Vi r Music Handbook

Distance teaching technology and possibilities

Created in 2010–2011 by Noa Nakai, Särestö Academy, Finland
Corrected on 25th April, 2012 ·· Proofread on 15th May, 2011

In co-operation with
Oulu University of Applied Sciences, Finland and
Luleå Tekniska Universtitet, Musikhögskolan i Piteå, Sweden

Contact

The Vi r Music website: http://virmusic.net/

To contact the authors of this Handbook, please send email to:
virmusic.blog@gmail.com

The Vi r Music project ends in January 2011, but the email address stays valid. If you find any mistakes or essential information missing, please report to the above mentioned email address. Since there are many specific details within a wide range of devices mentioned, there might be some errors or outdated information. It is recommended to look for other options and seek updated information before making decisions on equipment purchase.

Abstract

This Handbook aims to answer questions such as: What is possible in distance music teaching today? What is the best quality you can get and what are the limitations? What does it take to set up a distance teaching system? How should the student or teacher prepare for the lesson and does the lesson differ from a traditional local lesson?

The aim is also to cover the most popular, practical and highest quality solutions not only in standalone video conferencing but also the peripherals, software solutions and other related subjects such as networks and room design. Also video recordings in distance music teaching context and some methods to publish the videos are described. Many important matters that a teacher, a student, an organizer and an engineer should know are explained. A lot of focus and effort is put in finding the solutions which offer the highest quality of audio and video.

Contents

Introduction

This guidebook has been written for music teachers, students and music teaching event organizers alike. The idea is to offer practical information related to distance music teaching in a compact manner. Teachers and students find the distance music lesson productive and good if it feels similar to a traditional local lesson. Problems and limitations with the technology may make it difficult to hear or see the other one. Therefore the great challenge is a technical one: the system should be as transparent as possible, while some very useful aids are provided by the technology as a bonus to the conventional lesson. Since the challenges are mostly technical, this Handbook also deals largely with the technical matters. A large portion of the text is hopefully understandable also for readers without technical background, but some of the content is aimed especially for technical persons.

The fields of audiovisual technology and communications are developing relatively quickly so solutions may become obsolete quite fast. Some of the text concentrates on today's solutions and therefore if you are reading this in 2012 or later, chances are that there are new and updated solutions available. Some fundamentals on the other hand don't change so fast, but nonetheless it will be a good idea to consider all solutions as suggestions among others. Some lists of specific technology could be transferred to a wiki page where users can update the information as well.

The Vi r Music project official term was 01/02/2009 – 31/01/2011 and this guidebook has been written within the project. Dozens of successful master classes with several instruments and singing were organized during the Vi r Music project. Not all locations had the opportunity to acquire the best technology and framework, but overall the technical quality was good enough to provide meaningful pedagogy and learning, while the best classes were highly praised by students and others.

The persons working in Vi r Music have no connection to gear manufacturers other than during customer support and occasionally suggesting new features to the manufacturer. Special thanks go to Josh Chaffey from ANU School of Music for the collaboration in finding out feasible technologies.

This Handbook will hopefully benefit high quality global connectivity between organizations of music education. It would be a great achievement to have a technically top level, globally compatible system at all locations so that the teaching and learning experience is not any more hindered by technological problems or distances.

Online Extension of this Handbook

Content relevant to this Handbook is found online at the links below.

Standalone, PC and Mac solutions for video conferencing and streaming:
http://tinyurl.com/virmusic0

Relevant other hardware (microphones, audio interfaces, video hardware etc.):
http://tinyurl.com/virmusic5

AV applications and websites for processing and publishing videos:
http://tinyurl.com/virmusic4

Audio, video and video conferencing forums:
http://tinyurl.com/virmusic1

VLC streaming codec table:
http://tinyurl.com/virmusic2

These technologies are very rapidly evolving. Each room and occasion may require very different technical approaches and certain compromises need to be made to adjust to the situation. Thus any list of equipment will not be completely adequate. Still, these lists may give interesting pointers and ideas.

1 Music teaching technology and infrastructure guide

1.1 The best of distance teaching technology in 2010

In 2010, video conference and computer-aided, network-based music teaching is globally already quite common. There are various different technical solutions available and some of the platforms have already been developed for many years. For example the violinist Pinchas Zukerman was already teaching violin through video conference in 1994. In 17 years, the technology has certainly advanced a lot, but on the other hand for the industry, music teaching is still not as significant as business meetings. Largely because of that, it is still not at all guaranteed that the sound quality would be perfect and that the system would fit music teaching seamlessly.

Two examples[1] of distance teaching violin and cello in the Vi r Music project 2010:

http://www.sarestoacademy.org/demo-rudin2/ and
http://www.sarestoacademy.org/demo-svarfvar/

In those sessions, Tandberg[2] Edge 95 and C60 were used. The videos have resolutions 1272x372 pixels and 1024x358 pixels. The first video has stereo sound for local and mono for far end, the second video is mono only. In the best equipment available today, the typical maximum video resolution is FullHD (1920x1080 pixels if using a full-screen mode). For sound, the theoretical limit is almost to the same as in CD recordings. If the sound is not compressed at all, the limiting factor will be the microphone technology, acoustics and challenges related to the audio processing.

In the best case, distance teaching studios are skillfully built for the exact purpose and the equipment works as easily as switching the screen on and making a regular phone call. Once the lesson starts the image and audio will be sharp and clear throughout the lesson. The display is very good and the other party can be viewed in life-size. No intervention with the technology is needed. In case the camera should be zoomed to show details in hand, the technical supervisor can do it or teacher/student can also control both local and far end cameras if they wish. The lesson can be done even with multiple locations simultaneously, everybody seeing and hearing each other clearly.

At the high end optimal situation, video conferencing today is approaching the level where looking at the screen is similar to looking through a window, just with a small delay. At the best, image and audio quality can be approximately the same as in current advanced video technology in general.

In addition to instrument teaching, also for example music theory and music history are taught remotely, auditioning can be done via video conference and guidance for dance is possible alike. To get an overall view of what has been done, you can view some of the video links listed in chapter ‘3.6 Web sites with music learning videos’.

1.2 Challenges in distance teaching technology

In the previous section the optimal situation was briefly described. However, there are many challenges in instrument teaching through video conferencing. The technology can be unreliable if circumstances are not optimal. Usually the teaching situation is also a busy one, there might be audience and any time spent on fixing technical problems is away from the music lesson.

[1] The technological suggestions mentioned later in this Handbook allow even better quality than what you see in these videos (certain compromises had to be made when equipment for Vi r Music was acquired).

[2] As of 2010, Tandberg is now part of Cisco Systems, Inc.

In the Vi r Music project we did a lot of instrument teaching through video conference equipment and the most common challenges in today's technology were collected to the following web page:

http://www.sarestoacademy.org/technology

All video conference applications have some sort of feedback echo cancellation feature if they are intended to be used with speakers and not with headphones. The functionality is needed because the local speaking or playing will be played on remote speakers, and as the microphone is in the same room with the speakers, the microphone will likely pick up the signal from the speakers, when it becomes an unwanted echo in the other end. There are also many other elements causing problems in sound. Background noise, excessive distance to the sound source and large dynamic scale of an instrument are challenging for the system.

Preventive features like AGC (automatic gain control), AEC (automatic echo cancellation), noise fill and noise reduction are very important for business meetings, but if designed for speech only, they can be destructive to the sound when it comes to musical performance. If the room acoustics, microphone technology, audio processing and speakers are not carefully chosen, problems in the sound can be expected.

Other aspects that inevitably differ from a traditional local teaching situation include several problems caused by delay, the inability to fix posture by touch, the difficulty to see fast and small details like fingerings while playing, the extra effort caused by inability to share the same physical music sheet. On the other hand many times these problems are not crucial, though they may consume time more than in traditional situation.

For the transmission delay, there is a very fundamental[3] problem: speed of light[4]. If and when Internet works largely through optical fibers, the theoretical minimum delay for a connection to the opposite side of the Earth is about 90 milliseconds, but in practice at least 100-200ms just for the Internet fiber travel at the highest Earth distances. The maximum tolerable delay for duet playing is around 25-75ms. At 5ms, delay is usually not perceived. At 25ms it is possible to easily notice delay in critical applications. 50ms is already quite distracting for duet playing.

The speed of light in an optical network fiber is around 200 million meters per second[5]. In practice the speed is approximately 1ms per 100km (in the best case 0.5ms per 100km). Read more about delay on chapter '4.1 How to test latency (transmission delay) in a video conference?'.

Musical instruments and organic sound sources have a three-dimensional sound radiation pattern meaning that different frequencies are emphasized on different directions.

[3] There are scientific attempts to achieve superluminal communication. However the consensus is that faster than light communication, special relativity and causality cannot coexist – which is a fundamental problem for communication without delay.

[4] The speed of light: http://en.wikipedia.org/wiki/Speed_of_light

[5] Refraction in an optical fibre: http://en.wikipedia.org/wiki/Optical_fibre#Index_of_refraction

This results in a vivid experience of sound when the sound reflects from all surfaces in the space. The ear is very sophisticated at picking up directional information, so when recorded on a microphone and when played back on a speaker, the ear will in many cases easily spot the difference to original sound regardless of microphone and speaker quality. However, as many will agree, music recordings can still sound very detailed and rich, so also many subtle details in the sound can be reproduced through a video conference system as long as all the essential parts of the system are well-designed.

In many ways, the implementation of distance teaching technology is about balancing between different features such as audio and video quality, equipment price, setup workload, need for maintenance, need for live mixing and easiness of use. For example if the setup is carefully optimized for violin and the sound quality is optimized to a high level – this usually requires very extensive planning by the audio engineer. Also in that scenario the set up may be then relatively constricted: the system may be calibrated for violin only and changing the instrument will cause problems. Or the system may get hard to use if multiple scenarios are required. Designing an easy to use, universal studio may mean more freedom for the player to move furniture or change instruments, but this may be at the expense of quality. The goal is to come up with a perfect compromise where the experience is as close to natural local teaching session as possible. From the organizers point of view, an automatic system with easy and fool-proof interface – even usable by the student or the teacher alone, cheap but high quality, would be the optimum. For the engineer, the system should be extremely reliable; otherwise they will get complaints when something goes wrong even when there is nothing they can do about it. The technology is improving, but currently there is still a lot of manual work in order to achieve a good result[6].

As with many any advanced pieces of equipment today, user interfaces are often challenging. Not everything is automatic, so lot has to be done manually in the interface and that may become difficult for non-technical people. For video conferencing, simple calling, volume setting, camera controlling or other simple functionality may be familiar from cell phones and other well-known technology, but most of the deeper functionality and setting is practically restricted to engineers only. Other challenges are the reliability and stability, other wide issues in technology in general. Compatibility is also very important since there can be no connection at all if the protocol used does not communicate with the other end. For distance music teaching, several solutions[7] are used around the world, and they are not directly compatible which each other. H.323 endpoints should be compatible with each other, but in practice there are compatibility problems such as the audio codec mismatch, video quality decline, aspect ratio problems, connectivity problems and so on.

[6] There is a comprehensive thesis about the challenges, written by Alexander Carôt in May 2009, 'Musical Telepresence – A Comprehensive Analysis Towards New Cognitive and Technical Approaches': http://www.itm.uni-luebeck.de/users/carot/Docs/dissertation_AC.pdf

[7] Popular solutions are: hardware H.323 (by companies such as Polycom or Tandberg), ConferenceXP, DVTS and Skype

1.3 Standalone video conference hardware

The standard for regular business meetings and a commonly used way of doing distance music teaching is called H.323[8]. H.323 endpoints (also known as terminals or codecs[9]) allow for example multi-point video calls, camera control (local and far end) and presentation video channels. Microphones, video camera, display and speakers are the main peripherals connected to the endpoint. Most can be remote controlled via web browser and new models have a lot of versatile functionality and customizability.

Polycom has written a paper about their approach to music teaching, the Music Mode[10]. One of the most notable instrument teaching users of Polycom is Manhattan School of Music[11]. In Vi r Music, mostly Tandberg was used. Huawei, with their 256kbit AAC-LD audio is one of the other interesting contenders. More manufacturers and comparisons can be found on the Online Extension[12] of this Handbook. Additionally to other unmentioned companies providing H.323 equipment, there are solutions incompatible with H.323, explained in chapter '1.4 Software for distance music teaching'.

A standalone H.323 solution may be relatively expensive. However, it does bring certain important features that may not exist in other solutions such as H.323 video conference on a computer. The latest and best model of a standalone H.323 has traditionally been expensive, but it is also powerful. High processing power is needed to encode and decode today's high resolution video and that needs to happen fast because there is already delay in many stages of the transmission. An important quality of a standalone endpoint is its independence. In some cases it can be quickly turned on from standby mode via remote control and the call can be made immediately. Standalone terminals are relatively stable and since they are used only for the conference, the chances are less for settings being wrong or unit crashing because of external reasons.

On the other hand the stability on a computer is often worse since there are a lot more options which brings more room for unexpected problems. A computer may also struggle in encoding high resolution video unless the processor is extremely powerful, while typically the delay may still be higher than on a standalone endpoint which is optimized for video processing purposes. The relatively new technology, Scalable Video Coding[13] (SVC) has advantages and may allow lower latency (delay). SVC is available in products by Vidyo and Radvision among others. The problem with SVC today is incompati-

[8] H.323, the set of protocols for audiovisual communications: http://en.wikipedia.org/wiki/H.323

[9] Codec = coder-decoder device and in video conferencing it refers to the unit handling the incoming and outgoing video

[10] http://www.polycom.com/global/documents/whitepapers/music_performance_and_instruction_over_high speed_networks.pdf

[11] http://www.polycom.com/company/news_room/press_releases/2009/20090414.html

[12] Video conferencing and streaming comparison: http://tinyurl.com/virmusic0

[13] Scalable Video Coding: http://en.wikipedia.org/wiki/Scalable_Video_Coding or a video demonstration: http://www.radvision.com/Visual-Communications/Video-Communications-Technology/Scalable-Video-Coding/

bility with H.323. The industry around SVC is developing, for example Polycom has announced[14] developing SVC.

Besides the H.323 standard, there is another common protocol SIP, but that is less recommendable for instrument teaching since it is not compatible with H.323 (except via gateways, which may cause additional delay or other problems). However, some H.323 endpoints support SIP as well. Instead of H.323, SIP is used in products such as Tandberg Movi and LifeSize Desktop. They are computer software programs.

In Vi r Music we chose to use Tandberg Edge 95 in the beginning of 2009. At that time the newer Tandberg C90 was just released, but it was much too expensive. For instrument teaching, one factor in choosing between manufacturers is the compatibility of the audio standards which comes down to choosing between AAC-LD and G.722.1C. They are not compatible with each other but from the currently supported endpoints they are the best options in the sense that their cutoff frequency is 14kHz or more. Human ear ranges easily up to 16kHz and up to around 20kHz for most children and some adults. Therefore for the best quality, a cutoff frequency of 16kHz to 20kHz should be the minimum. Audio codecs[15] like CELT or FLAC are not yet usually supported even they could have some advanced features[16] like low algorithmic delay and unrestricted cutoff frequency.

Tandberg was a good option in Vi r Music because it has phantom powered XLR input connection for the microphones. Tandberg Edge 95 has 24V phantom and C series has the full 48V phantom[17]. When the endpoint has only line inputs, a mixer is necessary in order to use microphones intended for classical music recording. With an endpoint usually a table microphone intended for business meetings is included. That may be used in basic teaching, but is certainly not the best solution for high quality live music. If the acoustics, speakers and audio processing are top quality, then the microphone should certainly be a high grade microphone intended for classical music recording.

Avoiding extra peripherals is good for stability. When there are no knobs to adjust and no cables to disconnect, the risk of unexpected problems is lower. In this sense, Tandberg's XLR inputs help a lot. A mixer or other equipment could be hidden inside a cabinet so that nobody will accidentally change the settings. In a room used by many different students and teachers without a technical supervisor the importance of simplified interface is magnified. When there are no phantom powered inputs, it means there are also no microphone preamplifiers. That is the case with all models from Polycom and LifeSize in 2010.

[14] Polycom SVC: http://www.polycom.com/company/news_room/press_releases/2010/20101108_2.html

[15] Audio codecs comparison: http://en.wikipedia.org/wiki/Comparison_of_audio_formats

[16] There are many audio codec subjective quality comparisons available on the Internet; here is one interesting comparing codecs such as MP3, G.722.1C, AAC-LD and CELT: http://www.celt-codec.org/comparison/

[17] In our tests 24V was ok for Neumann TLM-103 and some other microphones

The functionality of an endpoint today is quite versatile and comparing products might not be simple. Usage of Polycom's Siren[18] audio codec is controlled by Polycom[19] while AAC[20] is more open and found in several free software applications. In that sense, AAC is more suitable for open, educational purposes and allows more space for future applications. For the distance education community, a mutual standard would be very practical. Today the community unfortunately consists of different types of technical solutions, some of which are not compatible with each other.

