“Towards a telematic visual-conducting system”

Alexander Carôt[1], Gerald Schuller[2]

[1]*University of Applied Sciences, Anhalt, Germany*

[2]*Fraunhofer IDMT, Ilmenau, Germany*

Correspondence should be addressed to Alexander Carôt (alexander.carot@inf.hs-anhalt.de)

ABSTRACT

Recent advances in networking technology (higher bit rates and lower transmission latencies) enable new applications where musicians can play together remotely, over the Internet. This application requires an audio coder providing sufficient compression to avoid overloading a connection and delay jitter, and also having a very low encoding/decoding delay. Often it is also desirable to have a visual connection, particularly for a conductor of an orchestra. But a parallel video connection often overloads a low delay connection, and usually also leads to more jitter and delay in the audio connection. Hence our approach is to design a special conductor transmission scheme, using a standard computer mouse, for this purpose. We found that the data from this transmission scheme can easily be integrated in the audio data stream without affecting jitter and delay. Experiments showed that, depending on the tempo of the music, the conductor and the orchestra could tolerate round trip times of about 75 to 150 ms.

1. INTRODUCTION

Since the year 2000 the domain of telecommunication has been influenced by requirements from artistic world: Generally distant realtime communication had been human voice conversations such as telephone calls and video conferences. This meant audio qualities up to 8 kHz and latencies up to 400 ms as generally accepted reference values [1]. However, with rapidly evolving multimedia capabilities of current operating systems, the possibility of writing custom music applications and increasing Internet bandwidths more and more artists and especially musicians considered using this technical infrastructure as a realtime interaction system – with latencies up to a maximum of 30 ms and linear 48 kHz, 16 bit audio quality [2]. In that context one major goal has been to musically perform with distributed persons anywhere in the world. As one of the major pioneers Chris Chafe et al took the approach of consciously choosing Internet2 [3] connections (with their corresponding high bandwidth and low jitter) and transmitting low-buffered, uncompressed audio packets across them. Since then, numerous distributed music sessions have successfully been performed as part of the Jacktrip project [4] between Internet2-connected locations all over the world. In that context they were able to prove that – despite its asynchronous transmission characteristics – the Internet is able to transport audio data in high quality and with low latency. Nevertheless, it had so far not been possible to evaluate this effect for narrow band networks such as A-DSL with its upload limitation of often no more than 500 kbps. For this reason we developed the Soundjack project with appropriate low delay audio compression algorithms: Especially by using the ULD [5] or CELT codec [6] it is possible to transmit a high quality audio stream with a compressed payload of only 96 kbps. In that context the author could show that also a conventional A-DSL endpoint connection meets the quality and delay requirements for distributed music performances [7].

2. PROBLEM AND PREVIOUS APPROACHES

Figure 1 shows the typical baton movement from the a conductor’s perspective. This figure represents the timing reference and hence – just like the audio signal – has to be transmitted as quickly as possible.

Numerous conferencing systems have a working video implementation but according to the previously mentioned telecommunication conventions the inherent latencies do not fulfill the strong requirements

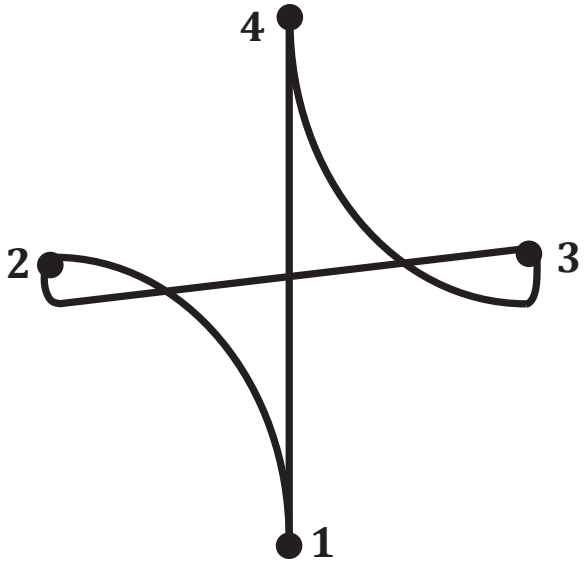

Fig. 1: Typical baton movement in 4/4 from the conductor's perspective. The numbers indicate the beats of the measure.

