed tempo curve is derived from the instantaneous tempi of both player's synchronized events.

2.2.6 Event detection

An automated procedure detected and time stamped true claps. Detection proceeded per subject (one audio channel at a time).

Candidate events were detected using the "amplitude surfboard" technique (**?**), tuned to measure onsets to an accuracy of ±0.25ms. 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. 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 refractory period, during which another cannot occur.

2.2.7 Event labeling, tempo determination

Inter-onset intervals (IOI's) were calculated from the event times. 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. Since only eighth and quarter-notes are present, we clustered the IOI's into two separate groups using the k-means clustering algorithm (**?**). The group of notes clustered with the shortest IOI is identified as eighth-notes and the one with the longest as quarter-notes. Only one quarter-note was miss-classified (Duo number 8, delay 9, start tempo 86) and it was corrected by hand.

After notes where identified, conversion to tempo (in bpm) is computed with:

$$\text{tempo}_{1/4} = \frac{60}{\text{IOI}}(\text{bpm})(\text{forthequarter} - \text{notes})$$

$$\text{tempo}_{1/8} = \frac{60}{2 \cdot \text{IOI}}(\text{bpm})(\text{fortheeighth} - \text{notes})$$

2.2.8 Acceleration estimation

We are interested in a metric of tempo acceleration (or deceleration) across each trial. This step establishes a smoothed tempo curve, merging both clappers into one curve. Smoothing is computed with a "local regression using weighted linear least squares and a 2nd degree polynomial model" (MATLAB's `smooth` function included in the *Curve Fitting Toolbox*). Then, to obtain a single quantity representing a trial's overall acceleration, we compute the average of the derivative of the tempo curve.

2.2.9 Trial qualification

Of the 17 sessions, 1 was discarded because of an inability to perform the clapping rhythm. Lack of competence was judged subjectively and confirmed by high tempo jitter. ANOVA and multiple comparisons of the mean tempo jitter of each session xxxxxxxxx\$bar{{s_{i}^{2}}}(i = 1, 2, ...17)\$: revealed a significant difference between the 1 discarded session and all other ($p = 1.0 \times 10^{-8}$). Hand fixesxxxxxxxx

2.3 Results

2.3.1 Database

Figure 3 presents the results for one trial. The example shows raw onset times, common beat synchronization points, instantaneous tempo of each event in both clappers, and the smoothed common tempo curve. The full set of trials is available online[4]. The database[5] of events detected in the recordings is also being maintained on a public server for continuing analysis.

[4] `http://ccrma.stanford.edu/groups/soundwire/research/clapping-experiment/`
[5] `http://ccrma.stanford.edu/groups/soundwire/research/clapping-experiment/`

2.3.2 Effect of Tempo

ANOVA and multiple comparisons of the mean tempo at each of the three starting tempi (86, 90, 94bpm) revealed no significant difference between these cases, ruling out a dependence on absolute tempo. All trials were shifted to a starting tempo of 90 bpm before further analysis.

The assigned rhythm creates points at which claps should be simultaneous (highlighted by circles in the diagram, Fig. 4). Disparities at these synchronization points are calculated to show the amount of anticipation (lead) or lateness (lag) of each player's circled event with respect to the other's. In Figure 5, the mean and variance of onset asynchronization are shown for each delay condition. The two qualities which we presume to be of greatest importance for a satisfactory performance are high synchronicity and low variance. To identify the "most satisfactory" region, we weighted onset asynchrony by variance (Fig. 6). Four regimes are distinct with respect to these qualities, and appear in order of increasingly greater delay: *lead* (asynchrony from anticipation), *stability* (the "sweet spot"), *lag* (asynchrony from lateness), and *deterioration* (dominated by variance). Evidently, stability requires a small amount of delay, but not too much.

