Fundamentals and Principles of Musical Telepresence

Alexander Carôt
Christian Werner
Institute of Telematics
University Lübeck, Germany

Abstract

The idea of playing livemusic with someone abroad represents a major challenge for musicians and sound engineers likewise. Cognitive, technical and purely musical aspects make high demands on a network music performance that should fulfill conditions of a realistic rehearsing scenarios in the same room. In turn the idea of a network music performance has generally been considered as an impracticable application. However, a precise and comprehensive analysis has so far not been published, which equally covers technical, cognitive aspects and their interdependencies equally. In order to give a final and valid statement about the feasibility of distributed real time music we take any of such relevant aspect into consideration and explain it accordingly to the reader. Finally we conclude that in recent wide area networks remote music sessions are possible assuming an awareness of latency conditions and appropriate interaction categories.

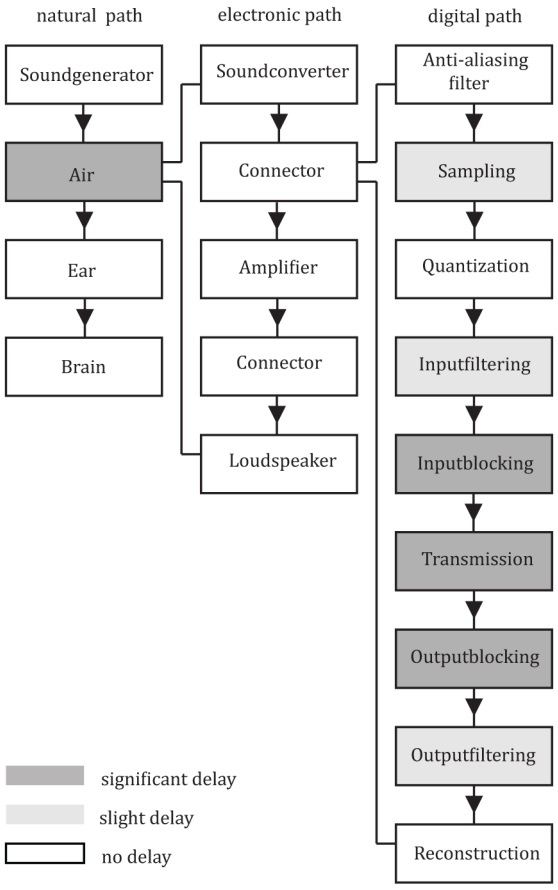
1 | Introduction

The process of hearing implies air waves of changing pressure to trigger a person's eardrum membrane in order to finally convert the signal to electric impulses in the brain [19]. If musicians are rehearsing in the same room they are separated by a certain distance and since the average transmission speed of air lies at 343 m/s, this introduces delays of about 3 ms/m.

Practical tests, in which the physical distance had been increased stepwise, showed that beyond 8.5 m of separation a rhythmical musical interaction without a conductor becomes too difficult in terms of keeping a common tempo.

Such separation corresponds to a one way latency of about 25 ms [5]. Considering the idea of placing two musicians abroad, technology would need to convert sound waves to electronic signals and transmit them on an existing network. Since nowadays audio technology and transmission systems process data digitally, the signal additionally requires a conversion from analogue to digital representation and backwards. Altogether this signal chain consists of the natural signal path, the electronic signal path and the digital signal path withnumerous corresponding stages as illustrated in figure 1.

Each stage is marked in eitherwhite, light gray or dark gray color as an indicator of latency.



01 | Signal paths and stages of the audio transmission process

Abstract

The idea of playing livemusic with someone abroad represents a major challenge for musicians and sound engineers likewise. Cognitive, technical and purely musical aspects make high demands on a network music performance that should fulfill conditions of a realistic rehearsing scenarios in the same room. In turn the idea of a network music performance has generally been considered as an impracticable application. However, a precise and comprehensive analysis has so far not been published, which equally covers technical, cognitive aspects and their interdependencies equally. In order to give a final and valid statement about the feasibility of distributed real time music we take any of such relevant aspect into consideration and explain it accordingly to the reader. Finally we conclude that in recent wide area networks remote music sessions are possible assuming an awareness of latency conditions and appropriate interaction categories.

1 | Introduction

The process of hearing implies air waves of changing pressure to trigger a person's eardrum membrane in order to finally convert the signal to electric impulses in the brain [19]. If musicians are rehearsing in the same room they are separated by a certain distance and since the average transmission speed of air lies at 343 m/s, this introduces delays of about 3 ms/m.

Practical tests, in which the physical distance had been increased stepwise, showed that beyond 8.5 m of separation a rhythmical musical interaction without a conductor becomes too difficult in terms of keeping a common tempo.

Such separation corresponds to a one way latency of about 25 ms [5]. Considering the idea of placing two musicians abroad, technology would need to convert sound waves to electronic signals and transmit them on an existing network. Since nowadays audio technology and transmission systems process data digitally, the signal additionally requires a conversion from analogue to digital representation and backwards. Altogether this signal chain consists of the natural signal path, the electronic signal path and the digital signal path withnumerous corresponding stages as illustrated in figure 1.

