The Quest for Low Latency*

Nelson Posse Lago† and Fabio Kon
Department of Computer Science — University of São Paulo
{lago, kon}@ime.usp.br

Abstract

Low latency processing is usually a goal in real-time audio applications; however, it is not clear how little latency is to be considered low enough. We discuss currently available experimental data on human perception and argue that somewhat high latencies (around 20–30ms) are probably perfectly acceptable for typical musical applications. We also argue that it should be possible to accept various levels of latency on a system if we can be aware of the effects of this latency on the users of the system; therefore, we still need more experimental data on latency perception to be able to better assess the effects of latency on musical applications. Such an experiment is suggested.

1 Introduction

In interactive systems, adequate latency and jitter characteristics are determined by the interaction with the user: high latency or jitter may impair the user's performance or, at least, offer a frustrating and tiring experience. So, in order to assess the quality of an interactive system regarding its latency and jitter characteristics, we need to understand their effects on the user's perception so that we can define maximum acceptable values for these parameters on such system. The acceptable limits for latency and jitter on an interactive system may vary a lot. Interactive multimedia applications usually require the lowest latency and jitter values, since they usually involve at least one continuous media that may be modified by the user's interaction. But even in multimedia systems there are differences on the acceptable limits for latency and jitter: human hearing has a higher time precision than vision (Repp 2003), and the time precision involving different stimuli types (such as visual and auditory or auditory and tactile) is usually lower than temporal precision with stimuli of the same kind (Levitin et al. 1999).

The higher timing precision of hearing and its relevance to music make the control of latency and jitter a very important part of the design of several systems for computer music. In many cases, systems are developed aiming at producing the lowest latency and jitter possible, which current cost-effective technologies put around a few miliseconds. However, many applications, especially those dealing with wide area network delays, cannot typically offer latencies under 10*ms*, and may be limited to much higher latencies; still, they are obviously very interesting and are, therefore, developed in spite of the supposedly suboptimal latency and jitter characteristics they are able to offer.

While latency and jitter have been discussed a lot and much is already known about how we perceive them, we still lack experimental research that enables us to understand better the various tradeoffs between latency, jitter, human performance, and perception. For instance, it would be hard to argue that "pop" music requires more strict synchronization between performers than slow-moving textural music; but what are the acceptable limits for latency and jitter in each scenario? And, more importantly, what about other scenarios? When are latency and jitter perceivable? When are they influential on the performance of a musical instrument? When do they degrade the user experience? When do they seriously impair different kinds of human performance?

In this paper, we try to show that there are a lot of aspects in human perception regarding latency and jitter that go beyond the usual "less is better" approach; that the naïve "perceptible is bad, not perceptible is good" approach to the problem may be inadequate in some circumstances; and suggest an experiment that could be carried out to help shed some light on the subject.

2 Effects of latency and jitter

Since we are able to use timing deviations as low as $20\mu s$ between ears as cues to determine spatial positioning (Pierce 1999, p. 102), variations in the typical 44.1KHz sampling frequency may affect our spatial perception. However, since this kind of jitter comes from hardware imprecisions, there is not much that can be done about it but to improve the hardware precision and maybe increase the audio sampling rate. Besides that, this kind of jitter is not directly related in any way to the interactive aspect of a system, and therefore will not be further discussed here.

Timing may also affect the perception of timbre, such as

*This work was partially supported by a grant from CNPq (Brazil), process # 55.2028/02-9.
†Partially supported by CAPES (Brazil).

in comb filtering or in tight drumming flams (Wessel and Wright 2002). Comb filtering effects only occur in situations in which an original sound is mixed with a corresponding delayed sound and both sounds are reasonably similar (for instance, the sound of an ordinary acoustical instrument and the same sound processed in order to increase its high frequencies). Even in such cases, ordinary latencies in real-world environments may also produce comb filtering effects, which suggests that this effect may be largely ignored in most situations[1]. The timbre of flams may be altered by timing differences as low as 1*ms*; so, in order to capture these, we need to guarantee a sufficiently high sampling rate and jitter below this level. As we will see later, jitter values close to this are relevant in other scenarios as well, which suggests that trying to achieve low levels of jitter (perhaps by trading it for added latency) is usually a good strategy.

Outside of these extreme examples, the problem with latency and jitter is usually a problem of perceived synchronization: they may prevent us from perceiving events that should appear to be simultaneous as such. This, in turn, may affect our interaction with the system. We may divide pairs of events that may have to be perceived as simultaneous in a musical system in three categories: an external and an internal isochronous beat (that is, the relation of a beat-based musical structure and the corresponding induced beat on the user), pairs of external events (such as pairs of notes or flash lights and note onsets), and actions of the user and their effects (for instance, what happens while playing a musical instrument)[2].