Leading or lagging at the synchronization points cumulatively effects the general tempo (Fig.7). Tempo maps (in beats per minute) are derived from the instantaneous tempo at each clap event. A general trend from acceleration to deceleration with increasing delay is apparent. We model this effect by measuring each smoothed tempo curve's mean acceleration and aggregating these means for each delay. Figure 8 reveals an orderly relationship as deceleration increases with delay. A linear model of the relationship across the range of sampled delays

$$\hat{y} = 0.03899 - 0.006775x + \epsilon \tag{1}$$

fits the data well, $R^2 = 0.94$. Of particular interest are the y-intercepts of the model and sample data means (Fig. 8) which again indicate that the region with no tempo acceleration ($\hat{y} = 0$) occurs with non-zero delay. The transmission delays in air which separate performers naturally create time delays of this same order. If there exists some intrinsic tendency to anticipate by this small amount, then short delays were required to balance this out.

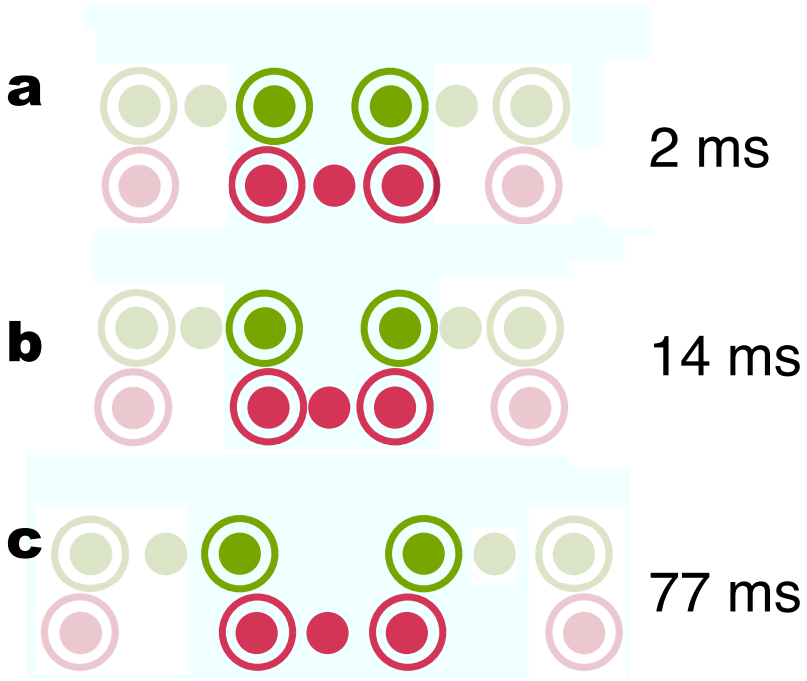

Figure 4: **Performed rhythms through time of one pair of clappers recorded at different delays.** Clapper A green, B is red. Ideally, each (circled, vertically adjacent) pair of events is synchronous. Leading or lagging by one subject with respect to the other at these points is related to delay. **a**, leading at 2ms. **b**, approximately synchronous at 14ms. **a**, lagging at 77ms.

The tendency to slow with increasing delay is moderated at the region of stability. In this interval, the deceleration measured from the sample data remains constant -0.1bpm/s (Fig. 8).