Each stage is marked in eitherwhite, light gray or dark gray color as an indicator of latency.

Following, we will describe each signal chain regarding its functionality and the introduced delay. Depending on the actual expected latency between two players we will introduce a number of interaction categories, which also allow compromised musical interplay beyond the 25ms threshold.

2 | Signal path stages

In spite of a relatively large set the majority of signal stages does not, or just slightly introduce, a delay. Within the natural signal chain it is just the physical distance, which has to be kept as low as possible due to the air's transmission speed.

In the electronic signal path the analogue processing and filtering is not associated with a delay. Excluding the analogue anti aliasing [22], reconstruction filter and the quantization process it is mainly the digital signal path, which introduces most of the delay. After passing the anti-aliasing low pass filter, a signal is sampled at a certain sample rate and quantized in a certain bit resolution. The delay of these stages relates to the sample rate and according to the standard of 48 kHz, samples are generated each 20.83 microsecond [22].

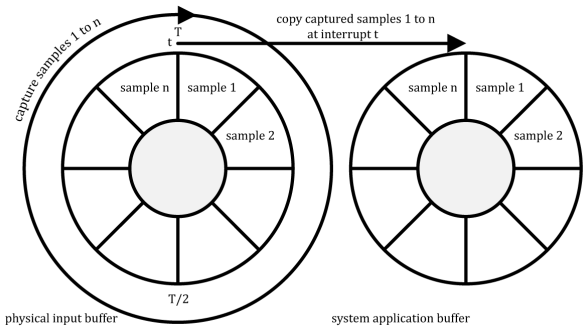
Generally soundcards also apply digital filters to the sampled signal in their input and output sections. What these filters precisely do is a vendor-, device and technology specific criterion but the general purpose is a reduction of signal distortions due to aliasing and quantization noise [22].

In any case filters require a number of input samples to work properly - the so called filter length. After the retrieval of this fixed amount of samples the filter generates output samples according to its actual filter architecture. In turn digital filters introduce delay depending on their filter length in samples. Hence, the higher the sample rate the lower this delay of a filter.

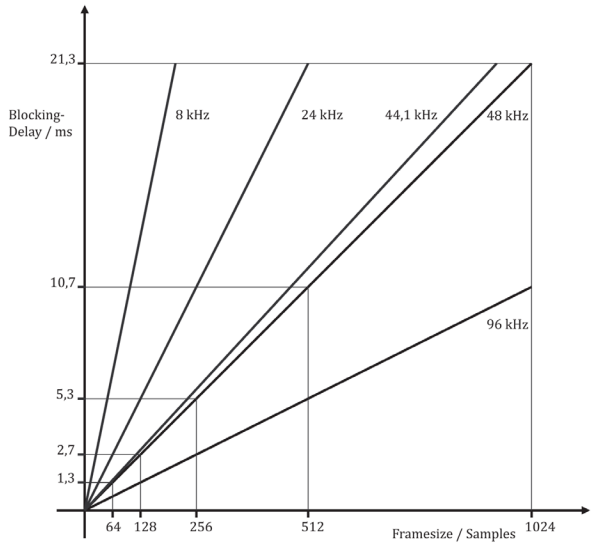
According to the professional RME Fireface 400 soundcard's datasheet the input filter requires 43.2 samples and the output filter requires 28 samples, which add to 71.2 samples in total and leads to a corresponding delay of 1.48 ms at a sample rate of 48 kHz - 0.74 ms at 96 kHzrespectively [4]. However, the major latency appears due to the blocking stages and transmission, which will be outlined in the following subsections.

2.1 | Audio blocking

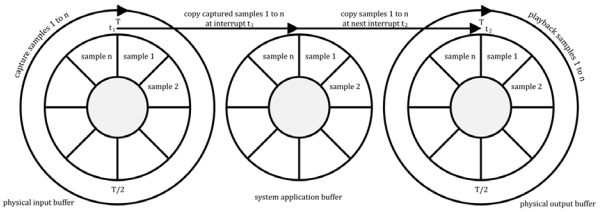
Soundcards can provide each generated sample to the next processing stage, however,current PC based sound systems are not able to process data with that processing speed due to internal scheduling and general architecture. Instead, they have to wait for an amount of samples before actually processing it. Hence, rather than sample by sample soundcards generate audio in blocks of a fixed number of samples [6]. A sample block also corresponds to the terms sample frame or audio buffer. The size of each processed block depends on a sound device system's performance and is typically limited to anumber of 64 samples. Since we generally look at audio data in 16 Bit resolution, that is one sample