The datasheets or specification sheets for endpoints are usually easy to read and contain lots of useful information. There may be some essential functionality that can not be read on the sheet such as local image delay and dynamic range for the audio gain control. Naturally it is best to test each potential machine in a real life situation. The easiness of the screen user interface and clarity of the hand remote control is often fairly similar between all modern machines. It is good to know that the endpoint functions can usually be controlled from an external web browser as well. That allows for example the changing of screen settings without having to display menus on the screen during teaching.

Some of the important features to test when choosing an endpoint could be as follows:

1. It should be compatible with the distance education community at maximum quality. (The data sheet specifications can usually be trusted; however full compatibility can only be verified by testing.)

2. For the best video quality and processing power, the latest models are the best. However the highest model may not be necessary since not great amounts of input and output connectors are usually needed for basic instrument teaching.

3. As low latency (delay) as possible is needed, for local image and for the total delay for receiving and transmitting. No delay for the local (self) image would be the best (but with today's digital technology, that is unfortunately rare).

4. The quality of audio processing: echo cancellation algorithm quality, noise reduction quality, automatic gain control and preamp quality. The processing should not remove relevant content from the sound, it shouldn't add noise and it shouldn't diminish the dynamic range excessively or at all. The audio compression and general quality should be high.

5. Integrated preamplifier or not: determines whether external preamplifier is needed (if microphones for classical music recording are to be used instead of the default table microphone).

[18] Polycom Siren codec: http://en.wikipedia.org/wiki/Siren_Codec#Licensing

[19] There is also G.719, more advanced than Siren22, but usage also needs a license, from both Polycom and Ericsson: http://en.wikipedia.org/wiki/G.719 (G.719 is not implemented in products yet in 2010)

[20] AAC codec: http://en.wikipedia.org/wiki/Advanced_Audio_Coding#Licensing_and_patents

6. The video frame rate should be at least 20–25 fps (preferably 30-60), otherwise details like fingerings may be impossible to see. There should be no video/audio desynchronisation even at more challenging conditions like in a multipoint conference.

7. The camera control should be smooth and responsive, it will be difficult to elegantly position and zoom the camera if there is great lag or the control is bad.

8. The participant video layout customization should be intuitive and versatile. Less of borders or other wasted screen space is good (cropping images may sometimes be beneficial). Especially on multipoint sessions, it is important to be able to easily place the images in a suitable way.

9. If you intend to use your endpoint as a multipoint bridge, high processing power is needed in order to send full quality video to all participants. Sending FullHD at 60fps to three or more participants is currently at the top of the line, but may be expensive.

1.3.1 Polycom vs. Cisco–Tandberg

Both Polycom and Cisco–Tandberg, especially the new models, are possible for distance music teaching. However, there are some advantages and disadvantages in both. Of course, some features are improved and updated from model to model or even as firmware updates so the situation is changing. Here some of the main differences between Polycom and Tandberg are reviewed.

Tandberg has phantom powered XLR input connections; Polycom has only line level audio inputs and a proprietary Polycom microphone input. The missing phantom powered XLR input makes the use of a mixer or a preamplifier compulsory whereas it is possible to directly[21] connect a high grade XLR microphone to Tandberg. Neither Polycom's nor Tandberg's default table microphones are good enough for high grade audio in music teaching. Polycom and Tandberg both support chaining endpoints for multimonitor Telepresence. Without chaining, Polycom's current maximum is 2 discreet DVI level video outputs while Tandberg's maximum is 4 discreet DVI level outputs.

Polycom has a very powerful remote control transmitter, you can point to the back wall and it would work fine. Tandberg requires more direct pointing. Polycom HDX has ping and traceroute commands available in the endpoint interface, they may come in handy. Polycom has 'People On Content'[22] green background chroma key mode and some

[21] On Tandberg Edge 95, the phantom is only 24V and not 48V, but with many microphones that is not a problem; the microphone works and sound quality does not degrade. Also on Edge 95, the digital gain range is slightly too limited, causing the signal to be too hot on sensitive microphones if the performer is playing loud. The digital gain range on Tandberg C series is wider reducing or eliminating that problem.

[22] Polycom People On Content:
http://www.polycom.com/products/telepresence_video/accessories/hdx_accessories/people_on_content.html

firewall traversal and port range locking functions. By default, Tandberg shows split-screen (side-to-side) images in 16:9, while Polycom strips the sides and shows two 4:3 images. The same logic applies also on multi-point images. Both will show 16:9 at full screen mode (with the default 16:9 camera and a setting of 16:9). Polycom's 4:3 split-screen uses more surface area of the screen, but has more black space between images and images have relatively bright, blue borders around the images where Tandberg is more subtle with grey. Both have good web browser control interfaces.

Tandberg C series offers some important flexibility not found in Polycom. The free Cisco TC Console[23] allows for example pixel-accurate positioning and resizing of layouts with a graphical user interface running over Java. The Video Compositor is extremely useful. In the same application, the Audio Console allows modular routing of chosen audio inputs to chosen outputs. There is also a graphical equalizer and 8 customizable presets which can be separately set on any input or output. Polycom has a simple bass and treble setting in the endpoint interface. Tandberg is currently more expensive than Polycom when similar level packages are compared.

Tandberg's adds flexibility with it's API[24], which can be used to set parameters like "EchoControl Dereverberation" or "EchoControl NoiseReduction". Like the Tandberg API, Polycom's API[25] works through telnet or serial port. Polycom offers commands such as "peoplevideoadjustment stretch" (stretch aspect ratio), "contentsplash" (splash screen toggle), "pip" (e.g. set near-camera to one of the corners), "camera" (set or get camera settings) or "mpmode" (sets or gets the multipoint conference viewing mode, e.g. discussion = split screen, fullscreen = current speaker on full screen or auto = if one site is talking for 15 seconds, the speaker appears full screen).

Tandberg's older model, Edge 95[26] doesn't have an optimal echo cancellation algorithm. There is a low-pass filter gate functionality, which will unsuitably for live music change the tone color drastically back and forth. Also the noise fill and other features are not optimal. However, the C series[27] has less or none of those problems and no gate problem at all. Polycom's music mode has noise fill, noise reduction and automatic gain control disabled. The same applies to Tandberg C except for the noise fill.

---

[23] Cisco TC Console: http://developer.tandberg.com/web/guest/tools/integrators/audio-console

[24] Tandberg API Guide for C90 version TC3.0: http://tinyurl.com/TandbergAPI-TC30

[25] Polycom's Integrator's Manual, chapter Using the API: http://supportdocs.polycom.com/PolycomService/support/global/documents/support/setup_maintenance/products/video/hdx_irm.pdf

[26] Some of the very important settings to use on Edge 95 for live music: Call Quality / All codecs (algorithms) on + max upstream the highest unless that causes problems + Video quality Motion, Audio / Inputs Mixer Mode Fixed + AGC Off + Inputs Level Settings Mic level settings so that it doesn't go to red at all. Carefully check the manual for stereo settings – it's a bit tricky but make sure you set it right. Note that AUX inputs do not have any echo control available.

[27] Based on our email exchange, we know that Cisco–Tandberg is developing their audio processing and are aware of the needs in distance music teaching. Currently noise fill cannot be turned off, but there might be a switch for that in a future software update.

According to our tests, the Polycom table microphone ('Polycom Microphone Array' that comes with HDX 8000) restricts frequency bandwidth to 16kHz contrary to the specification which says 22kHz. However, the VCR audio output and the main audio output do reach up to 22kHz when AUX signal input is used instead of the proprietary table microphone. Tandberg Edge 95 and C series reach up to 20kHz with the provided table microphone, custom microphones and AUX input signals.

1.3.2 Multipoint and bridges

The multipoint functionality of an endpoint usually costs extra, so the need should to be considered. It is mandatory only for the bridging endpoint, while other endpoints can connect to the bridging endpoint even without the functionality enabled. In the Vi r Music project, multipoint was used mainly so that other participants were listening and watching a point to point session. Then the floor was given to another location and the next student would start. Särestö Academy did also violin teaching to children at many locations simultaneously through the multipoint. The appearance of a multipoint call is typically 3-4 sites divided on one screen or one site per screen. Tandberg is currently one of the best to customize the screen layout in great detail. Master classes can be recorded and published in the Internet on the same day: the audience can view the whole class within a few hours in full quality. There are also solutions for live streaming of video conferences so that the audience could watch the master class live on the Internet.

The multipoint functionality is installed in an endpoint which will then act as a Multipoint Control Unit (MCU) aka bridge. This means other parties can call the MCU or the MCU can call the other parties and the MCU will send all participants' video streams to all participants. An external bridge can also do that. An external bridge can also connect devices otherwise incompatible with each other. For example via a bridge, ISDN can connect to IP (Internet Protocol). With bridges, make sure all bitrates are set to maximum, otherwise the audio and video quality may drop significantly. The bridge will naturally also cause delay.

As far as we know, on paper currently the Codian[28] (MCU 4500) bridge is the only solution for Polycom–Tandberg calls with higher than 7kHz cutoff frequency. Direct call Polycom–Tandberg[29] will use G.722 codec (7kHz cutoff) and since the Codian doesn't support Siren22, on paper the best result currently has the cutoff at 14kHz. However, we tested to different Codian bridges, mainly the MCU 4505 and the result was extremely questionable. The bridge did convert AAC-LD to Siren14 and vice versa, but that did NOT effectively raise the cutoff frequency. The direct Polycom–Tandberg was about

[28] Codian MCU 4500: http://www.tandberg.com/video-conferencing-multipoint-control/tandberg-codian-mcu4500.jsp

[29] We noticed that in a mixed multipoint call with Tandbergs and one Polycom, the Polycom sent an incompatible video aspect ratio: the image was too narrow and it couldn't be stretched to 16:9. This most likely has to do with Polycom preferring 4:3 at split-screen while Tandberg does 16:9 in split-screen. This seemed to happen only between Tandberg C-series and Polycom, not if Tandberg Edge 95 was used. A positive surprise was that the Tandbergs all received AAC-LD from other Tandbergs and G.722.1 from Polycom – so the system allowed for mixed audio codecs, maintaining the best one for Tandbergs and a separate codec for the Polycom endpoint.

7,5kHz cutoff and through the bridge, it was about 8,5kHz. Read an important note about frequency folding on chapter '1.3.6 Testing audio processing quality'.

As the MCU sends all streams to other parties, it also needs relatively much processing power. Multiple videos have to be decoded and sent back to all parties. Older endpoints typically don't have enough processing power to send multiple videos to many participants at maximum quality. Frame rate may substantially drop. Later endpoints may have enough processing power to provide 1080p/30fps or 1080p/60fps for all participants.

1.3.3 Audio hardware for standalone video conference

Mixer has a microphone preamplifier and also plain preamplifiers are available. One of these is needed with e.g. Polycom or LifeSize endpoints if a phantom powered microphone is used instead of the default table microphone. Usually a voice conference table microphone is not good enough for capturing live music at the highest quality.

Microphones, preamplification, audio interfaces and video recording hardware:
http://tinyurl.com/virmusic5

Microphones can be chosen quite largely based on the same reasons and principles as in classical music recording. However, new challenges are introduced if the microphone should pick up as little far-end feedback echo as possible due to echo cancellation related problems. Sometimes the microphones are not wanted to be seen in the image and in a mobile setup the mobility, acoustics or some other reasons may impact the requirements for the microphone type. There isn't a single recommendable general microphone setup that will work great on all possible distance studios and situations but audio engineer has to choose it based on the local requirements. However, the link above provides information to equipment that has been successfully used. Many sets of instructions to the art of stereo microphone techniques[30] or choosing microphones[31] and microphone[32] types[33] can be found on the Internet. One of the important principles is the proximity effect[34], which increases low frequency response when a sound source is close to a microphone.

A short summary of some basic microphone types for distance teaching:

- Large diaphragm condenser microphone: may have the most pleasing sound quality for a classical instrument, but as it is sensitive, it will also pick up room sound and echo (so good acoustics are required)surface and wall materials (further defines the acoustic characteristics)

[30] Stereo microphone techniques by Nuno Gama:
http://www.youtube.com/watch?v=_GU0pBuOrWs&feature=related

[31] Microphones overview by Jeff Towne: http://home.earthlink.net/~rongonz/home_rec/microphone.html

[32] Pickup patterns and microphone types by Ron Gonzales:
http://home.earthlink.net/~rongonz/home_rec/microphone.html

[33] Microphone types by Nuno Gama: http://www.youtube.com/watch?v=MACpIFBlGpg

[34] Proximity effect: http://en.wikipedia.org/wiki/Proximity_effect_%28audio%29

- Small diaphragm condenser microphone: compared to large diaphragm, it typically has less sensitivity, higher self noise, higher sound pressure level handling, higher dynamic range and capture high frequency content and transients well

- Dynamic microphone (small diaphragm): picks less echo and reverberation as it’s less sensitive, but the sound quality and detail are worse compared to condenser (transient response is weak and dull)

- Shotgun microphone: May have both fairly good sound and high directivity (less echo), but may allow less freedom of movement for the player (note that it will pickup echo and reverberation coming from the back wall behind the player and possibly some general room sound as well, depending on the model)

- Miniature condenser microphone (clip-on mic): Certain models pick only sound from very near sources, but the downside is less freedom because of the microphone attachment, cables and level adjusting (sensitive to distance and angle changes: if the button microphone is attached to collar and the person speaks upwards, the sound may vanish)

A different microphone for certain instrument may be beneficial, on the other hand one versatile kind of microphone may work well with many different instruments. Mono and stereo capturing and transmission through the endpoint are both valid for distance music teaching. Stereo will naturally sound better so it should be used when possible. In addition to the instrument microphone, a separate microphone for speaking was tested in Vi r Music, but the result wasn't perfect. Having a miniature microphone on the collar for example may improve speech sound quality and audibility, but it will cause hassle such as checking the levels, positioning the cables nicely and having to refrain from speaking away from the microphone. However, especially in certain cases, it may be a great challenge to get the speech sound level audible enough in comparison to the instrument sound. This problem may be apparent with loud instruments like a French horn or large instruments like a piano.

Using a highly directional shotgun microphone in a highly damped room (without reflective surfaces) may allow the disabling of echo cancellation. When doing that, the problems of echo cancellation lowering the audio quality will disappear, providing the feedback echo is not picked up by the shotgun microphone. This may not be easily successful unless the room acoustics are rigorously treated, but if the echo cancellation algorithm is not optimal, it is a potential solution. In general, either a cardioid (large or small diaphragm) or a shotgun microphone will produce the best sound quality in distance music teaching context. A dynamic microphone is also popular in situations where room sound and speaker sound leakage is avoided, but may have lower sound quality. Then again, echo is highly disturbing so lower sound quality will easily win over good sound quality with a confusing echo present.

The problem with echo cancellation degrading sound quality could also be removed if communicating or playing happened only in walkie-talkie manner, meaning that only one party talks or plays at a time while the other party has microphone off at the time. This is however highly unpractical, so it should not be considered an option. Walkie-talkie principle is what is used in EchoDamp, just in a optimized and softer way. The

same principle is often also more or less included in acoustic echo cancellation, though AEC uses principally a substraction of the filtered far-end signal in order to cancel out the unwanted echo.

1.3.3.1 EchoDamp (and other external echo cancellation)

EchoDamp[35] is a software program used to remove or lessen the video conference echo, which is a problem when local site is sending sound which is played back on the far end speaker and then again picked up by the far end microphone and sent back to local site unless the echo is "damped". EchoDamp aims to preserve very high audio quality but requires careful planning and set up for the whole system, including acoustics. EchoDamp can be used with standalone endpoints but also with any system using audio streaming such as Conference XP or JackTrip. EchoDamp requires an audio interface with a minimum of 4 discrete analog inputs and 4 discrete analog outputs, some cables and some time to set up and calibrate the system. If you're an audio engineer, the set up time is probably a few hours at maximum, including the calibration with receiving system. The software is sleek, stable and logical, although much more complex than just switching echo cancellation on or off.

EchoDamp runs on a computer (Windows XP to Windows 7 or OSX) and with slower computers it is recommended to use the computer exclusively for EchoDamp and not something else simultaneously. It is, in theory, possible to use for example ConferenceXP and EchoDamp on the same machine, but it may get tricky with the audio routing and CPU usage. However, in OSX, using JACK[36] network streaming and EchoDamp together on one computer was reported to work very well at low latency. JACK is also available for Windows and is popular for internal audio routing but not supported by all applications.