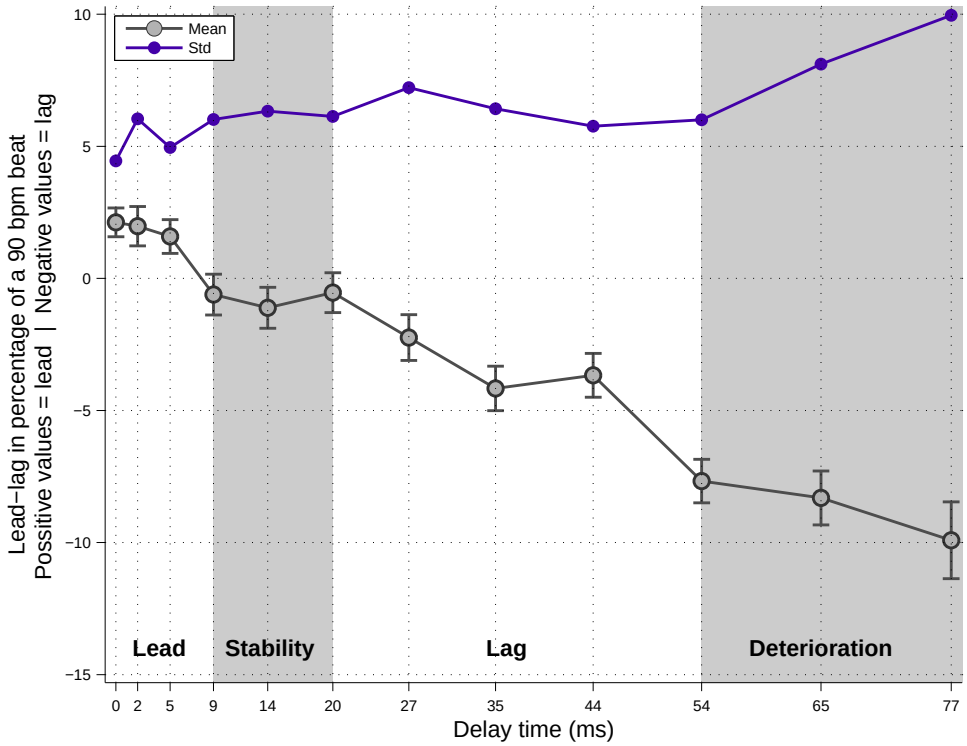

Figure 5: **Onset asynchrony measured at all beat points is compared for the set of delay conditions.** At very small delays, performances are dominated by a tendency to *lead*. Increasing delay improves synchronization (*stability*) until *lag* becomes prevalent. At the greatest delays, variance dominates (*deterioration*). Error bars are 95% confidence intervals for the mean.

3 Discussion

We know from experience that not all music will tolerate even a small drag on tempo. For example, an Internet experiment involving a string quintet playing the first movement of the Mozart G minor String Quintet (K516) (with the St. Lawrence String Quartet plus a former member) found that a separation of 25ms (−0.15bpm/s from Fig. 8) introduced perceptible variance in fast rhythmic passages[6]. Moreover, when two of the players experimented with letting the lag accumulate, it promoted an effortless *ritardando* (intentional deceleration).

[6]Recordings of these experiments are available online at `http://ccrma.stanford.edu/groups/soundwire/research/slsq/`