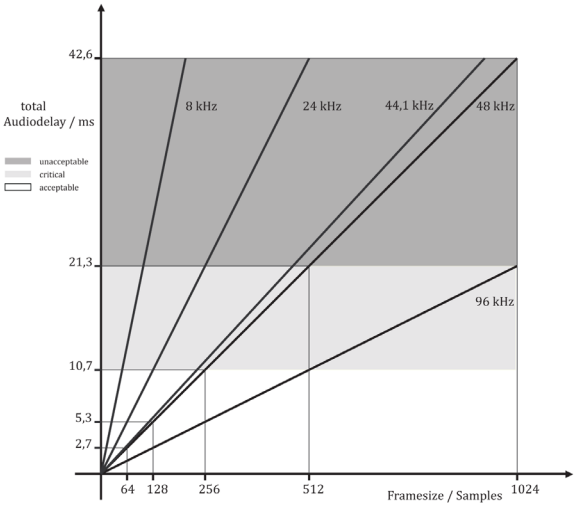
02 | Audio capture process



03 | Audio blocking times



04 | Audio playout process



05 | Total audio latencies

consists of 2 Bytes, the currently lowest possible block size equals 128 bytes.

Without performance optimization of a device's operating system in the form of a special low latency kernel [21] or further technical improvements such as the use of a realtime operating system, current personal computers can handle block sizes of 256 samples/block.

More recently, with increasing CPU speeds, some devices can work with 128 samples/block. In any case blocking implies a tradeoff between stability and delay: On one hand larger blocks require less computational power so that a more stable system behavior can be achieved; on the other hand the delay increases with larger blocksizes since it needs more time to generate an appropriate number of samples.

Given afixed sample rate the blocking delay is directly related to the block size. Assuming afixed block size and increasing the sample rate higher frequencies can be captured, whichleads to a higher audio quality. Additionally, due to smaller sampling intervals, it also leads to proportionally smaller delay times so that altogether the blocking delay depends on the block size and the sample rate. According to this the blocking delay can be calculated with the following equation :

$$\text{Blockingdelay} = \frac{\text{Framesize}}{\text{Samplerate}} \qquad (1)$$

After the soundcard has blocked for a certain amount of samples a filled block or buffer is available to be processed by a digital device or computer application. Practically this means that the buffer values can be modified in any way the user desires to.

In common computer operating systems (Windows, Mac OS X, Linux), so called "callback functions" are used by the audio device as a programmer's interface for the retrieval and modification of such audio blocks.

Their execution corresponds with a sound card's interrupt and following the previous equation this interval depends on the actual soundcard settings. As a standard case a sampling rate of 48 kHz with a framesize of 128 samples results in callback events every 2.7 ms.

Figure 2 shows the principle of the capture process, which is realized as a double buffer system. The digital application processes the filled buffer's *n* samples while at the same time another buffer is filled with the next *n* samples. Once the buffer is full, both buffers are swapped and the next period starts in the same way. At the moment *t* the callback event occurs after all sample values had been captured and made available to the application. The interval

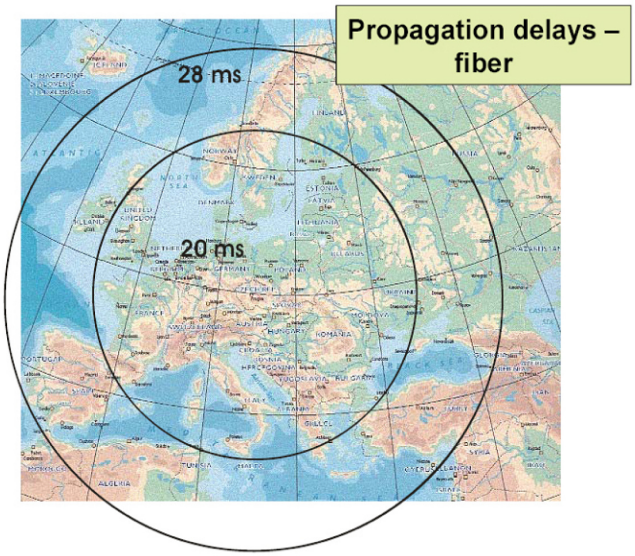
T represents the blocking delay.
At that state the soundcard has just captured audio data, which is now ready to be processed and actually ready to be sent to a desired destination.
Figure 3 graphically shows the relation between block size, sample rate and a sound card's blocking delay. E.g. a block size of 512 samples corresponds to a blocking delay of 10.7 ms at a sample rate of 48 kHz. However, the system has not created any audible feedback yet. The direct playback of a captured signal block also refers to the term "loopback" [4] and would require the buffer to pass the soundcard's output process as the next step. The soundcard's output architecture equals the input section and is as well designed as a double buffer system. While the second playout buffer offers its samples to the physical output, the application buffer holds the samples previously captured by the system's physical input, which as well introduces a delay of one audio block. As a result the blocking delay of a soundcard appears twice, firstly by the input double buffer and secondly by the output double buffer.
Figure 5 shows the total soundcard delay and marks the acceptable sound card configurations with aresulting total delay below the 25 ms delay threshold on a white, close to 25 ms delay ona light gray and beyond 25 ms on a dark grey background. As the user is supposed to figure the lowest possible latency a sound system can achieve, this graph helps to verifythe system's ability to suit the requirements of distributed music performances.