of network music performances. Examples are the iChat software [8], which implements the well-known H.264 standard [9] or Skype [10]. In that context it is mainly the large compression delay of several hundred milliseconds [1], which prevents participants from achieving time-accurate synchronization. As a consequence one of the first low-delay-video researchers Jeremy Cooperstock presented a video system, which transmits uncompressed video streams on broadband backbone connections of several Gbps [11]. Since the year 2008 the world opera project [12] has successfully been using this principle. Another experiment based on a broadband ATM network [13] was presented in [14]. Nevertheless, even in this broadband context a video technology-related problem remains: Although the conventional frame rate of 25 fps provides a fluent impression to the observer's eyes it might not suffice in terms of quick conductor movements. The image capture interval of 40 ms is not associated with a signal delay, however, it does not resolve faster movements within two capture moments. Figure 2 shows the vertical baton position over time – the abscissa is subdivided into intervals of 40 ms for a 25 fps video. It is obvious that due to the limited frame rate only a rough representation of the original signal can be achieved. Also none of the timing cue moments can be reproduced precisely. This effect is illustrated in an exaggerated manner: The slower the timing of a piece, the lesser the effect actually occurs.

Furthermore, the required bandwidth for an uncompressed video stream is extremely high: An average quality PAL standard video with a resolution of 768*567 pixels of 30 bit each and a frame rate of 25 fps already requires more than 330 Mbps, and – if possible – the bandwidth proportionally increases with higher frame rates.

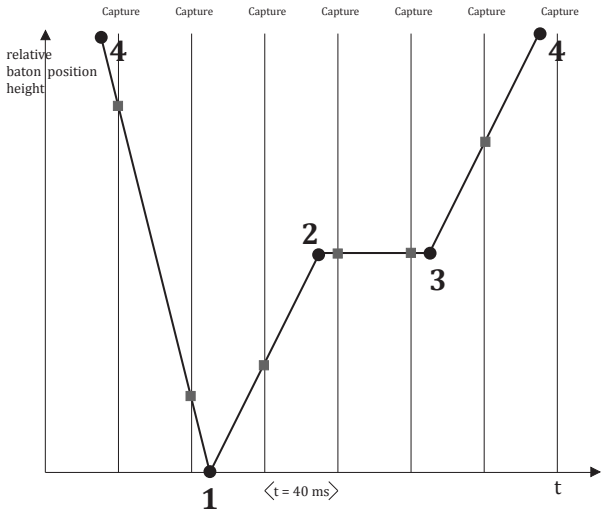

Fig. 2: Vertical baton position over time. The black circles represent the beats of the measure – the grey boxes highlight the actual capture moments.

In contrast, narrowband networks such as A-DSL with significantly less bandwidth are not able to carry uncompressed video streams at all: Often the A-DSL uplink does not offer more than 1 Mbps. Furthermore – even if compression is applied – video data represents a cross-traffic stream [15], which generates network jitter on the endpoint connection and in turn does disrupt the low delay audio stream with dropouts. Hence, in home consumer networks the choice is either a conventional video conferencing system with correspondingly high latencies for both modi or a low-delay audio-only solution [2].

3. GOAL

It is our motivation to find and develop a solution, which

overcomes the described problems and allows the combination of visual and audible low-latency transmission on narrow-band network connections. Since the concept of a remote conductor has not been evaluated for home-consumer networks yet the overall target is the development of a robust, cheap and effective solution. In that context as a first step we emphasize the transmission and representation of the conductor's timing rather than his or her gestures. In this paper we focus on the technical development and not on an extensive cognitive evaluation. The final prototype should offer frame rates beyond 25 fps, require a minimal bandwidth in order to support A-DSL upload limitations and should work with standard PC hardware.