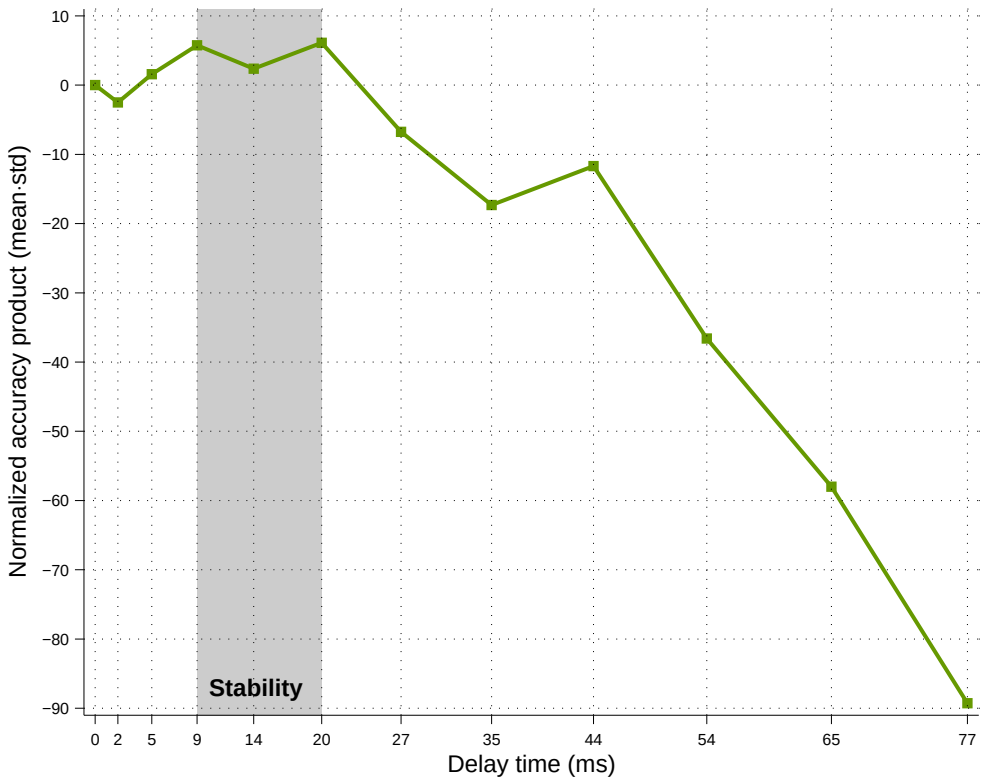

Figure 6: **Normalized accuracy product for lead-lag mean and std** (mean · std)**.** Ensemble competence is a function of greater synchronization and lower variance (at least, for the simple musical task studied here). The region of best rhythmic accuracy is defined as a product of minimum mean lead-lag weighted by variance.

In this sound clip, one hears that as the tempo slows, it eventually settles on a point where stability is achieved (at a much slower tempo). Our present clapping experiment tested only one tempo, moderately fast, at 90 beats per minute (with slight offsets introduced in the experimental protocol to avoid over-training to an absolute tempo, see Method). Experimentation with other, significantly different tempi will be required to include tempo in the present model.

Musicians use strategies to adapt to delay (**?**). Strategies include intentionally pushing (leading) the beat or ignoring the sound of part of an ensemble (to eliminate the recursion mentioned above). This latter seems to be a natural tendency when there is an imbalance

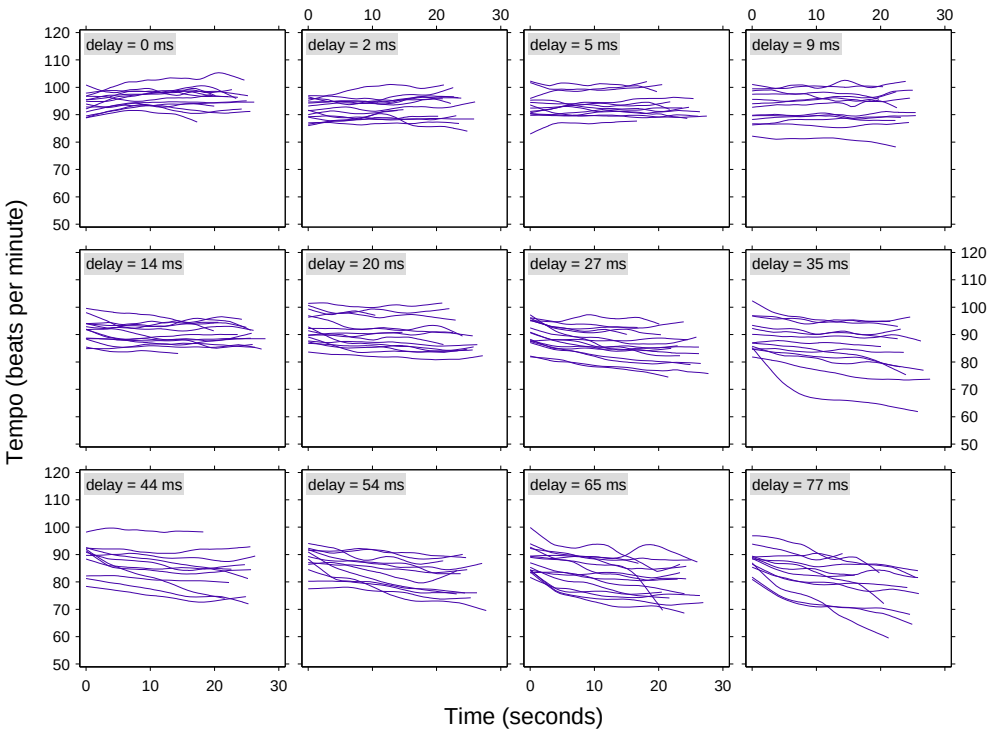

Figure 7: **All trials tempo curves grouped by delay.** Tempo acceleration during a given performance is tracked by measuring inter-onset intervals.