2.2 | Signal Transmission

Like analogue signals, digital signals can be transmitted on any kind of cable or radiowave, but especially for long distance digital connections, fiber optic lines are used, which transmit the binary data via light impulses. Their maximum speed lies slightly below the speed of copper at 0.7 * c [24].
In case two musicians were connected with a fiber optic link and if we deny the existence of the previously mentioned soundcard delays or additional sources of delay, the propagation delay would only depend on the speed of fiber and would allow a fiber length of 5250 km as calculated below.

$$\frac{300,000 \text{ km}}{1000 \text{ ms}} \cdot 25 \text{ ms} \cdot 0.7 = 5250 \text{ km} \qquad (2)$$

If a fiberoptic cable was to run straight across the european continent, it would lead to a maximal one-way delay of about 28 ms as illustrated in figure 6 [24].
Apart from the fact that sound data can be transmitted via fiber optic or copper lines with their appropriate signal speeds of 0.7 * c or 0.8 * c, a further transmission speed limitation has to be taken into account : Digital networks are equipped with network gear such as switches or routers, which basically guide each light impulse through a network in order to make it reach its correct and final destination. It is rather the capacity of a network device, which determines the delay for a signal to be mapped from one input port of a network device to an output port and this capacity is typically stated with the amount of bits per second (bps). The core of network nowadays consists of network devices with capacities of several Gbps (gigabits per second). These so called backbone connections are able to serve several millions of users directly or indirectly by linking traffic to or from interconnected sub networks.
Device and network capacities depend on the amount of users a network has to serve. The lower the amount of users, the lower the bandwidth capacity a network is generally administrated with.



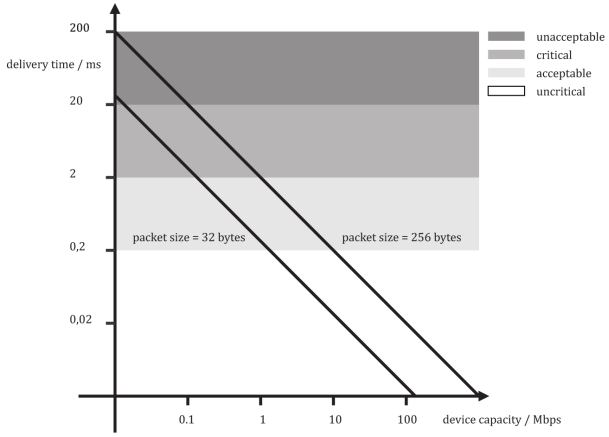
06 | Propagation delay of fiber cables

Nowadays current backbone links provide capacities from 10 Gbps to 5 Gbps down to 2.5 Gbps or 1 Gbps, in contradiction a user's end point connection is able to achieve about 20 Mbps down to 128 kbps or even 56 kbps in worst case. Within these two extremes further links typically range from 10, 34 or 100 Mbps up to 622 Mbps as common figures [27]. With respect to the sound data the total required bandwidth equals the product of the sample rate, the bitdepth and the number of channels, which equals 768 kbps for one audio stream with the standard settings of sample rate = 48 kHz, bitdepth = 16 Bitand a channel number of 1 as calculated in equation 3.

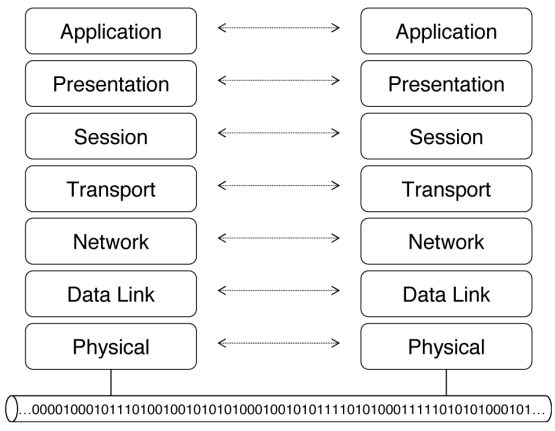
$$\text{Totalbandwidth} = 48 \text{ kHz} \cdot 16 \text{ Bit} \cdot 1 \text{ channel} = 768 \text{ kbps} \qquad (3)$$

This pure amount of audio data also refers to the term payload. As a precondition for a successful data delivery any involved network component has to be able to handle at least this amount of payload data. In terms of latency an audio chunk of 256 bytes (128samples * 16 Bit) already requires 16 ms to be fully transmitted on a 128 kbps link, in contradiction to just 25.6 ns of a 10 Gbps backbone device. This relation is illustrated in figure 7.
With respect to the 25 ms delay threshold critical delay are marked in gray, the darker the tone, the less acceptable the latency of the transmitting component. Hence, when estimating an audio block's transmission latency, it is firstly the speed of light multiplied with the sum of cable lengths between involved network components and secondly the sum of the respective packet propagation times, which finally determines this value (see equation 4).