EchoDamp has a lot of functionality and versatility, but the simplified main features are:

Downward Expander (for removing echo coming to the microphone)

Ducker (for removing your own voice if it comes back)

Latency calibration (beep-based, essential to the optimized performance)

EchoDamp includes also other functions, such as hi-pass filter or audience mixer, essential at certain situations. With EchoDamp, the acoustic preparation is very important. The microphone should pick as little sound from the speaker as possible. This means that room acoustic design, placements of equipment and objects, microphone type, sensitivity and polar pattern play an essential role.

EchoDamp can be used and calibrated even if the other party doesn't have EchoDamp or any echo cancellation at all or if they use alternative echo cancellation method. It is

[35] EchoDamp: http://echodamp.com/ and EchoDamp User Manual: http://echodamp.com/support/manual/manual1.html

[36] JACK low latency audio system: http://jackaudio.org/

also possible to use EchoDamp to eliminate echo for both directions from one side, having no echo cancellation at the other side at all, still effectively removing most of the echo. But, since EchoDamp only does expanding and ducking, double-talk (simultaneous talk of both sides) is not possible with zero echo, opposed to acoustic echo cancellation used in video conferencing endpoints.

In our test, it was possible to talk effectively back and forth with good quality, having EchoDamp only on one side and no echo cancellation on the other. This could be done even with a cardioid microphone (Neumann KM184) and just 2 meters distance from the speaker, but in such situation one has to talk or play extremely near the microphone (perhaps max 20–30cm) and very aggressive expanding and ducking is needed, so the situation is far from natural or unrestricted. EchoDamp is effective, but to use it with really high quality microphones, still having some freedom of movement, requires careful planning and acoustic treatment such as diffusing and damping. A miniature microphone attached to the instrument or a low-sensitivity dynamic microphone will pick up less speaker and room sound and thus make the cancellation easier, but they don't usually give the best possible microphone quality.

In addition to EchoDamp, there are other echo cancellation solutions. Access Grid Support Centre[37] has written a tutorial about external echo cancellation, which can be divided into four categories: 1) Headphones, 2) Desktop echo cancelling microphones, 3) Echo cancelling PCI cards, 4) Rack mounted external echo cancellers. However, many of products from categories 2–4 are not designed for high fidelity live music. (If you find a really good one, please let us know at virmusic.blog@gmail.com.)

1.3.4 Displays and video projectors

Latency is an important issue on video conferencing. On TV monitors and projectors, the delay consists mostly of two factors: the input lag[38] (video processing lag) and the response time[39] ("pixel" lag). In some cases, especially when heavy processing is set on the display options, the input lag may be the culprit of causing significant latency (and therefore bad synchronization to audio if the audio is connected separately and the display doesn't adjust the synchronization). Typically with LCD, the input lag is around 10–120ms and response time is around 2–16ms. For a monitor with a low total latency, the safest bet may be to use a digital signage[40] display. They usually don't have slow post-processing. Plasma has a lower latency on average and CRT has practically zero latency, but compared to LCD or LED, they are usually unpractical because of other properties they have. Currently LED is more expensive than LCD and not necessarily worth the difference. It is important to try out different features extensively in order not to end up with a display that has surprising problems with certain functions.

[37] Access Grid Echo Cancelling Guide:
http://www.ja.net/documents/services/video/echocancellingandagfinal.pdf

[38] http://en.wikipedia.org/wiki/Input_lag

[39] http://en.wikipedia.org/wiki/Response_time_%28technology%29#Display_technologies

[40] http://en.wikipedia.org/wiki/Digital_signage

The display should not have heavy post-processing turned on in case it greatly delays the image, instead a low-latency "Game mode" may be appropriate. Note that different video inputs may have different latencies. For example DVI or non-native resolution signal might be heavily processed but VGA signal might not. There are several methods[41] to measure the latency, but they tend to require a lot of set up and peripherals. In the chapter '4.1 How to test latency (transmission delay) in a video conference?', a slightly different test is introduced, but the principals are the same. A reference (such as zero-latency CRT display) has to be properly set and the delay can be measured with a stopwatch[42] and digital camera or with a reaction time game, or it can be done with a latency utility found in a music video game such as Rock Band 2+ (automatic) or Guitar Hero 2+ (manual). Or just using a USB cable mouse and looking at the cursor can give quick idea of the severity of the delay. To go even further with display tests, you can see and try the Lagom LCD monitor test pages[43], though for music teaching it is adequate if the picture quality is fairly natural and the display delay is 0–10ms.

When acquiring a TV or a projector, check all necessary video modes (720p, 1080p etc.) with the final video source. Sometimes video signals are incompatible and the picture may not show at all or it may be incorrectly stretched or cropped. This may be corrected in the settings but better to check for any hardware incompatibilities. Sometimes the signal can be "fixed" with a HDMI splitter for example. If the display is to be used in a school environment, it's good to check the standby functionality. Make sure that the display turns off to standby automatically after the endpoint has sent a black screen. Otherwise the display may unnecessarily run for days if nobody switches it off.

Compared to a monitor, a video projector may have these potential disadvantages:

- less brightness (for maximum quality, a fairly dark room may be needed)
- fan noise
- needs a projection surface and placement without obstacles between projector and screen (also camera cannot be within the projected image area since the projected light will glare the captured image)

If you know what you're doing, you might want to take the time to optimize your display settings, minding to stay out of any settings that add to the input lag. New endpoints and their cameras have fairly good automatic color balance and brightness settings, but it is possible to adjust such settings both from the display and the endpoint settings. In the best scenario, all participants have similar color balances.

[41] http://www.pcworld.com/article/183928/find_and_fix_input_lag_in_your_hdtv_or_monitor.html

[42] Online Monitor Tests, including an input lag stopwatch by FlatPanels.dk: http://tft.vanity.dk/ (mind that many types of processing, such as resolution conversion, will cause delay)

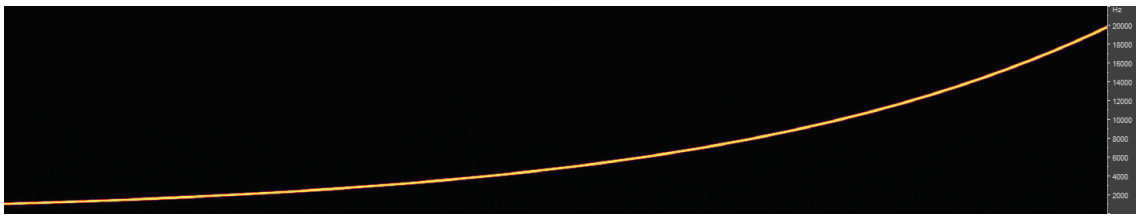
[43] http://www.lagom.nl/lcd-test/response_time.php

1.3.5 Peripherals for standalone video conference

There are a number of peripherals that could be meaningfully used in video conferencing. Solutions for viewing music sheets can be found on chapter '3.4 Music sheets'. For music theory or music history, a laptop can be used to present notation in a notation program or slide shows for music history, for example. Most H.323 video conferencing terminals have a presentation/computer video input for this purpose. Many video conferencing software programs support presentation too (check the Handbook Online Extension to find out which ones). External video cameras can be used as a means to instantaneously switch between different camera zooming positions, between students or between audience and student for example.

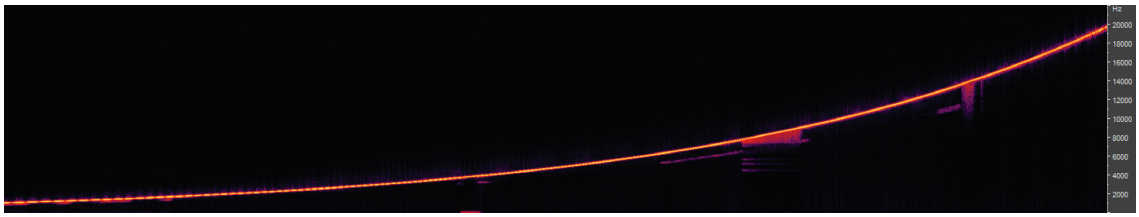
1.3.6 Audio processing quality: testing and results

The local player plays and the far end sends their audio signal eventually played through the local loudspeakers. The player and the loudspeakers are both generating sound waves entering the local microphone. That's where the chain of audio processing starts. The microphone signal is amplified by a preamplifier, either external or internal. Then we have AD (analogue to digital) conversion. Next is the fairly complicated internal processing and finally the processed audio is sent to the network in sync with the video stream. All of these steps can radically alter the sound quality. For the internal processing, some of the essential elements are: processing of dynamics and Automatic Gain Control (AGC), Automatic Noise Suppression (ANS), Noise Fill, Acoustic Echo Cancellation (AEC). We made a great number of audio analysis tests and comparisons with several types of test signals[44] between Tandberg, Polycom and other solutions. However, going deep into accurate analysis in this subject is a big task, and in this Handbook, at least this version, we'll not go there. However, here are some interesting examples of few tests:



A pure logarithmic sine sweep (linear spectrogram analysis)

[44] Test signals included a sine sweep, white noise with changing amplitude, one sample impulses, amplitude-modulating synthesized sounds to test dynamics, linear dynamic slides and number of live recordings of playing and singing. Live excerpts included both solo instruments and orchestral works.



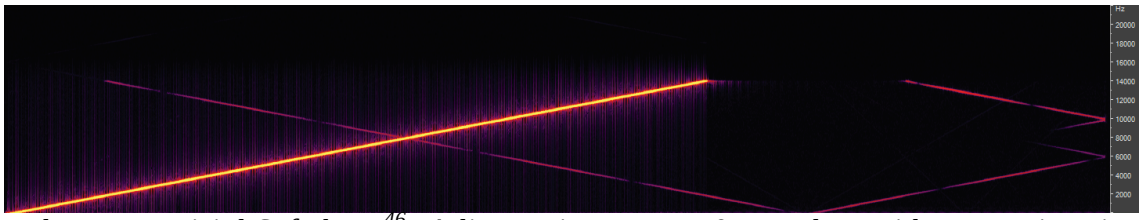
The same sine sweep as received through a video conference system with AEC, Noise Fill functionality and other processing enabled while the side receiving the signal is also talking at the same time making the situation realistic but complicated

Scientifically accurate testing AEC becomes complex when the whole back and forth loop is taken into consideration. The lower image above doesn't contain recorded talking, but the effects of talking at the side where the sine sweep was recorded. The visible noise occurred even as the sine sweep sending side didn't have their microphone connected. This shows that relevant measurement can be done even without having microphones on at both sides and not recording through a microphone but directly for line out instead. That way the results will be comparable to tests made elsewhere. Even a high quality recording equipment will affect the sound as well (especially if there is a problem like a ground loop problem), but not nearly as much as if a microphone is used to record sound from a speaker.

Completely other kind of approach, involving only subjective comparison by ear, is to A/B compare two systems with real live playing. That will reveal what happens in the ultimate, real situation.

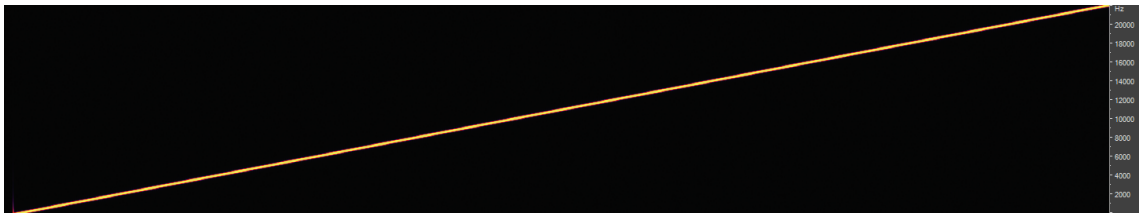
Besides problems with dynamics, artifacts and noise, there is one common problem for a great number of video conferencing products, including H.323 terminals and widely used commercial software. This came apparent when testing a wide range of products. The problem is in the frequency domain, called the Nyquist[45] folding or Nyquist aliasing. It causes the frequencies above the cutoff frequency to mirror back down. This is revealed when playing a sine sweep through a system, where the sine frequency range reaches over the maximum frequency or the Nyquist frequency of the recording system. The result will be a sweep sliding downwards after the cutoff frequency. That is a major flaw in the system and will dramatically distort the sound overall and make it sound messy. The correct way is to apply high quality low-pass filtering before sample rate conversion, but this seems to be neglected far too often. Also it is not uncommon to witness other highly distorting aliasing problems even with frequencies under the Nyquist frequency.

[45] Nyquist frequency and Nyquist folding: http://en.wikipedia.org/wiki/Nyquist_frequency

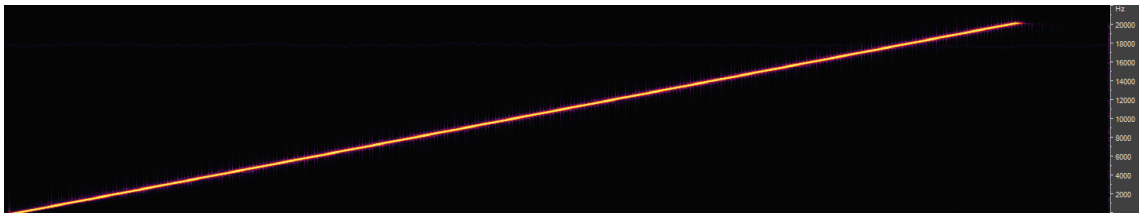


Polycom to Mirial Softphone[46]: A linear sine sweep 20Hz–22kHz with some mirroring problems and inadequate low-pass filtering above 14kHz, which is the specified cutoff frequency for the used audio codec, Siren14

Polycom–Polycom connection will not introduce any Nyquist folding problem. Similarly Tandberg–Tandberg doesn't have the problem either. But connecting Polycom–Tandberg is problematic. Not only does the audio codec drop to G.722 (about 7kHz), but Nyquist folding problem is also introduced. Frequencies around 8–13kHz are folded down, causing major distortion. Using a Codian bridge, converting Siren14 to AAC-LD and vice versa, does not change the situation much.



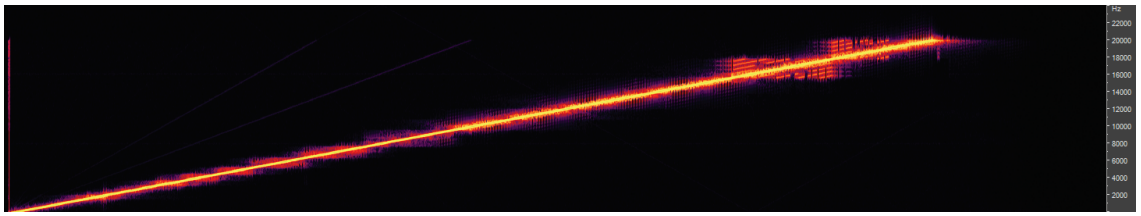
ConferenceXP–ConferenceXP, using uncompressed audio, sampling rate 44.1kHz, 1411kbps stereo (only one channel visualized), very clean[47] result (actual measured transmission throughput ~187 kilobytes per second)



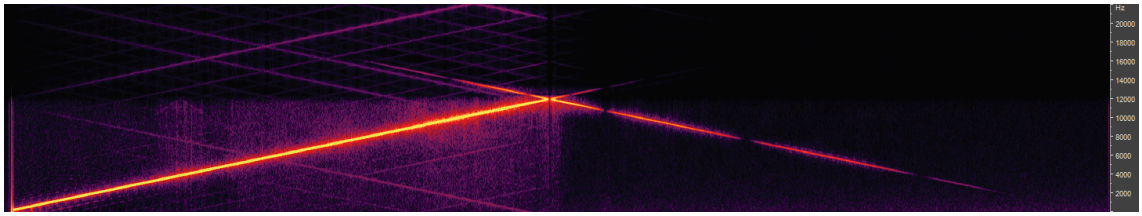
ConferenceXP–ConferenceXP, using compressed audio, sampling rate 44.1kHz, 64kbps stereo (only one channel visualized): cutoff frequency is at 20kHz and technical performance is good, although artifacts and degraded sound quality are apparent on live music (actual measured transmission throughput only ~12 kilobytes per second)

[46] RME Fireface 800 was used to play the sweep into the Polycom HDX 8000 AUX input, and after Mirial Softphone, Total Recorder 8 was used to record the result 'bit perfectly' (no conversion)

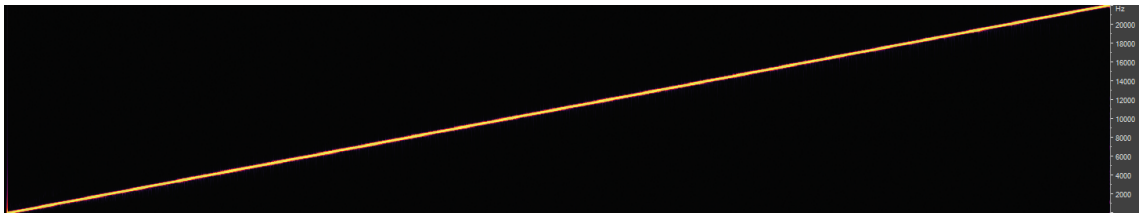
[47] There is a small wide-band transient present on the visual image at the sweep's start and stop. Those are a mathematically fundamental part of the formation of a signal and belong to the original signal as well.