in the structure of the ensemble. The weaker side (in terms of rhythmic function) naturally follows the strong one (whose rhythmic role, instrument type, and/or number of players dominates).

In terms of delay conditions actually tested in the present study, rhythmic stability has a maximum possible delay located at 20ms, beyond which tempo decelerates markedly. A minimum amount of delay, found at 9ms, appears to be necessary to cancel out a tendency to anticipate. Bounded by these conditions is a region in which neither lag, nor asynchronization, nor deceleration increase. For network music performance, the upper limit for this stable regime corresponds roughly to a path length of 2000km (if using present North American research internets provided by Internet2 and Canarie, as examples (**?**)).

The conclusion stands in sharp contrast to other studies where an upper acceptable limit

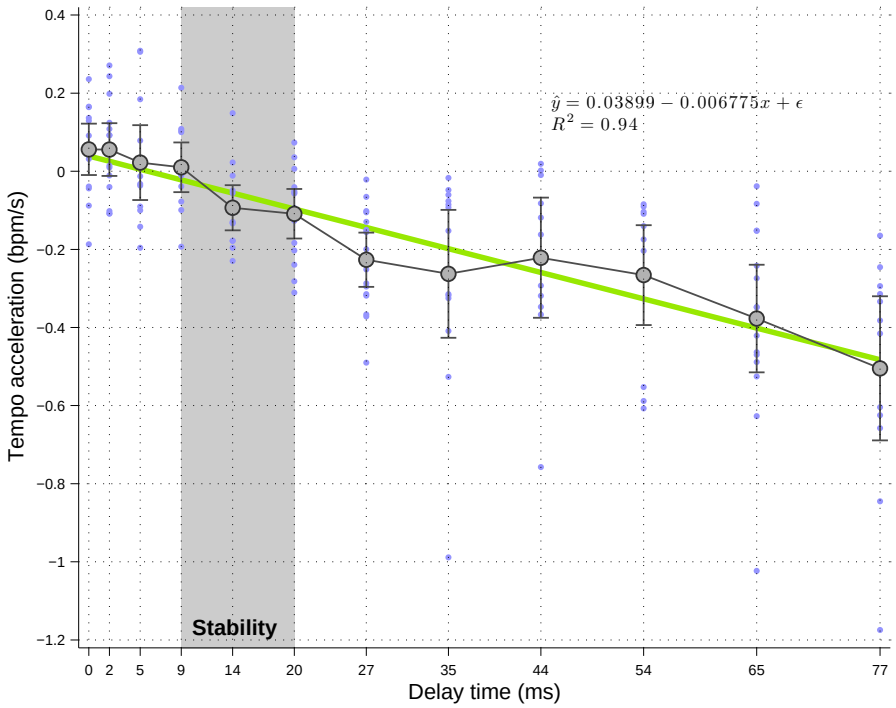

Figure 8: **A single measure of tempo acceleration (its mean) is compared for all performances.** A linear model (green line) correlates well with data sampled at the given delay conditions. Error bars show 95% confidence intervals for the acceleration mean. Single blue dots represents acceleration mean for each individual trial.

is 80msor greater. XXXXX

Acknowledgements

Many thanks to our study team at CCRMA, including students Nathan Schuett, Grace Leslie, Sean Tyan and the CCRMA technical support staff. Grant support from Stanford's Media-X program funded the (2004) data collection and Alberta's iCore Visiting Professor Program, the (2009) analysis. Stephen McAdams' comments on early drafts are gratefully acknowledged.