$$\text{Networkdelay} = \sum_{k=1}^{M} \text{Devicedelay} + \sum_{k=1}^{N} \text{Wiredelay} \qquad (4)$$



07 | Network device transmission latencies



08 | ISO/OSI Stack

Considering a lower audio quality with a corresponding lower bandwidth, audio blocks can be compressed or decimated, which in turn reduces the stated block sizes and delays by a fixed factor of 2 to 8 as described in [6]. As an example figure 7 additionally illustrates the propagation delays of a 32 bytes packet, which corresponds to a decimation or compression factor of 8 compared to the original 256 bytes packet.

3 | Transmission systems

Network technology is typically more than a simple cable connection between two devices. Apart from the pure delivery of electronic data on the cable, there is further complex logic behind a functioning network. In order to explain these logical structures we are using the ISO/OSI protocol stack [27], which is a very general and widely accepted network model. It is illustrated in figure 8. The model is divided into seven layers, eachone representing certain network functionalities. At the very bottom we see the physical layer that represents the direct link to the physical network media. On this layer we face hardware components and physical signals. The data link layer represents technologies for concurrently accessing the network media. It determines fundamental parameters how and in what sizes data is transmitted on an underlying physical medium and can as well define mechanisms for error detection. Then network layer is responsible for addressing packets (host-to-host) and for directing packets through the network, while the transport layer adds more fine-grained addressing functionalities and communication rules between two processes. Furthermore the session layer is needed for establishing a stateful communication between two processes. The presentation layer is responsible for adapting varying data representations between the communication parties and finally the application layer is providing an interface for the actual network application. In figure8 we see two instances of this protocol stack, representing two communicating hosts in the network. Through this protocol stack both host are able to interchange data.
Each layer can virtually communicate with its remote counterpart (horizontal arrows), but physically the data is always sent via the underlying layers, finally traveling bitwise through the physical media. After the remote side receives the data on the physical layer it is passed upwards again [18].
As the applied technology in the lowest two layers determines the basic transmission characteristics and rules in a network of interconnected devices it has the most fundamental significance concerning the fulfillment of the low delay restrictions of a distributed music system. As all layers have a hierar-

chical dependency a higher layer cannot be able to speed up possible delays in a lower layer and hence it is the Physical and the data Link layer, which characterize the basic transmission behavior of a network. Theoretically a pure and single cable connection between two devices represents this most direct and fastest case of device interconnection. However, the advantage of this simple and clear connection implies significant drawbacks.

With an increasing number of network hosts a network's complexity rises which in turn complicates the administration of *n* explicitly used lines and as well increases the probability of a line defect. Furthermore the explicit use of a single line can be considered as inefficient, since theoretically a transport medium can be used for the propagation of more than a single signal at the same time. The principle behind this is called multiplexing [27], which will be covered in the following.

Whenever a sender or a receiver consists of more than one entity data has to be multiplexed. Major multiplexing technologies are frequency division (FDM) or wavelength division (WDM) and time division multiplexing (TDM) [27]. By FDM a connection is separated into different frequency bands, which are finally used as separate transmission channels for numerous signals. Each signal in modulated according to the desired frequency band and decoded at the receiving end. This allows simultaneous transmission on a single cable. The most famous applied field of FDM is the transmission of radio waves trough the air, where each radio program has its own carrier frequency. As numerous destinations might want to receive different radio programs each radio receiver can betuned according to the desired carrier frequency. Similarly WDM follows the same approach but appears in context with fiberoptic lines. Rather than frequency modulation it uses different wavelengths (or colors) as carrier for the respective information channels on a fiber. Finally, the major multiplexing technology in terms of digital communication is the time division multiplexing (TDM), which transmits a number of signals in a time shifted and interleaved manner on the same transport medium. TDM can be broken down into the synchronous (STDM) and the asynchronous (ATDM) mode. For STDM the time domain is divided into several recurrent timeslots of fixed length, one for each virtual channel, the respective data block or STDM frame is transmitted in. After the last channel has sent the data in its respective timeslot the cycle starts again with a new frame, starting with the second data block of the first channel. Due to this synchronous clock based approach STDM also refers to the term "static multiplexing". In contradiction to STDM where network clients are served in their respective time slot even if no transmitting data is available, an ATDM network does not carry any data in case of a transmission break or interruption. This asynchronous and non-clock based approach refers to the term "statistical multiplexing". Based on the appropriate multiplexing techniques current networks can be divided into three categories: asynchronous,synchronous and isochronous networks [16], which are explained in the following subsections.