4. CONCEPT AND IMPLEMENTATION

Our idea is that conductors could use a conventional computer mouse as a replacement for a real baton and move it accordingly. The operating system polls the actual mouse coordinates in intervals of the current maximum event timer resolution. Unless a dedicated realtime OS is used, the maximum resolution is about 15 ms for Windows, OSX and Linux. In order to prevent network jitter the captured coordinates are not transmitted as a separate stream. However, according to the author's concept proposed in [16], they subsequently interleave the x and y - coordinates with the actual audio stream packets. As a result one single stream consists of audio and mouse coordinates without negatively influencing one another. Upon reception the data is extracted. Then the audio is played back and the remote host's mouse is updated with the received coordinates. In terms of a better visibility the receiving mouse pointer is replaced with an appropriate image. Figure 3 graphically illustrates the technical implementation: According to the 4/4 conduct figure the conductor controls the mouse, whose coordinates are captured and instantly added to the current audio stream packet. On the receiving musician's end the original mouse movement is reconstructed in order to use it as time reference.

Our solution does not generate additional overhead bandwidth because the amount of transmitted packets remains equal. With a reference configuration of 128 samples/block and a sample rate of 48 kHz a low delay audio stream consists of 371 packets per second [16]. Since our solution adds two 16 bit coordinates to each packet the total amount of additional bandwidth is lower than only 12 kbps.

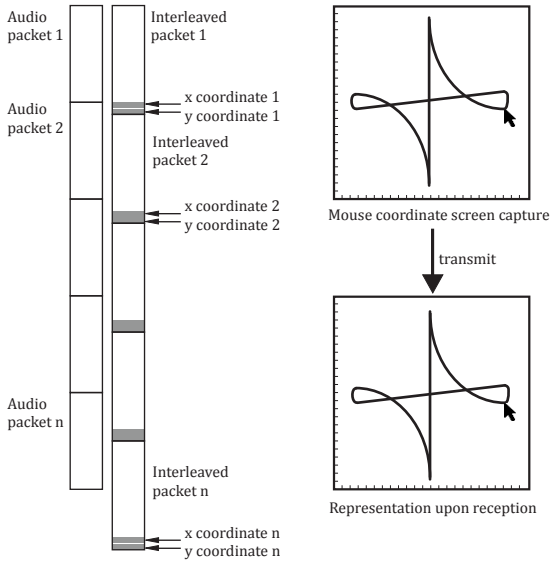

Fig. 3: Illustration of the technical concept

5. EVALUATION

Low-delay audio streams with the Soundjack software have been successfully evaluated over the last couple of years [7][5]. The proposed solution is embedded into Soundjack, which is why we first verified in how far the two interleaved data streams did possibly disturb each other. After several tests it was obvious that this was not the case at all. The audio was transmitted and received as usual. It also was clearly possible to remote-control another external peer's mouse after having extracted the coordinates from the Soundjack audio stream. Furthermore, a rough cognitive and musical evaluation involving conductors and musicians has been realized.

It is important to know that a conducted scenario does not compare with a conventional musical interaction as it is the case between e.g. a number of jazz musicians: Rather than listening to each other and adjusting the performance time accordingly a conductor takes care of the time himself and often "hurries" a number of beats before the currently played note. With respect to musical interaction categories this principle refers to the term Master-Slave-Approach (MSA) [17]. In this scenario the conductor represents the master, who directs the time, while the orchestra follows and in turn represents the role of the slave. In that context it is also important to point out that MSA implies the master to perceive the

roundtrip-latency rather than the one-way-latency: It first requires the one-way latency for the transmission of the mouse coordinates. Secondly, the receiving musicians perform according to the visualized mouse data. Thirdly, it again takes the one-way-latency for the transmitted sound to reach the conductors ears. On the other hand – since the musicians simply follow the received mouse movements – they do not perceive any latency at all. This phenomenon has been analyzed with unconducted music styles in [2]. Since a precise evaluation with conducted music would exceed the frame of this paper the authors perform a proof of concept, which is described in the following paragraph.

Three test connections were established between three musicians (two violins and one cello) and a conductor. As a simple test piece the conductor rearranged the piece "Inventio I" by Johann Sebastian Bach for the ensemble. In fact the participants found it "strange" to work with a computer mouse instead of a real baton. However, none of them complained about this principle and they agreed on the fact that it works for conducted music. At average performance tempo (90 bpm to 120 bpm) none of them considered the MSA principle problematic unless the one-way latency was above 50 ms resulting in a round-trip-latency (RTT) of 100 ms. Faster tempo beyond 120 bpm lead to RTT-thresholds of 70 ms, lower tempo below 90 bpm lead to RTT-thresholds of up to 150 ms. It was clearly obvious that the latency threshold falls with increasing performance tempo but also depends on the the conductor's taste and feel. These results directly correspond to how musicians perform using the MSA principle and that the upper threshold basically depends on the ability and willingness to think or play ahead of time.