2.1 Synchronization in rhythm

One important characteristic of human perception is rhythm, and rhythm is obviously very important in several applications involving music. This is an area where human performance and perception show extremely high precision, although not always in a conscious manner. It was shown that we can tap a steady beat with typical variations in inter-tap intervals as low as 4*ms* (Rubine and McAvinney 1990). Similarly, we can also adjust our tapping to compensate for variations of around $4ms$ in interstimuli intervals in an otherwise isochronous pulse sequence (Repp 2000) and detect consciously timing variations of around 6*ms* (Friberg and Sundberg 1995). If such variations are cyclic and a little higher, close to 10*ms*, we even spontaneously perform together with them (and not only correct our tapping after each variation is detected) (Thaut, Tian, and Azimi-Sadjadi 1998). This kind of adjustment, however, is done subconsciously. Still, it is not unlikely that such variations are perceived not as timing variations, but as some kind of fuzzy musical characteristic like the so-called "feel". In fact, there are strong indications that performers do introduce such variations in performances according to musical context (Bilmes 1993).

Experimentation suggests that this rhythmic perception is based on the comparison between the expected and actual time for each sound attack (Schulze 1978); this hypothesis is reinforced by the fact that such precision in tracking rhythmic variations is not significantly affected if we tap out of phase (that is, on the "upbeat") (Repp 2001). This in turn means that the perception of rhythmic variations of around 10–20*ms* is not based on auditory cues related to the slight differences in attack moments of close sounds. Instead, such high precision regarding rhythm means we are able to assess time intervals and attack times with around 4*ms* of precision in a subconscious level, and that discrepancies of this magnitude may affect the feel of some kinds of music (those that are based on a very steady isochronous pulse, like many forms of "pop" music). This makes a strong point for the case of trying to minimize jitter as much as possible in a computer music system if such kinds of music are to be supported.

2.2 Synchronization in external events

It would be tempting to conclude that such precision in perception means we need to guarantee that events that should be perceived as simultaneous should indeed happen with no more than around 4*ms* of asynchrony between them. However, asynchronies of up to around 50*ms* in supposedly simultaneous notes are not at all uncommon during ordinary music performance. In fact, the percussion and horn sections of an orchestra may be over 10*m* farther from the audience than the violin section, which results in asynchronies around 30*ms* for the public beyond the ordinary asynchronies between instruments. Even in chamber music, asynchronies of up to 50*ms* are common (Rasch 1979). In a similar way, dynamic differences between voices on pieces for the piano are responsible for what has been called *melody lead*: notes of the melody are tipically played around 30*ms* before other supposedly simultaneous notes[3]. In spite of the percussive characteristic of the piano sound (which results in short attack times and, therefore, very distinguishable attacks), these asynchronies are not perceived as such by performers or the audience. Finally, subjects asked to tap along with a metronomic stimulus virtually always tap about 10–80*ms* ahead of time (typically 30*ms*)

[1] In fact, there is not much that can be done about such effect of latency, since almost *any* reasonable latency value, high or low, will result in comb filtering in such situations.

[2] A special case of synchronization between an internal and an external beat exists when the external beat adjusts to the user's internal beat. A simple example is the rhythm synchronization between music performers, where the internal beat of all performers must be synchonized and, for each performer, the other performer's beat is external. There is already some work on this area (Schuett 2002), but the results were somewhat inconsistent; we will not address this subject here, but additional work must be done on this topic.

[3] There has been some debate as to wether such effect is only a reflection of the dynamic differences between voices or if it is subconsciously introduced by the performer in order to highlight the melody line. Recent research (Goebl 2001), however, leaves very little doubt that this effect is indeed a consequence of the dynamics; the perceptual effect of the melody lead effect also appears to be minor (Goebl and Parncutt 2003).

without noticing it (Aschersleben 2002). These facts suggest that latencies responsible for asynchronies in external events of up to at least 30*ms* may be considered normal and acceptable under most circumstances; music performance with traditional instruments is not impaired by them. In fact, such asynchronies are used by the ear as strong cues for the identification of simultaneous tones (Rasch 1978).

It may be argued that, even if such asynchronies are not consciously perceptible, they may have a musical role and be partly under the control of the performer; in fact, as just mentioned, they are at least responsible for better tone discrimination. Apparently, though, if such musical role exists, it is minor: not only perceptual experiments showed little impact of variations in artificially-induced asynchronies (Goebl and Parncutt 2003), but also performers apparently do not have such high precision in controlling note asynchronies. This stems from the influence of tactile and kinesthetic (usually called *haptic*) sensations that accompany the action.