3.1 | Asynchronous networks

Asynchronous networks are based on ATDM, which implies that signal transmission happens randomly depending on stochastic host transmission activities. Furthermore any network host is allowed to send a desired amount of data in a blockwise restricted manner, that is, if a sequence of bytes exceeds a certain upper size limit, this sequence will be split into chunks of that maximal size. Such chunks of data are also considered as packets, blocks or frames and their maximal size is called the maximum transfer unit (MTU) [27]. A drawback of variable packet sizes is, that larger blocks block the transmission medium for a longer time than smaller blocks. This delay variation is generally called network jitter. It has a negative impact on realtime traffic as we can imagine especially in context with the delivery of low size audio chunks as outlined in the previous sections.

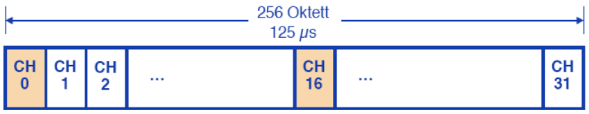
As the playout buffer of a soundcard expects data in constant time intervals depending on the actual blocking delay, sudden increases of the network delay would violate this precondition of constant and error free audio stream playback. Due to thisproblem asynchronous networks were originally designed as database or non-real-time communicationsystems, which do not make strong demands on the network in terms of delay or jitter. Examples of asynchronous communications systems include general purpose serial interfaces such as RS-232 and RS-485 or their successor the Ethernet, which is the most significant and currently most often applied local area network (LAN) technology [27]. Ethernet typically works with a MTU of 1500 bytes and its current transmission capacities range from a minimum of 10 Mbps to a maximum of 10 Gbps. Assuming a 10 Mbps Ethernet network component, a 1500 byte packet would block the medium for 1.3 ms. With a 100 Mbps connection the propagation time for a 1,500 byte packet is 10 times lower at 0.13 ms. Within the last years Ethernet has become the major data link layer technology for the Internet, which it self uses the asynchronous IP protocol on the network layer [23].

3.2 | Synchronous networks

A synchronous network hosts do not have such operational freedom as they have in an asynchronous network. Due to the clock driven approach of STDM each host is served at its reserved fixed time slot in terms of sending and receiving a fixed amount of data, which length and size depends on the given network specification. E. g. in the synchronous telephone network end point devices work with a sample rate of 8 kHz and in turn the network clock's frequency equals 8 kHz.

As a sample rate of 8 kHz results in a sample generation every 125 microsecond, the total duration of a respective STDM frame m atches this value in order to transmit the signal in a "sample by sample" manner, rather than packet by packet as in the asynchronous case [16]. This principle is illustrated in figure 9 for a 32 channel STDM frame. For one frame the number of served hosts depends on the network's transmission capacity, which technically determines the number of slots for one frame. The most famous example of a synchronous network architecture is the Synchronous Digital Hierarchy (SDH) [16], which is used in almost any telephone communication backbone all over the world and nowadays is mainly applied in combination with ISDN (Integrated services digital network) [26] in the top layer [27]. The most significant advantage of the synchronous approach is that due the fixed time slots and equal data sizes the problem of network jitter does not occur.

Each transmission has the same duration and is processed in constant time intervals.



09 | Principle of a synchronous TDM frame

3.3 | Isochronous networks

An isochronous network can be described as a compromise between a synchronous and an asynchronous network as it features aspects of both of them. Identically with anasynchronous network the network link is idle as long as no data is actually transmitted.

Furthermore hosts are not restricted to certain timeslots and fixed data sizes so that an isochronous network can identically be used like an asynchronous data network. However, in order to reduce the effect of network jitter and in turn equally provide decent conditions for synchronous realtime traffic it introduces the concept of cells, that is, any packet of data is split into a number of equally small sized cells, which require significantly less time to be transmitted [11]. Cells are generally not broken down to a single byte or sample as it can be the case in a synchronous networks, but sizes between 8 bytes and 100 bytes are not uncommon and can be considered as extremely small, especially if the corresponding full packet size might consist of 1500 bytes or possibly more. Furthermore an isochronous communication system is intended to deliver quantifiable performance, which is realized in a service agreement on the connection between communicating network nodes. This service agreement specifies criteria such as bandwidth, delivery delay and delay variation on a special protocol sublayer [11]. As a result isochronous transports are capable of carrying a wide variety of traffic and thus represent a versatile and flexible networking concept. Examples include IEEE-1394 (firewire) or USB (universal serial bus) as local device connectors. However, in terms of a wide area network with numerous hosts the best known isochronous network is the asynchronous transfer mode (ATM). ATM splits the data into cells of 53 bytes, which require 28 timesless transmission time than full Ethernet frames of 1500 bytes [18].