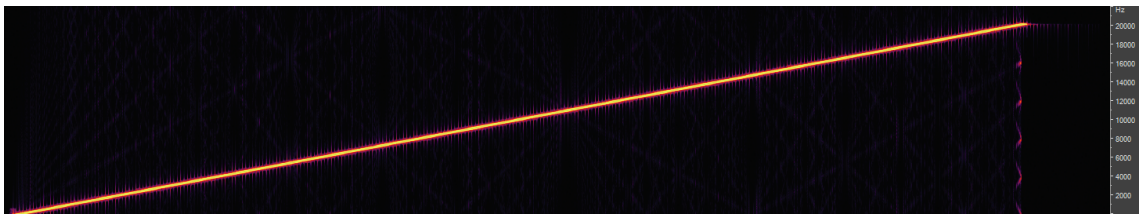
Soundjack–Soundjack, using CELT audio (high quality mode, ~25 kilobytes per second in mono mode), sampling rate 48kHz, cutoff 20kHz: For a low latency compressed solution, the quality is relatively good



Skype–Skype: cutoff is 12kHz and there is heavy distortion and Nyquist folding (automatic microphone and speaker level settings were disabled)



VLC–VLC, using AAC stereo at 128kbps: cutoff is at full 22.05kHz (sample rate is 44.1kHz), only very small artifacts (compared to uncompressed, lower sound quality will be apparent when listening to actual music or speech)



VLC–VLC, using WMA2 stereo at 128kbps: cutoff is at 20kHz and some distortion can be seen

Note that the sine sweep test is only a very small portion of proper audio quality analysis. There are many qualities in sound that are completely indefinable based on spectrogram analysis of sine sweep. A listening test as A/B comparison with different kinds of music clips is proper for overall quality comparison. However, some of these images do reveal clear flaws in the audio processing and are partially telling about the actual sound on music clips or speech as well.

1.4 Software for distance music teaching

There are amazing amounts of ways to connect audio and video over network, most however being quite unsuitable for distance music teaching due to limitations in audio quality, video quality, latency, stability and user interface among other functionalities. Here some of the most suitable software programs for high quality distance music teaching are briefly introduced. The web links and other information is available in the Handbook Online Extension[48]. Note that none of the solutions mentioned below are compatible with each other (for example ConferenceXP–DVTS or ConferenceXP–H.323 calls are not possible). Those video conferencing or audio streaming programs that don't have echo cancellation built-in have to use external echo prevention or headphones. The main downsides of headphones are the uncomfortable effects caused by the design (earpieces damping own sound, weight, cable etc.) and the fact that all who are in the room and want to hear need to use the headphones.

ConferenceXP

ConferenceXP allows fairly easy video conference with uncompressed stereo audio[49]. Moderate quality compressed audio is also available. Since the video bitrate can be adjusted, conferences at both low and high network bandwidths are possible. Minimum total latency is ~30ms end-to-end. Since there's no echo cancellation, the options are to have a carefully constructed shotgun microphone + acoustic damping/diffusion setup or headphones instead of speakers. Alternatively EchoDamp can be used. ConferenceXP supports USB and FireWire (IEEE 1394) cameras. It also supports end-to-end and multipoint through the provided Venue Service functionality. Multicast is supported to allow network optimization when applicable.

Pros: Free (and can work also with very basic equipment), uncompressed audio (no distortion), relatively easy to use, screen streaming, chat and other collaborative capabilities, multipoint capability.

Cons: No integrated echo cancellation, no camera control[50], minor audio clicks may occur even with good quality connections, Windows only.

Short manual: Choose your audio and video settings in Settings → Audio/Video and start the two-way conference at Actions → Start a Two-Way Unicast Conference. Both participants enter the IP of the other participant.

[48] Standalone, PC and Mac solutions for video conferencing and streaming: http://tinyurl.com/virmusic0 (press the tab buttons to find also all basic hardware and peripheral suggestions)

[49] ConferenceXP or other software running uncompressed audio with 0 or just 1 buffer is prone to minor clicks in audio if packets are lost due to network irregularities.

[50] If ConferenceXP supported features such as camera control, very low codec latency (<5ms), echo cancellation or the JACK API, that would simplify hardware setups a lot and thus lower the hardware cost considerably.

DVTS

Digital Video Transport System (DVTS) allows sending DV or HDV streams as it is. Audio is therefore uncompressed with 24kHz cutoff frequency. DVTS requires a dedicated network route at 30Mbps. DVTS sends DV frames, which include uncompressed audio and video together, and always runs at 0 buffers. Thus, if a packet is lost, you may immediately see a video artifact and hear a minor click in audio. DVTS does not include packet loss concealment for audio, so a highly reliable network is required for perfect audio without any clicks in the sound. Since (H)DV cameras output the compressed video (+ uncompressed audio signal) to FireWire at considerable latencies, a low-latency conference is not possible. However, DVTS has been successfully used a lot for music teaching.

Pros: Free, good audio and video quality.

Cons: No integrated echo cancellation, no camera control, not resilient to packet loss, high bandwidth requirement, relatively difficult to setup and use, relatively large minimum latency, no multipoint capability.

LOLA

LOw LAtency Audio Visual Streaming System is the first solution to allow audio and video sent over network at only 5ms latency. Video grabber and audio grabber are 5ms and in sync; LOLA can use 0 video buffers, 0 or 1 audio buffers and expects that network jitter is close to 0, e.g. it assumes the network is very reliable.

As LOLA is currently a research project by Conservatorio di Musica G. Tartini in Trieste and GARR, the Italian Academic and Research Network; the software is free, although not yet downloadable on the Internet by January 2011. A first public release is expected in 2011.

LOLA uses an industrial high speed (300fps) camera, black and white or color and currently runs on Windows XP and Windows 7 and certain video cards (BitFlow grabbers). The camera and video card are not expensive (less than 1000 €) but the software requires a high performance network service: either a dedicated network circuit or light path with a minimum 100Mbps up to 450 Mbps, or a GigaEthernet link, which is only available at certain locations. LOLA runs also on shared IP networks, at least in minimal configuration (95Mbps) but again it requires a very stable network service, currently provided only by the Academic and Research Network services. Note that even LOLA has very low latency, playing together properly is still possible only within 1 continent distances, e.g. a maximum of 3000km (since GE network latency is approximately 1ms per 100km).

Audio is 44.1kHz 16/24bit stereo or multichannel. Video is 30-60fps, 640x480 pixels. video jitter <3ms.

Pros: The lowest video latency available (5ms), at very low latency feedback echo is actually reverberation of the other space – echo cancellation may not be needed, not expensive even it needs special gear (high-speed camera, a video card and possibly a low-latency display), good audio and video.

Cons: Video is only 640x480 pixels, no camera control, no multipoint capability.

Soundjack

Windows version of Soundjack supports audio only, but OSX 10.6 has also experimental video built-in (currently only low frame rate and resolution). Minimum total latency of the audio, in case of a 32 sample buffer, is 2.4ms (excluding additional network latencies). A realistic practical value is 5.4ms as the default Soundjack setup. The interface is not extremely complicated, but some technical skills and experimentation are required. Requires ASIO sound card driver, but ASIO4ALL works so most cards can be used (real ASIO means less latency). The input or output routing are not built-in the software, you can only choose in/out devices and between 1 or 2 channels. With ASIO4ALL, you can choose the I/O routings as well. Soundjack uses either an external user list server or P2P mode without a server.

Pros: Free, good quality CELT audio codec meaning low-latency compressed audio (though some distortion and artifacts measured), multipoint calls for audio, fairly high audio quality (although certain distortion is measured).

Cons: Not the easiest to use, no integrated echo cancellation (and no camera control).

Audio only: JackTrip

JackTrip allows unlimited number of uncompressed audio channels to be sent over network at low latency. It is currently only available for OSX and Linux and uses linear sampling and redundancy to recover from packet loss, sending audio packets to the network as soon as the sound card can deliver them. No video or echo cancellation, but useful for any low-latency, high quality audio streaming purposes.

1.4.1 Simple solutions

Skype

Only very few solutions offer easy connecting to other participant at any circumstances with a random network connection. Skype is the most powerful in this sense because it typically can connect behind NAT[51] and with limited amount of open ports[52]. Skype also has the biggest user base and user directory. In Skype it is sometimes possible to find other users even if you know only their real name. In other ways too, Skype is one of the most convenient and easy to use solutions available. Minimum total latency is ~18ms end-to-end. Skype may use Internet bandwidth and computer processing power even when no calls are in place due to the peer to peer and supernode model the Skype network runs on.

[51] Network address translation: http://en.wikipedia.org/wiki/Network_address_translation

[52] Skype's firewall settings: http://www.skype.com/intl/en-us/support/user-guides/firewalls/technical/ – that will also give a basic idea on some of the connection issues and possibilities

However, the audio on Skype is nearly as good as on the previous chapter's programs. Skype's echo cancellation has quite a drastic walkie-talkie effect: loud double talk (talking at the same time) is not possible since loud talking will momentarily disable remote sound. The echo cancellation is effective but as it is drastic, sound quality is reduced. Even if 'Automatically adjust microphone/speaker settings' are off, it still has quite heavy processing on the sound. The cutoff frequency is 12kHz and the quality suffers from somewhat heavy distortion and Nyquist folding. Skype supports USB cameras, newest version may also support IEEE1394 aka FireWire.

Pros: Free, relatively user friendly, widely used, supports Windows/OSX/Linux.

Cons: Relatively bad sound quality, sound processing is automatic and user friendly but audio is quite distorted, multipoint is not free.

Mikogo, Vsee and TeamViewer

In addition to ConferenceXP, the best free (at least for non-commercial use) web collaboration tools are probably Mikogo, Vsee and TeamViewer. Mikogo allows efficient desktop sharing and remote control (no sound) and Vsee is an easy to use multipoint video conferencing application, though sound quality is poor. TeamViewer is excellent for remote control and it has webcam video capability as well, though sound quality is very poor.

1.4.2 High quality streaming solutions

VLC

VLC has strengths in versatility: It is very popular, continuously developed, has a wide range of functionality from simple playback to streaming with subtitles and post-processing effects. It includes a wide array of codecs built-in. VLC is available for all major operating systems. The versatile nature comes with downside though. As there are so many features and possible combinations how to use it, it gets complex and the settings get slightly out of hand. There are so many settings that finding the optimal ones gets quite tricky. However, to simplify, here are some suggested methods how to use VLC for streaming. Look at the VLC codec table[53] and you'll get an overview of what is possible.

According to our test, one of the best combinations producing good video and good audio at low latency while still not using too much CPU power was to use WMV2 and WAV (requires more bandwidth) or WMA2/MP3 (it probably depends on audio material which is better at given bitrate). So here is how it was possible to stream compressed live webcam and audio to another computer at 550ms latency:

Sending computer: Start VLC. Press ctrl+c for Capture Device streaming. DirectShow is correct. Choose your Video device and Audio device (not everything is compatible,

[53] VLC streaming codec table: http://tinyurl.com/virmusic2

for example a USB microphone may not work). Press alt+s to move to Stream Output (or click the small arrow at bottom to do so). Press the Destinations block. Choose to Display locally and choose your streaming method (for example MS-WMSP or UDP). In this case, we need to use HTTP. Press Add. Now press the Edit selected profile at Transcoding options. In this case, encapsulation has to be ASF/WMV. From Video codecs tab, [x] Video and [ ] Keep original video track is correct. Codec should be WMV2. Choose a low bitrate, for example 800 kb/s for starters. Frame rate 0,00 fps means use original frame rate. For the resolution, the simplest way is probably to use a fixed Width. So enter 200 for the Width, for example. Scale and Height are left 0, which means they will be automatically adjusted according to the Width you entered. At Audio codec, [x] Audio and [ ] Keep original audio track is correct. Codec should be WMA2 for starters. Bitrate can be 128 kb/s for example. Channels is 2 and the Sample Rate is a tricky one: you have to choose correctly between 44100 and 48000 or otherwise there may be no sound. VLC has problems with sample rate conversion. So choose 44100 for starters (and come back to this if there was no sound). Press Save. Now your computer should be streaming (more precisely, waiting for somebody to connect). You will not see any network activity yet.

Receiving computer: If you have heavy restrictions in your firewall, disable it (sending and receiving computer, also routers if necessary). It may work without disabling, but if you encounter problems, firewall is something to check. Start VLC. Press ctrl+n to open Open Media / Network. The URL is “http://SenderIP:8080” (without quotes). Press Play. You should start to see the stream in a few seconds. The quality should be good and latency should be about 550ms + network latency (which should be only a few milliseconds if the computers are very close). Feel free to adjust bitrates and resolution at this point. They were set low just to ensure that CPU and bandwidth usage won’t cause problems.

USTREAM

Besides VLC, there are a large number of ways to stream video. One of the simplest solutions is to use USTREAM[54]. It is currently free and supported by advertising revenue. By the time of testing, it used Flash for playback and streaming. At the simplest method, user can use their webcam and microphone to stream on a simple web browser interface. There is no restriction on the number of viewers. In other words, all the server work is conveniently provided by USTREAM. The minimum latency is about 2–3 seconds. The quality settings are customizable and allow high quality.

More alternatives to VLC and USTREAM are listed on the Handbook Online Extension.

[54] Description of USTREAM: http://en.wikipedia.org/wiki/Ustream

1.4.3 Peripherals for software video conferencing

Video cameras

Some software programs, such as Mirial Softphone, are capable of controlling remote camera within the H.323 protocol (more specifically H.224/H.281 far-end camera control), but software implementations often don't enable the controlling of local camera. At this stage of computer video conferencing involving a consumer level webcam, one may have to do with a manual camera control. Webcams such as Logitech QuickCam Orbit AF do have motorized tracking, but as far as we know it cannot be controlled by the other party. Compatibility for all cameras has to be checked since there are many standards. USB 1.1 is relatively slow compared to FireWire for example, so low frame rates can be expected on USB when the resolution is high. "USB 2.0 High-Speed" has also proven to be usually slower than FireWire-400 in practice. For video transfer comparison, it is relevant to find out what exactly is transferred: compressed or uncompressed and what kind of processing and latencies are involved already before the signal is leaving the camera. For H.224/H.281 remote camera controlling, look at PTZ (Pan-Tilt-Zoom) cameras. Good ones are expensive, but can be far-end controlled if the software fully supports it. There are also remote controllable IP cameras.

FireWire is a relatively safe standard for video capturing but still not supported by all software. HDMI can be converted to FireWire (DV or HDV[55]), refer to the Handbook Online Extension[56] for that. Lower quality composite and s-video connections may be used in certain situations as well. Some methods for video transfer, including FireWire and USB will introduce latency. Check the latency before purchase.

Audio peripherals

To achieve appropriate sound quality, you need good loudspeakers or headphones, one or two good microphones and a way to connect the microphones to the software. For loudspeakers, headphones and other suggestions, please take a look at the Handbook Online Extension. For microphones, you have mainly two options: a USB microphone or very good XLR microphones via audio interface with Phantom 48V XLR inputs. USB microphones include the preamplification and analogue to digital conversion thus eliminating a need for the audio interface and extra hassle. However, for best possible quality, you should consider a good audio interface, two good XLR microphones on an appropriate microphone stand (table or floor, with double microphone adapter). The audio interface can usually be any model, USB or FireWire or even a mixer with USB audio interface functionality, as long as it has Phantom, good quality preamplifiers and analogue to digital conversion, low latency, audio level meters and gain controls. The meters and gain controls will help a lot at calibrating the correct recording level.

Note that if you don't have echo cancellation, you cannot use speakers and have to use headphones instead. Headphones can be awkward for a musician, but are cheap com-

[55] DV: http://en.wikipedia.org/wiki/DV and HDV: http://en.wikipedia.org/wiki/HDV

[56] Microphones, audio interfaces, video hardware etc.: http://tinyurl.com/virmusic5

pared to speakers and eliminate the need for echo cancellation. In case of headphones, all participants in the room have to use them; otherwise they won't hear the far end.