2.3 Synchronization in haptics

This brings us to the most interesting aspect of latency and jitter for multimedia and music applications: the perception of the latency between an user action and the corresponding reaction. In this respect, our perception once again shows a very high degree of precision: it was shown that variations in feedback delay of 20*ms* are, although not consciously noticed, compensated for in the same manner as we can adjust tapping to a slightly disturbed beat sequence (Wing 1977). It is reasonable to expect similar mechanisms to be involved in both cases; in fact, it is most likely the same mechanism that is involved: subjects create an expectation for the moment in time for the feedback, detect the feedback disturbance and try to compensate for it.

In spite of the similarity, in such situation there are three elements at stake that make matters more complex: the user's motor commands, the user's corresponding haptic sensations, and their relation to the external feedback. These elements are important because there is very strong evidence suggesting that the moment we recognize as the moment of start of external feedback can be widely influenced by several factors, including the haptic sensations (which are themselves a form of feedback) (Aschersleben 2002). This means that events that actually happen simultaneously may be perceived as asynchronous, even if only at a subconscious level.

As mentioned before, subjects typically tap together with a metronomic stimulus ahead of time. The amount of anticipation, however, is dependent on the characteristics of both auditory and haptic feedback. Auditory-only feedback produces perfect synchronization; haptic-only feedback produces reasonably large anticipations; both forms of feedback together produce relatively small anticipations; and finally, normal haptic feedback combined with delayed auditory feedback produce anticipations that grow in accordance with the amount of delay (Stenneken et al. 2003; Aschersleben and Prinz 1997; Mates and Aschersleben 2000). Excluding the very special cases of auditory-only feedback, such measured variations were of about 15*ms* for auditory feedback delays between zero and a little less than 30*ms* for subjects that proved to show very little variability due to previous training. Anticipations also tend to decrease in contexts where there is sound data in between beats. Such variations give further indication that, while asynchronies in note onsets are used as cues to tone discrimination, their role in musical expression is probably very limited.

The most important aspect of this is the fact that we can subconsciously adjust our performance to compensate for such different feedback conditions. During experiments with delayed feedback, subjects clearly altered their behavior according to the characteristics of each trial, forcing the researchers to introduce control trials between each pair of trials (Aschersleben and Prinz 1997; Mates and Aschersleben 2000). In piano performance, the time elapsed between pressing a key and the corresponding note onset is around 100*ms* for *piano* notes and around 30*ms* for *staccato*, *forte* notes (Askenfelt and Jansson 1990). Even if we assume that the pianist expects the note onset to happen somewhere in the middle of the course of the key, it is very likely that latencies will be different for different dynamic levels. Still, pianists have no problem dealing with such different latencies; since voices in pieces for the piano usually have dynamics that change continuously, the performer has the opportunity to adjust himself to the corresponding changes in latency. When there are abrupt changes in dynamics, they usually are related to some structural aspect of the music, which brings with it large interpretative timing variations. Finally, modern music may make use of dynamic changes that do not fit well with interpretative timing variations; however, such music is usually not based on a clear and steady beat, making the effects of sudden variations in latency much less perceptible.

In fact, since our motor system cannot react instantaneously, we must issue motor commands ahead of time in order to perform "on time"; it is not hard to believe that the various feedbacks for our actions are used to calibrate how much ahead of time commands are issued. In tapping experiments, latencies of up to around 30*ms* were adjusted for, resulting in final asynchronies between stimulus and response variations of about 10*ms* (Mates and Aschersleben 2000), which, as previously stated, we believe are mostly irrelevant.

3 Conclusion and the future

We hope we have been able to argue convincingly that somewhat large latencies, maybe up to 20–30*ms*, are pretty much acceptable for most multimedia and music applications.

Jitter, on the other hand, can be a bigger problem; but it is generally possible to trade jitter for added latency. This does not mean lower latencies are not of interest; quite on the contrary, since latency in different parts of a system accumulates. For instance, the mere positioning of loudspeakers at around 3–4m of distance from a user adds 10ms to the total perceived latency of the system; many DSP algorithms add significant latency; etc. Therefore, aiming at the lowest possible latency in each part of a system helps keep the overall latency under control. Still, tradeoffs are acceptable.

Currently available data is still insufficient to determine clearer limits for latency and jitter as well as to confirm much of what was said here in a musical context, making it difficult to assess the quality of musical and multimedia applications regarding temporal precision. This is so because much of the current research on music and timing perception makes use of non-musical stimuli. Since timing is so tightly tied to still unmeasurable aspects of music such as "feel" and since at least part of our timing perception occurs outside of consciousness, we need more experiments performed on actual music. Such experiments would face many technical challenges, some new, some of which have already been dealt with before (Bilmes 1993; Levitin et al. 1999).