4 | Distributed music in current networks

Until the late 1960's, wide area networks were commonly available in terms of voice telecommunication only [13]. Asynchronous data networks played no significant role in that context as they were mainly intended for data retrieval in local area networks of companies or as control instances in manufacturing processes. Hence, in the past we could clearly distinguish between synchronous telecommunication networks and asynchronous industrial data networks. Although in the first decade of the 21st century this separation still exists, it is fairly not as strict as it used to be. Following McLuhan in [20], that sooner or later every kind of medium will finally end up in a single multipurpose medium, customers more and more asked for multimedia and communication services for audio and video on one hand and at the same for services in data retrieval and non-realtime communication on the other hand. In order to be prepared for the new mixture of real time and data services telephone providers consequently had high hopes in a new broadband ISDN (B-ISDN) [26], which preferably applied the isochronous ATM in its data link layer. Things, however, developed differently than expected as people became more and more attracted to the Internet, which nowadays offers an enormous amount of data hosted on globally interconnected machines and can already be considered as the major commonly available source of information. Since the Internet has such an impact on our daily life, the idea of a network that covers

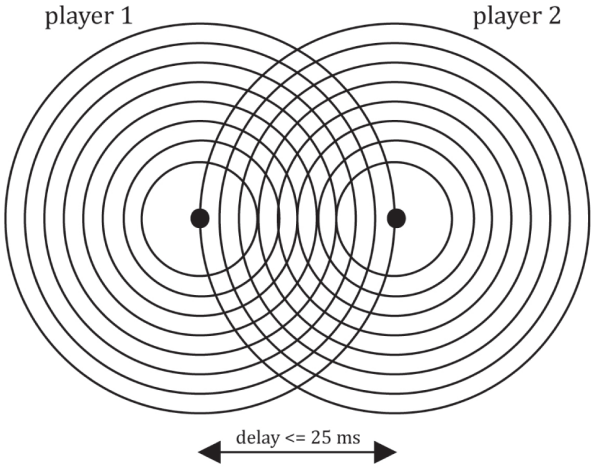
any desired service, has become attractive to users and eventually to the industry. Hence and despite the drawbacks of asynchrony in 1994 first services offering Internet radio came to life, followed by Internet telephony (VoIP) and video conferencing[28]. These services have been improved constantly over the years and about 10 years later they have become extremely reliable, so that realtime communication on the Internet can nowadays compete with services such as email or web browsing [15]. The famous VoIP tool "Skype" [3] represents a good example for this tendency.

The common approach to compensate the effect of network jitter is the application of a jitter buffer at a receiver's end. By storing up an amount of audio packages in the network queue the audio process can still providea solid playback in case of late arriving packages. The drawback of this principle is a higher latency as packages are not processed right after the reception [17]. There is no doubt that the Internet as an asynchronous medium is not the predestinated choice for the carriage of synchronous data and especially not for low delayed audiostreams with their according low block sizes. Nevertheless with constantly increasing bandwidths and respectively increasing propagation speeds the effect of network jitter has been loosing significance over the years, which appeared to make the cell based ATM become obsolete and finally lead to the consequent replacement with the Ethernet technology. Furthermore, due to the Internet's large distribution and its strong sociocultural establishment the Internet already beats the outdated synchronous telephone network so that further investigations in distributed music systems will use the Internet as the prime transmission system of choice. However, apart from the applied network technology it is clear that even in case of a direct fiber optic link between two players the maximal distance might not exceed 5250 km as outlined in the previous section. On top of that cable detours due to signal routing, device propagation delays and soundcard delays decrease this maximal playable distance typically below 1000 km. Especially in a worldwide scenario delays of more than 100 ms are likely to expect and thus making distributed music impossible. However, taking further cognitive aspects of musical interaction into account it is possible to find compromised ways to cope with latencies beyond the 25 ms threshold.

In that context we generally have to distinguish between a solo instrument and a rhythm instrument. Apart from that the placement of the so-called "rhythm section" is of significant importance. As an example with drums, bass and saxophone, a simple case is present in which drums and bass form the "rhythm section" while the saxophone player represents the "solo section". Though of course the solo section and any musician must have a sense of rhythm, it is basically the interplay of bass and drums, which forms the essential fundamental groove of an ensemble which allows other solo instruments to play upon. In this scenario the saxophone player relies on and plays on the groove that is produced by the rhythm section [8]. Due to the fact that rhythm and synchrony are the main fundament of groove based music, the following subsections put emphasis on rhythm based instruments and the groove building process. In classical music things are more complex. Here we usually cannot precisely distinguish between "rhythm" and "solo" sections [8]. Anyhow, in most pieces of classical music an analogical categorization is feasible, but should be more fine-grained and more dynamic. Also the concept of a conductor has to be considered here. In the following, in order to present our concepts as clear as possible to the reader, we will focus on applications in the field of rhythmical music and continuously use the according terms "solo section" and "rhythm section". Based on the actual delay between two players, we can separate the possibilities of amusical interplay into four main categories.