1.5 Network requirements

Network requirements depend on the video conference or streaming method used. Simple Skype video calls at low quality are possible at almost any home level connection, even with less than 1Mbit up and down and with considerable packet loss. The most demanding applications, such as LOLA, may need over 100Mbps of bandwidth and as direct fibre connections as possible, no packet loss and minimal network jitter. Look at chapter '4.2 Network tools' for means to measure your network performance. It is difficult to say what connection is good enough since the network reliability also changes by the time of day and from day to day. Therefore all new connections have to be extensively tested before any actual session such as music teaching takes place. It is always a good idea to test with the final equipment with as final circumstances as possible. If there is no player or singer present for testing, it'll be good to simulate one with a cd-player or such. There are many pitfalls such as unexpectedly loud instrument sound. In far too many situations, the session is delayed because the technicians are still fixing quality problems as the teaching should have begun already.

Domestic level DSL connections are not especially recommended for video conferencing since they may introduce greater packet loss (especially on upstream) and unstable bandwidth not matching the specified maximum. However, successful lessons have been made on such connections (for example 24/2Mbit ADSL, ConferenceXP, Skype or H.323), so it may not be impossible either and is worth testing. Dedicated light paths or dedicated paths containing least amount of non-fiber connection will be of the highest grade. For non-research or small organizations a dedicated line may be more challenging due to the price or difficult availability. If the path is not dedicated and if there is something extra on the way such as tunneling, more or less unexpected problems may occur, such as sudden disconnection and sudden packet loss problems.

1.5.1 Firewall and private networks

To understand the basics, one must be familiar with the basic terminology. Terminals (endpoints), Multipoint Control Units (MCUs), Gateways and Gatekeepers are explained in the Wikipedia[57]. In IP (Internet protocol) based video communication, a critical question to answer is how to connect two terminals or IP addresses together without obstacles. On this chapter, we'll concentrate on how to connect two H.323 terminals or software such as ConferenceXP together.

With H.323 terminals, if both are able to connect to Internet, two questions remain: 1) If there is a firewall, is it open enough to let all needed ports and other functionality work properly? 2) Is at least one of the terminals connected directly to a public IP? If so, a

[57] http://en.wikipedia.org/wiki/H.323#Multipoint_Control_Units

successful connection can be made. All terminals will be able to show the IP they're getting. If the IP is inside private network address pool, such as 192.168.x.x, then the other terminal must have a public IP in order to establish the connection. Only a public IP can be called unless both are in the same private network (same room, for example). With software such as ConferenceXP, both IPs need to be public (unless both are in the same private network).

H.323 uses a few "fixed" ports and few dynamically allocated ports[58]. In some systems, they can be customized[59]. The dynamically negotiated ports are tricky for the firewall because it may be difficult to anticipate what ports are needed. If there are heavy restrictions in firewall that can't be switched off due to reasons like company policy, the easiest way may be to purchase a tunneling solution. Video conference companies provide firewall traversal as services. They may make things easier or in some cases that type of solution may be necessary in order not to compromise the network security. This kind of service often comes with dial plans, security, protocol translation etc. However it costs.

Simple IP based dialing is adequate for most distance music teaching. This may be achievable with correct network settings. With home level DSL routers, if you want that your terminal can be called to, you need to use the bridge mode and turn off the NAT (Network Address Translation). That way your terminal connected to the router can get a public IP. If you're stuck with the private IP, things get more complicated, but you may still be able to use the NAT feature in the endpoint to overcome the problem. In that case, you have to know the public address is and put that into the endpoint. With DSL, that address is often not fixed and my change every few days or weeks. You can also try using DMZ and port forwarding. More information about that is available at PortForward.com[60].

Some endpoints or software programs have ways for firewall and NAT traversal. Examples of such methods are H.245 tunneling[61] and H.460[62] traversal. The authorized network person should always be at hand when setting up the system. In order minimize network problems it may be a good idea to start with no firewall at all. One recommendable thing is to have UPnP (Universal Plug and Play) turned on[63]. Any small blocking elements may cause strange problems to the H.323 connection. Skype and other basic programs have better chances at working properly even with some firewall blocking.

---

[58] IP Ports and Protocols used by H.323 Devices (c21video.com): http://c21video.com/firewall.html

[59] Administrator's Guide for Polycom HDX Systems (version 3.0 page 2–23 or PDF page 51), description of Port settings:
http://support.polycom.com/global/documents/support/setup_maintenance/products/video/hdx_ag.pdf

[60] PortForward.com: http://portforward.com/

[61] H.245: http://en.wikipedia.org/wiki/H.245

[62] H.460: http://en.wikipedia.org/wiki/H.460

[63] In one of our tests at certain location this was mandatory, otherwise there was only audio but no video.

1.5.2 Latency and playing together

The question of maximum latency for local and remote ends playing together is a subjective one. In many situations chamber music or duo playing has been successful over the Internet, however already a relatively small delay can be destructive for the music. A small delay will start to feel different, strange and perhaps difficult compared to local situation. A large delay will destroy the possibility to react to musical cues and eventually make any rhythmical exactness impossible.

The following is a subjective view of latencies per effect. Opinions from different sources have been taken into consideration while making this table. There is a thesis study by Schuett[64] (2002), where the subject is tackled within ensemble performance scenarios. According to Schuett, 20-30ms is the maximum latency for playing together.

5ms: Quite impossible to notice any delay or difference between audible, visual or tactile cues. Playing together will not be hindered because of delay.

15ms: Audible, visual or tactile delay can already be perceived and may cause some, usually only slight discomfort.

25ms: Delay is clear for a musician. However, playing together may still be fully successful, tempo will not get confused and musical approach will not be hindered too much for meaningful rehearsing or performance.

35–45ms: Playing of simple, evenly rhythmical music is still possible. Delay (audio and image) can be slightly confusing and requires some getting used to. Also more complex music is still possible, but also the audience may note the slight delay.

45–100ms: Not recommendable for chamber music. With small children playing together some simple pieces may still be meaningful. Less rhythmically exact music can be tried, but the delay will cause more or less confusion.

Over 100ms: Clear delay, better not to even try playing together rhythmically critical music such as classical music.

Note that the speed of sound through air is approximately 340m/s, which means that 10 meters distance through air causes a delay of 29ms so one meter distance equals to ~3ms delay. It is better for the sound to arrive to the ear after the visual cue (image) because the brain is used to receiving visual cue before sound but not the other way around.

The limits for noticing bad synchronization between video and audio is subjective, but usually around 20–40ms difference is already quite distracting especially if sound arrives before video or when the synchronization is changing (delay is varying).

[64] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.7795&rep=rep1&type=pdf

1.6 Room layout

The more you plan to organize teaching and the more on-site audience you will have present, the better your distance teaching room planning should be. The room layout makes a lot of difference. A mobile video conference unit can, in theory, be carried to any space with a network connection, but surprising problems may occur in new circumstances. Therefore a room dedicated for distance teaching only is recommended. Mind that different instruments have very different characteristics and may well require very different technical setups. Some of the essential elements in room design are:

- room dimensions (defines overall spaciousness and basic acoustics)
- surface and wall materials (further defines the acoustic characteristics)
- window placements (sunlight is very bright and different to lamp light), curtains
- lights: no dark shadows should appear on face or background and image should look very detailed on the display
- door placement (audience and technician should be able to exit without disturbing the class or without being seen in the display)
- placement of acoustic damping and diffusion elements[65]
- placement of displays, cameras, microphone stands, speakers, equipment racks, chairs, cables (cable canals), mirrors (not compulsory at all but in some cases can be used to allow different simultaneous views)
- air conditioning and other sources of unwanted noise
- appearance (simple and beautiful will look good also on screen, background should be calm)

Polycom has written their recommendations about room design and layout on their Integrator's Reference Manual[66]. That is written from the speech conference perspective, but can give some basic ideas. Unfortunately we in Vi r Music don't have an example of a perfect room for distance music teaching, but some things can still be said. Depending on microphone type, generally the acoustics[67] should be somewhat dry and free of 'ugliness' caused by standing waves. Typically, what sounds bad to the other end, is unclear speech caused by too much reverberation and 'small room booming sound' caused by standing waves.

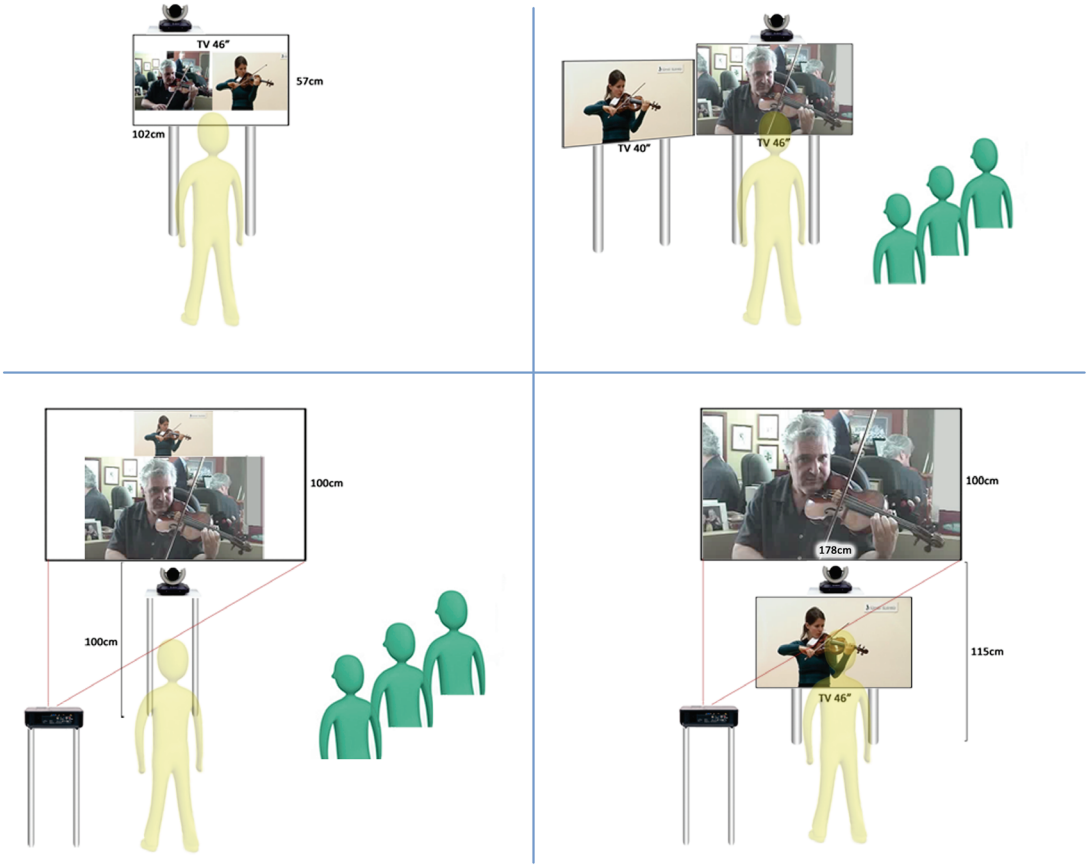
[65] Chapter '1.8 Audio optimization and acoustics' lets you know the basics of acoustic treatment

[66] Integrator's Reference Manual by Polycom, chapter Room Design and Layout: http://supportdocs.polycom.com/PolycomService/support/global/documents/support/setup_maintenance/products/video/hdx_irm.pdf

[67] Room acoustics: http://en.wikipedia.org/wiki/Room_acoustics

The background behind the performer, as seen on the display, should be calm and thus no audience or other moving objects should be placed there. A simple light or dark, soft toned wavy curtain is good. Behind that also any diffusers can be placed, hidden by the curtain. If there are lots of light surfaces in the room and lights are not spot lights, fewer shadows can be anticipated because of the high amount of light reflected around the room. Any coloring, such as room filled with brown wooden surfaces or red curtains, will change the tone captured by the camera. This can be corrected to only certain point by the camera's color settings. The tone is of course also heavily affected by the color properties of the lights used.

Here are some examples of how the displays can be used and how the on-site audience can be placed:



The yellow figure is the performer. In the lower images a video projector is shown, but that is not recommended for fixed setups. Instead, for traveling, it sometimes may be useful since it's easy to carry it while a big display is not. Check the Handbook Online Extension for hardware recommendations. For the camera placement it's important to place it as near as the far-end person's eyes as possible to ensure more natural eye contact. However, without projection to 45 angle surface and camera behind it, it's unfortunately not currently possible to achieve real eye contact. Viewing self-image is optional and is used for three main purposes: 1) Correct controlling of own camera, 2) Seeing yourself helps to understand your posture and compare it to the teacher, 3) If self-image is side-by-side on one display, that image is easy to record for later viewing (whereas if

local and far end are recorded separately, more work is needed in order to combine the images together).

1.7 Lighting

Three-point lighting is standard method used in film. It is described well in Wikipedia[68] and other web pages[69]. It involves the use of key light, fill light and back light. For video conference, use three-point lighting or any means that provide a good result in the screen. The camera captures the image differently than the eye and will have problems with great intensity variation within image. If there is direct sunlight in the background, any dark details in a shirt for example will be lost. Luckily, video conference cameras have quite good automatic calibration, so with luck, an average room light may even work. However, it may well be very bad as well. Test it under all conditions: windows open, closed, day time, night time and with different kinds of lamps. Lamp flickering, lamp heat and lamp noise are some of the potential causes of problems.

White background picks up shadows especially well and that is not wanted. In an optimal case the person has no ugly shadows or too bright or too dark areas, the clothes are visible and detail is seen on them too. Especially the face, the fingers and torso should look detailed and natural without blocking shadows. It is advisable to put something calm and beautiful in the background such as a soft toned curtain. Both light and dark backgrounds work, just mind that too much intensity variation will be challenging for the camera.

If a professional lighting person is not available, start simple and add lights as necessary until the performer's face and body is completely visible and looks detailed and sharp on the screen, without masking shadows or too bright spots.

1.8 Audio optimization and acoustics

Largely the same principles for microphone recording techniques apply to instrument distance teaching as in studio or live recording. If the microphone is too close or two far away, the sound will not be optimal. Not only the distance but exact location and directional positioning play an important role. When recording in stereo, the distance and angle between microphones makes a great difference. With many microphones, the movement of the player can affect the sound a lot. Sound and frequencies may radically diminish if the player or singer moves or turns away from the microphones.

Cameras have problems if within image there are great brightness variations such as sunlight on background and therefore a much darker face in front. In same way, microphones have problems if there are huge sound level variations within a session. Depend-

[68] Three-point lighting: http://en.wikipedia.org/wiki/Three-point_lighting

[69] 3dRender.com provides a good demonstration (three-point lighting): http://www.3drender.com/light/3point.html

ing on microphone type, distance and angle to the microphone can drastically change the sound level. When the sound level is way too much, microphone will pick unusable, distorted noise. When the direct sound level is way too little, only background noise or unfocused reverberation is picked. For audio hardware tips, please check '1.3.3 Audio hardware for standalone video conference'.

Since different instruments and singing may have a loudness variation of tens of decibels, it is usually not adequate to have a fixed sound level setting and assume everything will automatically be alright regardless of what instrument is played, who is talking or singing. Your equipment may be capable of some automatic gain adjusting for different situations, but then again it is often completely turned off since it is usually not that intelligent and may completely ruin the dynamics of the sound. Therefore audio level should be manually set according to each situation. If you only plan to play violin and talk, standing in one spot, it'll be adequate to have a one fixed level setting and one fixed microphone setup. All participants in a multipoint conference have to carefully adjust their levels to appropriate since others probably have automatic gain control also off, meaning participant's levels may drastically differ from each other if the levels are not calibrated.

Acoustics make a great difference in many ways. There are a few videos that will help understanding basic room treatment (watch the videos at 720p high quality mode):

'Hearing is Believing' (diffusion and damping in a small room as presented by Doug Ferrara from RealTraps):
http://www.youtube.com/watch?v=dB8H0HFMylo

'All about diffusion' (as described by Ethan Winer from RealTraps):
http://www.realtraps.com/video_diffusors.htm

'The Ultimate Home Studio'[70] (Ethan Winer and Doug Ferrara from RealTraps):
http://www.youtube.com/watch?v=ZSX14geMw-c

If the room acoustics are not optimized, for reasons such as using a mobile system at random rooms, there are two essential stages where the audio can be substantially further improved after all endpoint settings are already optimal and microphone type and placement is chosen wisely: 1) Speaker equalization or digital room correction[71] and 2) Microphone equalization. Find suggestions for hardware for this in the Handbook Online Extension. Microphone equalization is less important and doesn't probably doesn't need an external device if the room characteristics are proper or if the endpoint itself has good quality equalization features.