As an example, one such experiment would be, on a small ensemble, to subject one of the instrumentists to different feedback latencies to assess their effect over the perfomer. This should be repeated with feedbacks provided by earphones, loudspeakers, with and without artificial reverberation. Also, different kinds of music (such as tonal classical music and percussion-rich "pop" music) might have an impact. Finally, running such tests in different rooms would be useful, since acoustical ambience may affect the performer.

References

Aschersleben, G. (2002). Temporal control of movements in sensorimotor synchronization. *Brain and Cognition 48*, 66–79.

Aschersleben, G. and W. Prinz (1997). Delayed auditory feedback in synchronization. *Journal of Motor Behavior 29*(1), 35–46.

Askenfelt, A. and E. V. Jansson (1990). From touch to string vibrations. I: Timing in the grand piano action. *Journal of the Acoustical Society of America 88*(1), 52–63.

Bilmes, J. A. (1993). Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, MIT, Cambridge. Available at: <http://www.icsi.berkeley.edu/~bilmes/mitthesis/mit-thesis.pdf>.

Friberg, A. and J. Sundberg (1995). Time discrimination in a monotonic, isochronous sequence. *Journal of the Acoustical Society of America 98*(5), 2524–2531.

Goebl, W. (2001). Melody lead in piano peformance: Expressive device or artifact? *Journal of the Acoustical Society of America 110*(1), 563–572.

Goebl, W. and R. Parncutt (2003). Asynchrony versus intensity as cues for melody perception in chords and real music. In R. Kopiez, A. C. Lehmann, I. Wolther, and C. Wolf (Eds.), *Proc. of the 5th Triennial ESCOM Conference*, Hanover, Germany, pp. 376–380. Available at: <http://www.oefai.at/cgi-bin/get-tr?paper=oefai-tr-2003-11.pdf>.

Levitin, D. J., K. MacLean, M. Mathews, and L. Chu (1999). The perception of cross-modal simultaneity. Available at: <http://ccrma-www.stanford.edu/~lonny/papers/casys1999.pdf>.

Mates, J. and G. Aschersleben (2000). Sensorimotor synchronization: the impact of temporally displaced auditory feedback. *Acta Psychologica 104*, 29–44.

Pierce, J. (1999). Hearing in time and space. In P. Cook (Ed.), *Music, Cognition, and Computerized sound: an Introduction to Psychoacoustics*, pp. 89–103. Cambridge: MIT Press.

Rasch, R. A. (1978). The perception of simultaneous notes such as in polyphonic music. *Acustica 40*, 21–33.

Rasch, R. A. (1979). Synchronization in performed ensemble music. *Acustica 43*, 121–131.

Repp, B. H. (2000). Compensation for subliminal timing perturbations in perceptual-motor synchronization. *Psychological Research 63*, 106–128.

Repp, B. H. (2001). Phase correction, phase resetting, and phase shifts after subliminal timing perturbations in sensorimotor synchronization. *Journal of Experimental Psychology: Human Perception and Performance 27*(3), 600–621.

Repp, B. H. (2003). Rate limits in sensorimotor synchronization with auditory and visual sequences: The synchronization threshold and the benefi ts and costs of interval subdivision. *Journal of Motor Behavior 35*(4), 355–370.

Rubine, D. and P. McAvinney (1990). Programmable fi nger-tracking instrument controllers. *Computer Music Journal 14*(1), 26–40.

Schuett, N. (2002). The effects of latency on ensemble performance. Available at: <http://www-ccrma.stanford.edu/groups/soundwire/performdelay.pdf>.

Schulze, H.-H. (1978). The detectability of local and global displacements in regular rhythmic patterns. *Psychological Research 40*, 173–181.

Stenneken, P., J. Cole, J. Paillard, W. Prinz, and G. Aschersleben (2003). Anticipatory timing of movements and the role of sensory feedback: Evidence from deafferented patients. Available at: <http://jacquespaillard.apinc.org/deafferented/pdf/stenneken-et-al-ms-03.pdf>.

Thaut, M. H., B. Tian, and M. R. Azimi-Sadjadi (1998). Rhythmic fi nger tapping to cosine-wave modulated metronome sequences: Evidence of subliminal entrainment. *Human Movement Science 17*, 839–863.

Wessel, D. and M. Wright (2002). Problems and prospects for intimate musical control of computers. *Computer Music Journal 26*(3), 11–22.

Wing, A. M. (1977). Perturbations of auditory feedback delay and the timing of movement. *Journal of Experimental Psychology: Human Perc. and Performance 3*(2), 175–186.