6. CONCLUSIONS AND FUTURE WORK

We have successfully developed an efficient and standard-PC-based conducting principle and integrated it into the Soundjack software. In terms of simplicity a computer mouse is used as input medium, whose coordinates are constantly polled and interleaved within our low-delay audio stream. As a consequence Soundjack transmits the audio and the mouse coordinates in parallel without any negative impact regarding latency or jitter. Analogue to audio related aspects the software is now prepared for a precise timing evaluation of conducted remote music scenarios. In early-stage cognitive evaluation studies we realized that a conducted remote music scenario corresponds to the Master-Slave-Approach (MSA) and the respective latency thresholds. Nevertheless, conducted scenarios suffer less from latency since conductors typically think and act ahead of time anyway. The final delay threshold also depends on the individual person. These cognitive relationships require more detailed scientific attention, which will be covered as part of future work.

7. REFERENCES

[1] International Telecommunication Union (ITU), *Recommendation H.323: Audiovisual and multimedia systems – Infrastructure of audiovisual services - Systems and terminal equipment for audiovisual services – Packet-based multimedia communications systems*, 1998.

[2] A. Carôt, "Musical Telepresence – A Comprehensive Analysis Towards New Cognitive and Technical Approaches," Ph.D. dissertation, Institute of Telematics – University of Lübeck, Germany, 2009.

[3] "Internet2 - website," Mar. 2008, http://www.internet2.edu.

[4] C. Chafe, S. Wilson, R. Leistikow, D. Chisholm, and G. Scavone, "A simplified approach to high quality music and sound over ip," in *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*, Verona, Italy, Dec. 2000.

[5] U. Krämer, J. Hirschfeld, G. Schuller, S. Wabnik, A. Carôt, and C. Werner, "Network music performance with ultra-low-delay audio coding under unreliable network conditions," in *Proceedings of the 123rd AES-Convention*, New York, USA, Oct. 2007.

[6] J. M. Valin, T. Terriberry, C. Montgomery, and G. Maxwell, "A high-quality speech and audio codec with less than 10 ms delay," *IEEE Transactions on Audio, Speech and Language Processing*, 2009.

[7] A. Carôt, U. Krämer, and G. Schuller, "Network music performance in narrow band networks," in *Proceedings of the 120th AES convention*, Paris, France, May 2006.

[8] J. McKenna and S. Florell, "Cost-Effective Dynamic Telepathology in the Mohs Surgery Labo-

ratory Utilizing iChat AV Videoconferencing Software," *Dermatologic Surgery*, vol. 33, pp. 62 – 68, 2007.

[9] International Telecommunication Union (ITU-T), *Recommendation H.264: Advanced video coding for generic audiovisual services*, 2007.

[10] "Skype - website," Aug. 2008, http://www.skype.com.

[11] J. Cooperstock, "Real-time networked media: http://www.cim.mcgill.ca/sre/projects/rtnm/," Oct. 2002.

[12] "World opera project - website," Mar. 2011, http://theworldopera.org.

[13] M. P. Clark, *ATM Networks – Principles and Use*. Wiley-Teubner, 1996.

[14] D. Konstantas, Y. Orlarey, O. Carbonel, and S. Gibbs, "The distributed musical rehearsal environment," *IEEE Multimedia*, vol. 6, pp. 54 – 64, May 1999.

[15] A. S. Tanenbaum, *Computer Networks*, 4th ed. Pearson Studium, 2003.

[16] A. Carôt and G. Schuller, "Networked music performance: State of the art," in *Proceedings of the AES 44th International Conference on Audio Networking*, San Diego, CA, USA, Dec. 2011.

[17] A. Carôt and C. Werner, "Network music performance – problems, approaches and perspectives," in *Proceedings of the Music in the Global Village – Conference*, Budapest, Hungary, Sep. 2007.