It'll be good to remember that all equipment with fans will produce unwanted noise. Mind the fan noise and consider some acoustic protection for the endpoint, any computers and hard drives. However, total silence is probably not necessary so small fan noise may be tolerable.

[70] A distance music teaching room should have least amount of modal ringing problems and no problems in the acoustic frequency response but generally perhaps slightly more beautiful reverberation compared to a mixing studio room

[71] Digital room correction: http://en.wikipedia.org/wiki/Digital_room_correction

1.9 Technical personnel

Many of the previously mentioned issues require wide-ranged professional skills. The goal is to create a fixed, stable studio where distance teaching is possible by simply turning on a main switch and dialing the other party or answering a call from them. However, in practice, at least one technical engineer is needed for certain tasks. Those tasks naturally depend on setup type and certain setups are inoperable without a real-time audio engineer for example. However, in a typical setup, some of the main jobs for an engineer would be:

1) Real-time zooming and panning of camera (local and also remote camera, if remote location doesn't do their camera control).
2) Audio levels: if the remote party sends exceptionally loud or too quietly, listening level might have to be varied in real-time to accommodate the situation.
3) Fixing unpredicted problems like lowering audio input level if a new instrument is much louder than previous instruments causing the audio level to hit the maximum and sound getting distorted. Or if wrong button is pressed and critical problems occur, there is an unexpected problem with network or other situation where only the engineer might understand the cause of the problem.

Much can be done by the student or the teacher by themselves as well, but to maintain the best technical circumstances, an active engineer who accepts only the best quality is needed. It is very typical that one of the peripherals or pieces of equipment malfunctions sooner or later, certain parts easier than others. But the biggest, most challenging and most important task for the engineer is to plan and build the studio. Then it is important to follow the development of technology and to update and maintain the studio.

If the studio is very well executed and there are no changing variables: the other party also has a stable, familiar and unchanged setup, no new instrument types are introduced, network quality is extremely high and nothing new to previous, such as unexpectedly needing to view music sheets remotely, is needed, then an engineer is not needed during the lesson (assuming that the student or teacher can operate the basic functions such as camera control and dialing by themselves.)

Some of the most important skills an engineer for distance education technology can have are:

1) Microphone recording techniques (especially microphone types)
2) Wide knowledge on audiovisual standards (audio and video technology)
3) “Golden ears” (including wide range of spatial sound understanding)
4) Understanding of visual composition and video quality
5) Knowledge on IP (Internet protocol) related issues and global networks
6) Personal experience on wide audiovisual work and knowledge on what others have done and are doing globally in network performing arts
7) Experience on video conference units and streaming technology
8) If the classes are recorded: Experience on video post-processing and publishing

9) And above all: The desire to understand and to improve audiovisual quality wherever it is possible

Typically commercial video conferencing companies do business meetings; distance music teaching is not their main field of business. Therefore typically one cannot simply order the construction of a distance music teaching facility from a regular video conferencing company. Much wider knowledge is needed. That makes the finding of a correct type of technical engineer perhaps the most important, and perhaps in some areas the most challenging part of starting a distance music teaching facility. In a few years the level of know how will get better and systems will incorporate more automation such as room acoustics and spatial audio calibration, but at the moment much has to be done through trial-and-error and a lot has to be customized with great precision and detail in order to achieve natural sound and image.

1.10 International standard

In 2010, there was no such thing as a mutual standard which all organizations of instrument distance teaching would use. The most common standards are presented in this Handbook, but most of them are unfortunately incompatible with each other. If music teaching studios are using different standards, it considerably reduces the connectivity. In other words, a studio could connect to many more other studios than today if there were a reduced number of standards in use. An option is of course to implement several setups in one studio. But in order to wisely develop the global distance music teaching network, it will be a good idea to team up and plan the standards together.

TERENA's[72] mailing list[73], consisting of people working in the fields of art and networking, is possibly one of the best communities where to discuss the connectivity. Hopefully there will be discussion and as result, music schools and independent studios as well as ordinary schools and other organizations can purchase high quality equipment, set it up properly, have good technical and social connections and the hardware doesn't go obsolete too quickly.

1.11 Summary: Requirements for distance music teaching

This section is for quick reference answering the question "What are the most important things needed to start a proper distance teaching facility?". We try to combine all the previous information in this Handbook into simplified sets, sketching some basic information needed in starting a distance music teaching studio or facility. Please note that much is left out for the sake of simplicity. Prices are also of course changing and not all cables and certain expenses specific to each location are included in these examples. Please read through the whole Handbook in order to get more a accurate view. Note that

[72] TERENA, Trans-European Research and Education Networking Association: http://www.terena.org/

[73] TERENA Network Arts mailing list, subscribe at http://www.terena.org/mailinglists.php (the email address of the list is network-arts@terena.org)

in addition to the equipment, budget consideration has to be done also on studio acoustics, furniture, engineer fees, maintenance, upgrading and so on.

Recommendable basic specifications in 2010:

- Video features: Resolution 1080p (or at least 720p) at 25 frames per second at minimum (50+ fps preferred), two screen support (useful but not compulsory)

- Multipoint bridge functionality if you need it. It costs more and complicates things as more displays may be needed. You might also be able to add multipoint functionality later. Note that only one site in a group of sites needs to have the bridge: others can connect to the bridge in any case.

- Audio cutoff frequency[74]: 16kHz or more

- Audio codec bitrate: 128kbit or more (for stereo), 256kbit+ or lossless codec preferred)

- Audio processing: Disabling AGC (auto gain) and Noise Fill should be possible and feedback echo cancellation should be as transparent as possible

- Microphones: Broadcast grade studio or live recording microphones (with a high quality preamplifier if needed)

- Compatibility: Maximum quality with other participants should be ensured

- Room: Carefully planned room solely for distance teaching purposes, acoustically quite dry but good sounding, practical furniture and equipment placement (mobile units are also possible but bring new challenges)

- Technical supervisor: A person or persons who are devoted and capable of supervising and developing all aspects of the technology (microphones, speakers, video conferencing units, computers, network, video production, lights, usability and reliability of the system)

Current popular 1080p/30fps H.323 systems more or less compatible with the above technical requirements, including the multipoint bridge functionality (720p/60fps except where stated) and the compulsory support package for one year:

– Polycom HDX 8000 (estimate 14800 €[75] with a microphone preamplifier)

– Tandberg C60 (estimate 24300 €)

[74] In the data sheets cutoff frequency is also known as frequency bandwidth. However, cutoff as a term means exactly where the highest frequency is (to be exact, the cutoff is steep but gradual and not sudden in the frequency response) and doesn't say what the lowest frequency is. Frequency bandwidth means exactly the band from lowest to highest frequency, and if the lowest is not specifically reported, it is assumed to be usually around 20Hz.

[75] Prices in this section are without taxes.

– Tandberg C40, multipoint bridge 576p only with 1 display support only (estimate 17000 €)

– Tandberg C40, multipoint bridge 576p only (estimate 18200 €)

Note that Tandberg AAC-LD and Polycom Siren 22 audio codecs are not compatible and Tandberg–Polycom connection will revert to 7kHz audio (cutoff frequency). For other competitive H.323 systems and other solutions, please check the Handbook Online Extension. Next we will list examples of typical setups. Note that H.323, ConferenceXP and Skype are not compatible with each other. Also always make sure the network is reliable and fast enough for the purpose.

1.11.1 Example setup 1: Fixed H.323 studio

A fixed studio is the most reliable, best performing solution. The studio or space can be small or big, as long as it is suitable for distance music teaching. Since there are pros and cons for both Polycom and Tandberg, we have to list both options. Perhaps in a great simplification, two questions stand at the top when choosing between the two: 1) Which system are the other participants in the network already using? 2) Do you need pixel-accurate image screen layouts customization for multipoint (only Tandberg is capable of that)?

Polycom HDX 8000 with 1080p/30fps or 720/60fps capability, multipoint bridge 720p/60fps and 6Mbps line rate = 14500 € + RME QuadMic Phantom 48V (preamplifier for the microphones) = 300 €. Total: 14800 €.

OR

Tandberg C40 with 1080p/30fps or 720/60fps capability, multipoint bridge 576p/30fps, secondary video output enabled, has Phantom 48V microphone inputs = 18200 €.

AND

2 x Neumann TLM-103 cardioid pattern condenser microphones = 1560 €.
1 x Microphone stand with stereo adapter = 50 €.
2 x Genelec 8030A loudspeakers = 800 €.
2 x Stands for the loudspeakers = 100 €.
1 x Samsung 460MX-2 46” LCD low latency display = 1500 €.
1 x Stand for the H.323 terminal and display = 500 €.
1 x DBX DriveRack PX Auto-EQ for speaker calibration = 380 €.
1 x Pinnacle Video Transfer[76] + 1TB USB2.0 ‘silent’ hard drive = 200 €.
Miscellaneous cables = 200 €.

Total: 20090 € with Polycom or 23490 € with Tandberg

[76] For very basic video recording (for HD recording, please look a the Handbook Online Extension)

Remember the additional expenses such as engineer fees, acoustic design, lighting, furniture, music stands and upgrading.

1.11.2 Example setup 2: Mobile H.323 unit with wheels

This is similar to fixed setup except the microphones are better to replace with shotgun microphones. That will make it easier to catch lesser part of the unwanted random noise from audience, air conditioning, room reverberation and so on. Shotgun makes the situation easier for the echo cancellation processor as well. The idea is that the mobile unit can be easily transferred inside a school to any room with an appropriate Internet connection available. The mobile setup, compared to the fixed one, is as follows:

– 2 x Neumann TLM-103 cardioid pattern condenser microphones = 1560 €.
– 2 x Stands for the loudspeakers = 100 €.
+ 2 x Røde NTG-2 shotgun microphones = 320 €.
+ 1 x For use with Polycom only: Behringer Ultracurve Pro DEQ2496[77] = 200 €.
+ Extra structure and fastening for the mobile rack = 300 €.

Total: 19250 € with Polycom or 22450 € with Tandberg

1.11.3 Example setup 3: Computer solution (relatively cheaper)

This setup is recommended as a cheaper alternative to replace the H.323 system. It may often be a very good idea to add a secondary system besides (or integrated into) the H.323 system. That will enable compatibility with a greater number of partners. A basic setup for software like ConferenceXP (uncompressed high quality sound and video), Soundjack (very low latency audio conference using CELT audio codec) and Skype (easy conference though with distorted and much less dynamic audio):

1 x The fastest/quietest PC currently on the market (laptop or desktop) = 1300 €.
1 x RME Babyface USB audio interface = 440 €.
2 x Neumann TLM-103 cardioid pattern condenser microphones = 1560 €.
1 x Microphone stand with stereo adapter = 50 €.
1 x Samsung 460MX-2 46” LCD low latency display (or use a shared display) = 1500 €.
1 x Sennheiser HD 600 headphones = 240 €.
1 x Logitech HD Pro Webcam C910 = 80 €.
Miscellaneous cables = 120 €.

Total: 5050 €

[77] This is a EQ processor meant to put between preamplifier and the Polycom terminal, and the idea is to remove low frequencies or any other problematic frequencies. Tandberg C series has a digital parametric EQ so that functionality eliminates the need for the external EQ.

This system allows very high quality audio (ConferenceXP, Soundjack). But headphones are not convenient for classical music. The musician should hear the sound of their own instrument without obstacles. In many situations, loudspeakers would be better. But then we have to add in a high quality echo cancellation system to be used with software like ConferenceXP and Soundjack, which don't have any echo cancellation built in. Since, as far as we know, there are no high fidelity external echo cancellers (suitable for live music without quality loss) other than EchoDamp. We must use it, although it does add lot of complexity to the system. External echo cancelling will surely get simpler in the following years, but currently we have to do it like this.

We may have to sacrifice the high quality of Neumann TLM-103 microphones to make echo cancellation easier for EchoDamp. Therefore we switch to shotgun microphones (they pick less room sound). To be on the safe side, you would use a separate computer for EchoDamp. However, the following setup has only one computer doing both EchoDamp[78] and video conferencing. The system with loudspeakers would be as follows[79]:

1 x The fastest/quietest PC currently on the market (premium level laptop) = 1900 €.
1 x Samsung 460MX-2 46" LCD low latency display (or use a shared display) = 1500 €.
2 x Genelec 8030A loudspeakers = 800 €.
1 x Logitech HD Pro Webcam C910 = 80 €.
1 x Roland Cakewalk UA-25EX USB audio interface (for 'codec') = 160 €.
1 x RME Fireface 400 FireWire audio interface (for EchoDamp) = 650 €
2 x Røde NTG-2 shotgun microphones = 320 €.
1 x Microphone stand with stereo adapter = 50 €.
Miscellaneous cables = 120 €.

Total: 5580 €

Cons compared to H.323 system: Video quality is not as good, no camera control, EchoDamp is not easy to calibrate for others than audio engineers (in principle it can be set once and then not touch it, but that is not a completely safe solution).

1.11.4 Example setup 4: Minimal computer setup

With this setup you can use ConferenceXP with headphones and Skype with the loudspeaker. Of course almost any PC laptop nowadays is capable of running ConferenceXP or Skype successfully, but the default microphone and speaker quality is usually very bad. This setup gives relatively good sound quality when using ConferenceXP uncompressed audio and headphones. Skype sound is quite distorted but makes things easy because of its effective echo cancellation.

[78] EchoDamp does moderately utilize CPU, but for a very fast new computer, the CPU usage won't be significant. An Intel Core2Duo computer bought in 2008 runs EchoDamp at 35% CPU usage.

[79] Note: This system contains some relatively new hardware and is not yet tested in this particular setup. There is a risk of unexpected compatibility problem. This text will be removed as soon as we have a chance to test this exact system and confirm there were no compatibility problems.

1 x A fast and quiet PC laptop with high quality 18.4” display and webcam = 1100 €.
1 x Shure PG42-USB Microphone (mono), with shock mount = 230 €.
1 x Simple table microphone stand = 20 €.
1 x Genelec 8030A loudspeaker (naturally mono when there is only one) = 400 €.
1 x Sennheiser HD 600 headphones = 240 €.
Miscellaneous cables = 30 €.

Total: 2020 €

If you build systems based on these recommendations, please email virmusic.blog@gmail.com and let us know of your experience.

2 Video recordings

2.1 The advantage of video recordings

Master classes and pedagogical demonstration videos can be greatly beneficial for students, their parents and others who are interested in the topic. If the subject taught or examined on the video comes across clearly enough, then it can be said that the video has successfully executed its purpose. However, to get the best result, there are many issues and steps to carefully plan and carry out in video recording, post processing and publishing.

Recording step: The requirements are similar to commercial film production – all recording equipment and studio preparation have to be of best quality, including audio, video, camera operation, light, acoustics and so on. Distance master classes can be recorded out of the video conferencing equipment VCR output or similar. The recording computer has to be fast enough to record the highest quality (FullHD) without any dropouts. Everything should go as perfectly as possible at the recording stage, because many kinds of problems may be daunting to fix in the post processing. Problems in audio and video synchronization are an example of that. It may be a good idea to already index as much as possible while recording. That can be done for example simply with pen and paper, writing down what happens at what time stamp and what a particular section is dealing with.

Post processing step: The recorded videos are transferred to video editing station. Video and audio quality is fine tuned, videos are edited and unnecessary bits are left out. If videos or audios were recorded by several devices simultaneously, they are now connected together. When needed, the previewing process is one of the things requiring time: to find the best bits a lot of material may have to be watched. When everything is chosen and edited, it is time to publish the videos.

Publishing step: Sharing the videos on the Internet is obviously good for the music education community. What makes it even more accessible is if the videos are indexed, categorized and described nicely. For example in YouTube there are many videos about violin playing, but videos about certain bowing techniques may not be easily found because either the videos aren't there, the essential index words are not set or the interesting part is a part of a long video and there is no way to directly find the specific section.

2.2 Video file formats for storage and Internet

For storage, it is usually best to save the original, often huge video file. It will be slow to transfer to Internet or even to backup hard drive, but original quality will be preserved and on the other hand encoding (compressing) the original long/huge file may take vast amounts of time as well.

On the Internet, for many years the video players have been mostly using Adobe Flash technology, lead by widely used services like YouTube. However, in 2010, the trend is changing due to instances like Apple, who are putting Flash down and bringing up new technologies like HTML5 instead. Since HTML5 is relatively new, we shall focus here on Flash technology, though it is a good idea to consider the HTML5 players[80] due to future and current compatibility. Two most common Flash video players with a free version available are JW Player and Flowplayer. They are regularly updated and have additional features like custom subtitles etc. Also Adobe Flash itself is regularly updated and that means the video codecs are improving year by year.

http://en.wikipedia.org/wiki/Flash_Video

In 2010, the two popular video codecs inside the Flash video containers FLV[81], F4V[82] and SWF are the On2 VP6 and H.264[83] (also known as MPEG-4 AVC) video codecs. The Sorenson Spark[84] (also known as Sorenson H.263 and FLV1) codec may be used when little CPU usage for decoding is needed, however the quality is considered lower. High resolution videos can be problematic when GPU acceleration is not enabled. Playback may stutter, especially at full screen mode, even on fairly fast computers.

Some of the most important features in a video file for Internet are the following:

1. Good quality (it is assumed that the original untouched video file is of professional quality):

---

[80] Some of the promising HTML5 player technologies: http://www.net-kit.com/20-html5-video-players/

[81] FLV specifications: http://en.wikipedia.org/wiki/Flv

[82] F4V container for encoding/muxing is not supported by FFmpeg (only FLV is)

[83] H.264 and On2 VP6 short introductions: http://help.adobe.com/en_US/AdobeMediaEncoder/4.0/WSF866FB02-31F2-4bab-99F3-E4D8653759D1.html

[84] Avidemux version 2.5.4 only encodes Sorenson Spark alias FLV1 for the FLV container.

1a. Resolution (high enough resolution so that important details are not lost)

1b. Frame rate (to keep quality, it may be better to avoid lowering frame rate from original, however unnecessarily high frame rate may waste bandwidth and CPU usage considerably)

1c. Video codec (as an example, H.264 has high quality video but CPU usage is high)

1d. Audio codec (mp3 or aac or other recommended codecs are usually good above 192kbps)

2. As small size as possible (so that real-time playback is possible through Internet)

3. As little CPU usage for decoding as possible (otherwise may stutter on slow computers)

4. Streaming ready format (only very specific file structures can be streamed so that timeline seek is possible immediately without downloading the whole file)

5. Compatibility (codecs should be as open and compatible as possible so that files won't have to be re-encoded when switching to another platform, also the same files should be playable on standalone common integrated operating system players or living room video capable machines if possible)

6. Quick encoding (otherwise long videos may take days to encode)

2.3 Basic video editing and post-processing

Some of the very basic tools[85] and programs for video editing could be introduced as follows:

The two very important free video software programs are Avidemux and FFmpeg. Avidemux can decode and encode various formats, includes a timeline and basic post processing. Avidemux is prone to crashing[86], but at best does a good job. There are various ways to do also more complex editing[87] on free software. For commercial products, Avid Media Composer and Adobe Premiere Pro are some of the usual programs. When using those, it is possible to have certain video cards do both real-time viewing/processing and offline encoding/processing a lot faster. Avidemux may be useful for certain basic tasks even if larger programs will do other things better.

[85] The links to these tools are listed here: http://tinyurl.com/virmusic4

[86] Unfortunately Avidemux 2.5.4 (and older versions) can crash in strange ways: doing something may crash, but doing the exact same thing again may suddenly work (so it may be worth trying again)

[87] For example you may want to do a video fade in/out. That can be done with Avisynth for example. See: http://forum.videohelp.com/threads/48579-How-to-edit-with-Avisynth#fade

With FFmpeg you can do a wide variety of things though it is a command line program (when no GUI or graphical user interface is used to control it). Here are some useful things to do with FFmpeg:

If you have a corrupted video file, you can try to retranscode it:
```
ffmpeg -t 00:0:05 -i "input file.avi" -vcodec copy -acodec copy "output file.avi"
```

H.264/aac/.mp4 will give you good quality at a given bitrate/filesize. You can encode to that, then re-wrap into .flv with FFmpeg (no video re-encode, in this example audio is converted to 44100kHz and 256kbit/s but choose not to convert when possible):
```
ffmpeg -i input.mp4 -ar 44100 -ab 256k -vcodec copy output.flv
```

To strip audio from a video file, to 44.1kHz 16bit stereo WAV:
```
ffmpeg -i input.flv -vn -acodec pcm_s16le -ar 44100 -ac 2 outputaudio.wav
```

To strip original audio:
```
ffmpeg -i input.flv -vn -acodec copy outputaudio.mp3 [or change mp3 to what it really is, find out that with MediaInfo for example]
```

```
More FFmpeg parameters: http://www.ffmpeg.org/ffmpeg-doc.html or http://howto-pages.org/ffmpeg/
```

MediaInfo is a good program to identify the containers and codecs and other metadata information of a video or audio file.

If and when the audio is not perfect as is, it may be important to edit it with an audio editor. A simple free program for that is Audacity but that may not be adequate for certain tasks. For more programs you can use Google with for example the search terms DAW (digital audio workstation), audio editor or sequencer.

When making a basic .mp4 video file from a digital video recording[88], the process with Avidemux might be something like this:

1) Open your .avi (or some other) video to Avidemux.

2) Choose MPEG-4 AVC (= H.264) as the video codec, AAC as the audio codec and MP4 as the format (= container).

3) If you need to do edit your audio, you can do that at this stage for example. Choose from menu: Audio/Save and the audio track will be saved as is (if it is for example AAC, you may have to decode AAC to WAV with faad.exe for example). After the edited audio is ready, you can put it back to Avidemux at Audio/Main track.

4) Video/Configure: Choose your video bitrate (for example 2500kbit/s CBR).

5) Video/Filters: Add a proper interlacing filter if necessary.

6) Video/Filters: Add Transform/Crop (and adjust properly).

[88] Looking for hardware to record video conference calls with? Some basic recording hardware are listed here: http://tinyurl.com/virmusic5

7) Video/Filters: Add Transform/Resample fps if the real frame rate is lower than the frame rate on the original video, because unnecessary frames may still eat CPU usage and grow the file size.

8) Video/Filters: Add filters to fix colors and image quality when necessary.

9) Audio/Configure: It may be best to use the highest bitrate.

10) Audio/Filters: You may want to consider resampling in certain situations and adding Gain mode: Automatic if you didn't optimize the audio externally.

11) If you already set your starting and ending locators, you're ready to encode your .mp4 video (press ctrl+s to do so).

You can publish a .mp4 file on your website. If you do so, you probably want to have the "moov atom" metadata located at the beginning of file. Otherwise the video has to download completely before anything is played. A recommendable program to fix the metadata is MP4 FastStart[89]. However, that is replaced by another technique in the Flowplayer pseudostreaming solution explained next.

2.4 Seekable streaming

For example this video: http://www.sarestoacademy.org/demo-rudin2/ uses Flowplayer and a simple PHP technique to stream the file jumping to any file location even the page was just opened and not much of the file is yet downloaded. In other words, the video supports forward seeking without downloading the whole file first. Note that this solution requires Flash and therefore does not work on iPad for example.

The previously mentioned video uses Flowplayer as the player and Richard M. Bellamy's streamer.php[90] to pseudostream a video from a PHP enabled server. That method does not support streaming from an external site, so the storage site must support PHP. Another common Flash player would be the JW Player. To get started with Flowplayer forward seekable pseudostreaming with just simple PHP, take a look at the html source of the previous example. The detailed instructions are found at these sites:

http://flowplayer.org/plugins/streaming/pseudostreaming.html

http://richbellamy.com/wiki/Flash_Streaming_to_FlowPlayer_using_only_PHP

The recommended format for Internet is either FLV (On2 VP6 or H.264), F4V (H.264) or MP4 (H.264). Choosing between VP6 and H.264 is mainly a question of encoding software (and hardware) speed and playback CPU usage.

[89] MP4 FastStart: http://www.datagoround.com/lab/

[90] streamer.php: http://richbellamy.com/wiki/Flowplayer_streamer_php (you may want to disable the XMOOV_CONF_LIMIT_BANDWIDTH in order to stream fast enough)

FLV needs to have the metadata injected to the beginning of the file. Otherwise the pseudo-streamer will not do the forward seeking. FLV MetaData Injector[91] is a recommendable injector[92]. When using the GUI version of FLVMDI, you have to choose the option to "Include 'keyframes' object", otherwise it doesn't help. The "Inject onLastSecond event" can be enabled as well, but that didn't seem to make a difference. If the GUI version hangs when doing batch processing, use the command line version instead.

2.5 Audio and video synchronization on recordings

One of the most time consuming tasks is to fix the audio and video synchronization if it has failed on the original video feed. It will be especially slow to fix it if the delays are changing throughout the video. One notable aspect is that different players may produce slightly different synchronization between audio and video. That means you also have to make sure that the synchronization is correct on the final player you're going to view the videos with. Hopefully that is not a problem, but it's important to acknowledge the possibility if you want to achieve a technically flawless result.

2.6 Streaming for an audience – Live teaching or file streaming

At master classes, there is often an audience at site but many others could benefit of the teaching as well. Multipoint functionality is used to let passive parties connect and participate only by listening. However, as time zones may cause problems and multipoint will add general complexity for the technical setup, there is the alternative to either stream the session live on the Internet or record the session and later stream the edited recording.

The advantage of multipoint conference is that audience can also interact for example by asking questions. The advantage of file streaming is that the less interesting parts can be edited out and distribution becomes easier as viewers can view the video file at their own time, as opposed to many watching at once, creating more challenge to the server technology. However, video editing requires extra work and the feeling of a live event is lost. Pedagogically it is very beneficial to have well-edited videos carefully indexed on the Internet, as then the audience can very easily watch the videos at their own time and pick the ones that are most relevant to them.

For streaming, it's good to start with understanding of routing, specifically unicast and multicast. They are explained in Wikipedia. If unicast is used, huge network bandwidth is required for the server sending the stream as all viewers will need their share of the bandwidth. Some of the current streaming solutions are explained in chapter '1.4.2 High

[91] FLV MetaData Injector: http://www.buraks.com/flvmdi/

[92] Others you can try are flvtool2 and YAMDI, but flvtool2 crashed on a flv created by ffmpeg (didn't crash on a flv created by Adobe Premiere) and YAMDI completed successfully but the file didn't seem to have metadata correctly injected.

quality streaming solutions'. Some endpoints such as Tandberg Edge 95 may have streaming capability built-in. The video conferencing companies also usually offer streaming of a video call as a service.

2.7 Music theory and music history

For music theory or music history distance teaching, a means to send high quality audio streams from both microphone and a music player is required. Also either the video camera should be able to transmit very sharp images so that notation and other small details can be easily viewed on the screen, or this can be done online on a computer. Alternatively there can be a combination of both: a high resolution video camera feed plus online collaboration on a computer. This can be set up on both ends (or all ends in case of multipoint). There are several methods how to achieve this. One is to use H.323 video conferencing, having a laptop connected to the presentation input. When computer desktop sharing is used, one can effectively show pedagogical material at high resolution, for example showing notation in Sibelius notation software.

Another solution is to combine your favorite streaming, desktop sharing and communications software and use them simultaneously to create a convenient online classroom. The basic configuration can vary. For example it can be done as follows: teacher–students, where only the teacher has capability of sending high quality speaking voice and music (such as excerpts from classical music), or it can be teacher–student–student, where the students are each at their own locations and each participant can have the same equipment and capability of participating with same way and same quality. One relatively simple software solution is ConferenceXP, which includes all previously mentioned streaming, desktop sharing, communications and more. There are many ways to mix the speaking voice and playback music together. A traditional mixer can be used and the music can be played from an external CD player. However, the mixing can be done in software too. The music can be played on any media player and if the audio interface supports 'stereo mix', the microphone input can be mixed to the media player sound in a software mixer, while outgoing audio is the 'stereo mix'. Or JACK can be used in certain situations to mix sound between applications.

An alternative to ConferenceXP is to use for example the following: 1) TeamViewer for high resolution desktop sharing, multipoint basic webcam video and text chatting, 2) Soundjack for high quality low latency multipoint audio when using headphones or Skype/Mumble when using Speakers (echo cancellation). Or, in some situations, the teacher can use for example VLC to stream the audio. The VLC stream is easy to view on many media players, but very low latencies will probably not be achieved that way. Please look at the Handbook Online Extension for more alternatives.

3 Good to know for teacher, student and organizer

3.1 Camera usage

The engineers may control the movable camera at both ends or one engineer may control both local and far end camera. The engineer may also optimize sound levels and other settings. However, if there is no engineer available or if the teacher or student want to control some of the simple functions by themselves, it is certainly workable. The video conference unit's remote control have very simple buttons for sound level up or down, mute or camera control. The camera can be tilted in all four directions and zoomed. Camera presets are also not usually difficult to set or use. That works by the machine's own logic, but usually it involves pressing down a number key to set or get a specific position for the camera. Camera preset functionality may become very handy when there are certain often desired zoom positions such as close up on piano keys or close up on violinists left hand and back to the full view.

Below are some examples of recommendable positioning for violin and cello. It is important to always zoom as near as possible, still maintaining visibility to essential parts. As a generalization, the most common mistake is to zoom out too far and have unnecessary blank space above head or too much of legs or floor in the image.

3.2 What to expect

With best setups, you can hear and see very close as much in detail as if you sit on the same room, except on current systems the video will not be three dimensional and the sound will have the three dimensional radiation information lost as well. However, not all systems are optimized to the highest level. In those cases, there might be problems of several natures. Hopefully you don't have to experience these problems, but here is a list of some basic quality problems that may sometimes rise in a video conference:

Audio

- The total delay may cause slight discomfort (same with self-image delay)
- The timbre or sound tone may not be as detailed as in a traditional classroom situation
- The dynamics (distinction between piano and forte) may not be as great as in a natural situation
- The synchronization between sound and image may break
- Due to varying distance from microphone, the sound might get too distant or low (or too close and too powerful)
- There might be some dropouts in sound or image
- You might hear echo of yourself

Video

- The image may be zoomed or positioned in a non-optimal way
- Framerate will drop meaning the video will be jumpy and fast finger movement is lost making it impossible to tell which fingers were actually used
- Difficulties with music sheets (if tried to view it through the camera)
- Image may be too dark, too bright or too blurry to see properly

These are all problems which are avoided if the system is properly designed and maintained. However, if some of the problems do occur, all is still not lost. For example after getting used to, the communication delay will no longer feel as problematic. Lot can be still taught even the sound isn't as accurate as it should. However, the engineer should fix the problems if they are dependent on the settings. Where the system is already completely optimized to the last resort, the only option will be to upgrade the equipment or other fixed elements such as acoustics.

In the best case you can except to see as good image as a FullHD display can produce and to hear as good sound as on a very good classical recording, played back on very good speakers. As a teacher or as a student, you may need to simply walk in the studio and start the lesson as if the display was a window to the other room. To see one example of the distance teaching situation, you can watch the video below. In this video Alexander Rudin is teaching a student in Oulu, Finland and the set up is similar at the other end. Two versions are provided since the high quality version requires a very fast computer in order to see the video smoothly.

Low quality (720x576 30fps):
http://www.sarestoacademy.org/demo-rudin3-h264-720x576-2688kbps-30fps

High quality (1280x720 60fps):
http://www.sarestoacademy.org/demo-rudin3-h264-1280x720-4059kbps-60fps

3.3 Restrictions and tips

The network and the transcoding introduce latency, in other word delay. For short distances it can be minimized to unnoticeable in the best scenario. However, for long distances it is theoretically impossible to communicate without any delay. Therefore it is good to acknowledge what the delay means. Most notably it means waiting for a short moment (for example on fifth of a second) for the other party to respond. It may slightly make the conversation uncomfortable, but it will not make it impossible. Instead, tapping tempo or playing together will become impossible. These activities are possible only at short distances and setups optimized for very low latency. In theory one can play the accompaniment and the other one can play the solo, but as soon as the accompanist starts actually listens to the soloist, the timing will completely break. In other words there can exist no musical collaboration in real time when the delay exceeds a certain limit (around 45 milliseconds). However visually showing emotion or expression to the student may still be possible even the timing is slightly late.

Moving the student's arm or fingers, especially important with children, is not possible. Also moving around the student to see the back of the hand or other special angles is not possible by walking around. However, it is not that difficult to ask the student to turn around or to zoom the camera. It may be slow and the student might not immediately understand which way to turn to. But it is manageable and you should be able to see even minute details in the finger if the system supports high grade video features such as FullHD at 60 frames per second.

In a multipoint call, the microphones of participants listening only should always be switched off. Having them active will create unwanted noise for all participants and if a voice activated mode is used, the big image will also change according to the loudest sound source possibly switching to wrong sites. If voice activated video switching is used, it's good to know that it may take some seconds before the video is switched.

It's good to realize that it is challenging for the system to remove the feedback echo resulting from the microphones picking up what the other party just said. Because that echo is removed (unless echo removal is done completely by acoustic techniques or if there is no delay and the echo is turned in to reverberation effect), the sound quality may get lower when both parties are speaking or playing simultaneously. If that happens and the quality drop is disturbing, simply try to refrain from talking or playing at the same time with the other party. (This problem may not be noticeable at all with the best systems.)

There are a number of new opportunities the technology can add to a traditional lesson. The practicality and usefulness of them depend on things like how well the system has been built and the readiness of the student or teacher to use it. As a distance lesson already involves cameras, it is often not a big effort to also record the lesson. How to pub-

lish or view the video is another question, but some of the ways to view the video include: watching after the lesson from the same equipment, getting the video transferred to the Internet and watching the video on any web browser, watching the recording on a tape or disc at home or elsewhere. The video recording and archiving is pedagogically very useful and if the videos are published widely, a great number of students, teachers and others can benefit from them.

A video conference system offers an opportunity to see yourself on the screen. It differs from a mirror because the horizontal plane is not swapped as in a mirror. Self-image is needed for the one in your room who is going to correctly position your camera (unless it is assumed that the far end will control your camera). The self-image can be pedagogically very important, especially for the students who work with their posture and benefit from analyzing their own movement. When the system has the teacher on the left and student on the right, it is easy to compare the two postures. When posture is in question, it is good to tell the student to turn at the same angle as teacher. With violin for example, it is very useful and relatively easy to ask the student to point their violin neck exactly towards the camera or point their bow exactly to the camera, carefully positioning the violin at exactly 90 degrees so that the bridge appears as a thin line on the screen. This type of teaching allows for very accurate visual comprehension. If your self-image is slightly delayed, don't let it disturb you too much. It is a problem related to camera, codec and display latency and may not be possible to remove without upgrading the equipment.

For those who are already very fluent with computers, it is possible to use also some other functionality, basically just transferring things like metronomes or tuning meters to a computer. Sometimes it may be pedagogically interesting to share links to videos such as videos about some playing technique or great performance found on YouTube for example. For instruments like jazz guitar, the computer accompanying is popular. However, the use of computers or complicated systems is certainly not necessary in distance teaching. They should be just used where it is reasonably comfortable and when the systems are truly practical and good and don't hinder the highest level of traditional music making and pedagogy.

3.4 Music sheets

In a traditional local lesson teacher may check the fingerings or make markings on student's music sheet. This becomes challenging in distance teaching. It also may cause some additional trouble to try to pinpoint places in the score. However, things get much easier when both have quick access to copies of the same sheet at both locations.

It's often essential to prepare the following for distance lessons: Student should scan and send the scanned music sheet to teacher side where it is good to print on paper with clear quality so that markings do not vanish in the process. Instead of printing, there are other well-working ways of viewing the scanned sheet "offline". Most obvious methods are opening the scanned file, such as pdf, on Apple iPad or a laptop. If doing a lesson on a computer with a one widescreen display, student on one side and the music sheet on

other side works fine. The current version of Apple iPad is slightly small and the resolution (1024x768 pixels) is not too high for music sheet viewing purposes, but it is handy at turning pages and zooming and the image quality is high enough for practical use.

The sheets can be sent or video streamed online or over the video conference by several different methods. Aiming the camera at the sheet will often produce quite bad results. If the lighting conditions are not optimal, the result may be just a white page. Aiming the camera really carefully and waiting for the auto-focus and auto-brightness to work may result in readable music, but that method is often troublesome. Much time is wasted moving the camera around. The high quality video conference methods are having a document camera[93] or a laptop video output connected to the video conference endpoint's presentation video input. That will allow presenting the second video onto different positions in the video layout. Usually the presentation can be set within the same screen layout logic as multi point calls: Picture in Picture (PiP), Picture outside Picture (PoP), Side by Side or to the second (or third) display in full screen mode. It's also possible to have any external camera to point at the sheet and connect the camera to the endpoint by a relatively low quality S-Video connector for example.

The sheets could be also presented in an external online service, but in many cases that would be impractical and the extra hassle doesn't make it worth the trouble. However, the free Web conferencing software listed in http://tinyurl.com/virmusic0 will work fine at online collaboration and functions such online PDF presenting or whiteboard drawing.

3.5 A dedicated studio

The importance of stability of the system should not be overlooked. This means that the studio or room should be as dedicated to distance teaching only as possible. The same equipment should not be used for other purposes and parts of the equipment should not be moved to other location temporarily. The software settings should not be changed after they are optimized. Any changes will often cause problems; a cable is accidentally put back to wrong connector, a volume setting is forgetfully left at the maximum position and so on.

3.6 Web sites with music learning videos

http://www.pzvln.com/ (Pinchas Zukerman violin lessons online)

http://www.violinmasterclass.com/ (Sassmannshaus violin classes)

[93] Wikipedia will show what a document camera looks like: http://en.wikipedia.org/wiki/Document_camera#Desktop_models (the highest quality models support FullHD 1920x1080 pixels and HDMI and can cost somewhere around 200–2000€)

http://violinmasterclass.com/vm_live.php (the previous site has started offering live video conferencing lessons as well, using ooVoo or Skype)

http://www.youtube.com/results?search_type=search_playlists&search_query=violin+master+class&uni=3 (Example of YouTube violin master classes playlists)

http://www.pickstaiger.org/video-library (Davee video library music classes and performances)

http://www.rockway.fi/ (a Finnish site for learning rock instruments playing with features like video commenting, integrated accompaniment tracks, similar video suggestions)

University lectures and YouTube

http://oyc.yale.edu/ (Yale courses: Astronomy, Music, Philosophy etc.)

http://www.youtube.com/education?b=104&t=m&s=edu&lg=EN&cr=US&p=3 (YouTube EDU)

Examples of other live arts through video conferencing

http://www.youtube.com/watch?v=ijMGKDxtMyI (a dance rehearsal using a triple screen system, filmed at both ends)

http://www.youtube.com/watch?v=aWjb7Q42uUg (Bradley University doing a theater and multisite production called The Adding Machine)

New technology related to video conferencing

http://www.youtube.com/watch?v=EndNwMBEiVU (True 3D Display Using Laser Plasma Technology)

http://www.youtube.com/watch?v=L7kJv2aLnBk (DVE Telepresence Stage)

http://www.youtube.com/watch?v=jAIDXzv_fKA (DVE Immersion Room technology)

http://www.youtube.com/watch?v=pSICZ_7hpho (Pepper's ghost projection technology explained)

http://www.youtube.com/watch?v=9OvTLg4i2_U (rollable OLED screen)

http://www.youtube.com/watch?v=f8S8tbQMp2k (another bendable OLED screen)

http://www.youtube.com/watch?v=3seTlvQtIgc (touchable holograms)

http://www.cim.mcgill.ca/sre/projects/ (Shared Reality Lab, McGill Centre for Intelligent Machines)

4 Good to know for the engineer

4.1 How to test latency (transmission delay) in a video conference?

A digital stopwatch[94] with milliseconds is required. Point your local camera at the stopwatch, zoomed into the numbers. The far end points their camera at their screen. On your screen you will then be able to see both the stopwatch you're sending and the delayed stopwatch you're receiving. Then freeze the screen or take a photo.

The steps creating significant delay are as follows:

1. Actual local event (e.g. human clapping) ->
2. Delay: Local machine capturing ->
3. Delay: Local machine encoding ->
4. Delay: Traveling through Internet nodes ->
5. Delay: Far end decoding ->
6. Delay: Far end TV or projector processing ->
7. Original event reproduced at far end

The time difference in the stopwatch screenshot will be the delay between phases 4–7 (sending) plus phases 2–5 sites reversed (receiving). It is possible to simply divide the difference in two to determine the one-way delay. To add in delay caused by local machine capturing (and encoding), point local camera to the stopwatch and then take a photo on the stopwatch + screen. The time difference in that photo will be the local machine capture (and encoding) delay.

For the low-latency solutions like LOLA (video and audio, 5ms) or JackTrip (audio only, almost no latency) the typical transmission delay for a few hundred kilometers distances would be around 15-30ms. For today's H.323, typically the delay for those distances would be 80-200ms and above for models with high latency. The Internet delay is increased by route distance and routers. Jitter buffer[95], adding to the latency, goes up when connection quality goes down. Wireless local area network will add delay (perhaps 1ms and above) and considerably lower the reliability.

One way to measure the total system roundtrip latency is to use EchoDamp. The preparation is fairly complicated, but for this purpose it's enough to set it up to one side only.

[94] You can use this or some other stopwatch on laptop, just make sure that it runs smoothly: http://www.online-stopwatch.com/full-screen-stopwatch/ or with refresh at eg. 10ms intervals you can use the PC software Xnote Stopwatch: http://www.xnotestopwatch.com/

[95] Jitter buffer: http://en.wikipedia.org/wiki/Jitter#Jitter_buffers

Once it's set, it's easy to use the latency beep function to measure the exact latency (which according to EchoDamp usually somewhat varies in time).

See the next chapter for tips on how to measure the network latency.

4.2 Network tools

The most basic tool is perhaps Ping[96], a very common utility found in for example Windows OSX and Linux installations. It can measure the roundtrip latency to any IP not blocking ICMP (Internet Control Message Protocol) Echo Request. Here are some examples for Windows (OSX version slightly differs):

ping {site}	= ping 4 times using 32-byte packets
ping -n 10 {site}	= ping 10 times using 32-byte packets
ping -t -l 1024 {site}	= ping forever using 1024-byte packets

You can discover your own IP at the command line using ipconfig (Windows) or ifconfig (OSX). To find out latency to each route hop, use Traceroute[97]. You can also use third party servers found at traceroute.org to do the traceroute. To find out the physical location of the server by IP, you can use for example The 81Solutions Server Location Lookup[98]. To find out the global network topology and backbone bandwidths, take a look at Internet topology maps, e.g. GÉANT[99] Maps.

To find out which ports are open[100] and which are closed ('Stealth' means the incoming packets are ignored), use ShieldsUP![101] by Gibson Research Corporation. Or you can use Netcat[102] (nc[103]) for network debugging and investigation[104]. To find[105] out

[96] Ping utility: http://en.wikipedia.org/wiki/Ping

[97] Traceroute: http://en.wikipedia.org/wiki/Traceroute

[98] Server Location Lookup: http://www.81solutions.com/server-location.html (try also the Visual Traceroute found at the same page)

[99] GÉANT Media Library / Maps: http://www.geant.net/Media_Centre/Media_Library/Pages/Maps.aspx

[100] You can also do the following to list open ports: Windows = netstat -an | find /i "listening", OSX = sudo lsof -i -P | grep -i "listen", Linux = netstat -atp | grep -i "listen"

[101] ShieldsUP! firewall port scanning: http://www.grc.com/x/ne.dll?bh0bkyd2

[102] Netcat: http://en.wikipedia.org/wiki/Netcat

[103] You can download nc.exe for Windows at: http://joncraton.org/blog/netcat-for-windows

[104] To check whether a port is open, you can use: nc -v IP PORT (where -v means verbose, IP is the host and PORT is the port number to check)

[105] As a commercial solution for in-depth network analysis, one example is Apparent Networks: http://www.apparentnetworks.com/

throughput, jitter and datagram[106] loss, use iperf[107]. Schroder's article 'Measure Network Performance with iperf'[108] will get you started. Here are two examples for iperf:

iperf -su	= start iperf server at UDP mode
iperf –c {site} –u –b 54m –d	= connect to the server in UDP mode with 54Mbit bandwidth, testing simultaneous upload and download

5 General acceptance of distance teaching

5.1 Standpoints of teachers, students and organizations

Many organizations such as Manhattan School of Music, Canada's National Arts Centre, IRCAM, TCU School of Music, Conservatorio G. Tartini, The Australian National University School of Music and of course the participants in Vi r Music have successfully conducted many master classes and other distance music education. In Vi r Music, the general acceptance of the technology has been good, although all moments of faulty quality are easily noticed by the professional musician or a student. When systems are well built, teachers and students have been satisfied with the system and the image and sound has transmitted without degrading the quality under a level, which would significantly hinder the lesson's purpose. Some teachers and students are understandably skeptical, but after seeing a good distance lesson, they have been seen to change their minds. Of course after seeing a technically failed lesson, they remain skeptical for a good reason. Attitudes have changed to the positive after seeing and hearing how detailed the sound can be and how wonderful an opportunity it is to connect to far sites with great image and audio detail.

5.2 Subjective evaluation of perceived quality in Vi r Music

This is a summary of subjective quality evaluation based on the experiences in the Vi r Music project, written by Dan Nyberg and Jan Berg.

Evaluation of subjective quality of online master classes has been carried out between the autumn of 2009 and 2010. The technology used for the online classes included vide-

[106] Datagram: http://en.wikipedia.org/wiki/Datagram

[107] iperf: http://en.wikipedia.org/wiki/Iperf – download iperf for Windows at http://sourceforge.net/projects/iperf/files/jperf/

[108] Measure Network Performance with iperf: http://www.enterprisenetworkingplanet.com/netos/article.php/3659616

oconferencing equipment, LCD screens, studio microphones and loudspeakers. During these online master classes questionnaires and interviews were distributed and conducted. The locations involved were Oulu (Fin), Helsinki (Fin), Olos (Fin), Rovaniemi (Fin) and Piteå (Swe). The results presented here are the major findings from the subjective evaluation of perceived quality study. The results presents only the experiences of those involved and the results are not necessary generalisable outside the framework of this study. They do however illustrate important topics that are essential to be aware of when engaging in online master classes.

Questionnaire Results

From the questionnaires observations related to these major topic areas were made:

- Sound quality
- Sound and video quality
- Teaching
- Communication

The sound quality related topic contained statements that described the sound as good, natural and as intelligible. The sound and video quality related topic contained statements, which indicated that lack of synchronicity between video and sound as well as delay between the locations was present. The teaching related topic statements described the ability for students to meet other teachers and that the need for travel was eliminated and that both are perceived as positive. The communication related topic contained statements that indicated that there can be problems related to the communication with the teacher and problems for the teacher to indicate to the student when to stop playing.

The number of attained statements from the questionnaire, labeled and sorted based on topics and attitudes, can be seen in table 1. The table illustrates that there are a large quantity of positive statements from the questionnaire and they are related to teaching (Tch) and sound quality (Sqr). It also shows a large quantity of statements that does not have attitudinal content, here referred to as blank statements. The topic containing most blank statements is the sound and video quality (Sqr&Vqr).

Table 1: Attained statements from the questionnaire. Bold numbers indicate high number of occurrences.

	+	-	+&-	Blank	Tot.
Tch	**35**	4	3	1	**43**
Sqr	**17**	4	8	7	**36**
Sqr&Vqr	3	4	5	**14**	**26**

Diverse	6	3	1	8	18
Com	4	5	1	7	17
Vqr	2	1	0	6	9
Tec	0	5	0	1	6
Tch&Com	2	0	0	1	3
Tot.	**69**	**26**	18	**45**	

Interview Results

The topics presented in this section are the major results from the interviews. The interviews were conducted as a second data collection stage, after the first data collection stage, the questionnaire.

The perceived audio and video quality

The teachers and students described the perceived audio quality in the interviews as: metallic, boring, contained no room sound and having a good dynamic. The teachers could however imagine how the instrument sounded based on experience. One could also perceive the small efforts in the students' performance that were not audible.

The video quality was perceived as sufficient for online master classes. The users could distinguish between the system's limitations and the students' limitations when playing.

Perceived problems and possibilities on online master classes

The perceived problems during an online master class included: a lack of synchronicity between audio and video (often audio leading), a delay between the locations, small details in the music is become inaudible, an inability to control muscles and hard to perceive the students' playing techniques.

The perceived possibilities were: ability to control several musical parameters as; tempo, intonation, articulation and phrasing. A small delay is not perceived as problematic if the user is aware of it, thus the user can work around it. It is also perceived that an online master class increases the ability to connect with more people, attain new input on playing techniques and performance, receive comments from other teachers, save money, travels and time.

Perceived differences and similarities between regular and online master classes

Differences; One cannot control the playing technique physically. Instead one has to verbally explain new playing positions. There is also a perceived difference in creating a connection/relation to the student over distance.

Similarities; Meeting with the student and discussing. One can also conduct master classes online seen from an pedagogical point of view, but one has to work around and adapt the teaching to the system's limitations.

Conclusions of the quality evaluation

In conclusion, the presented results show a large quantity of positive statements related to teaching, and to sound quality. Where online master classes gives opportunity to meet and connect with other locations (schools) and people (teachers/students). The system used has a sufficient sound quality for online master classes. It can also be concluded that the video quality is sufficient for online master classes. Teachers can distinguish between the students' efforts and the systems limitations. This latter point indicates that the video conferencing system used is somewhat "transparent" to the user; it does not affect the evaluation of the students' performance. However the delay between the locations and lack of synchronicity between audio and video is perceived as problematic. The results also show that the teaching is not the same as a regular master class and the teacher needs to adapt the teaching to accommodate the systems problems/limitations.