4.1 | Realistic Interaction Approach (RIA)

A realistic musical interaction, as if in the same room, assumes a stable one-way latency of less than 25 ms [5] between two rhythm-based instruments such as drums and bass. In this scenario both instrument's grooves merge into each other and the real musical interplay can happen [9]. From the perceptual point of view the delay appears to be as not existing, which is similar to musicians playing with a maximal physical distance of about eight meters in a rehearsing space, where the speed of sound is the limiting time delay factor. The RIA is the only approach pro-



10 | Realistic interaction approach

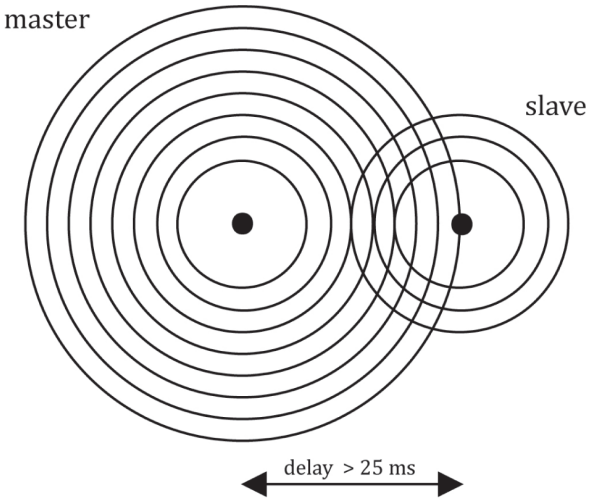
fessional musicians accept without any compromise since it is the only scenario, which exactly represents the conventional process of creating music in groups or bands.
Beyond this threshold of 25 ms, the groove-building process cannot be realized by musicians anymore and thus different compromises and categories have to be applied [8]. Figure 10 shows that below 25 ms delay both players are able to play at the same instant and receive each otherís signalsas if no delay was existent.
Due to technical difficulties in applying the required RIA conditions, RIA has so far not turned into a commercial entity but has mainly been examined in research projects, such as SoundWIRE by Chris Chafe of CCRMA [10] our Soundjack system [7].

4.2 | Master Slave Approach (MSA)

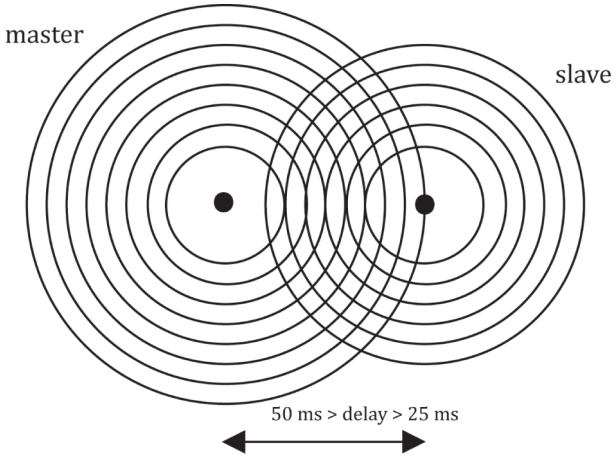
Assuming an attendance to compromise and to step back from musical perfection and
ideals, it really is feasible to perform with two rhythm-based instruments such as drums
and bass, even when exceeding the 25 ms threshold, simply if one of the musicians keeps track of his rhythm and does not listen to the incoming high delayed signal anymore. In that situation the remote side can perfectly play to the incoming signal since the other side doesn't care about the response anymore - a change in the musical interaction is happening, which here is called the “Master-Slave Approach”. The first musician takes the master role since he is producing the basic groove while the remote musician simply relies on it and hence takes the slave role [25]. Of course the higher the delay, the more difficult the ignorance of the delayed input can be realized by the master since shorter delays will easier establish a musical connection to the previously played notes. In terms of delay MSA generates no latency and perfect sync on the slave's side but on the other hand it delays the slave with the round trip delay on the master's side. While the slave musically depends on the master but has a perfect sync, the master has musical independency but an unsatisfying sync [8].
Figure 11 shows a situation with a delay beyond 25 ms delay between two players. Due to the high delay the slave has to wait playing until the master’s signal has arrived, which finally leads to a roundtrip delay on the master’s end.
In general the master role is taken by a rhythmic instrument in order to let solo instruments play on its groove in slave mode. An exception can happen when a rhythmic instrument suddenly starts with a solo part. In this case it will require the other instrument to take over the leading rhythmic role, which in turn leads to a switch of roles. MSA can be applied with any system that allows the transmission of realtime data on the Internet. This could be tools for IP telephony or video conferencing, which do not put emphasis on low delay signal transmission, but as well high speed audio transmitters in an intercontinental setup. In the latter case the main source of latency is the long physical distance.

4.3 | Laid Back Approach (LBA)

The Laid-Back Approach is based on the”laid back” playing manner, which is a common and accepted solo style in jazz music. Playing “laid back” means to play slightly behind the groove, which musicians often try to achieve consciously in order to make their solo appear more interesting and free. The



11 | Master/slave approach

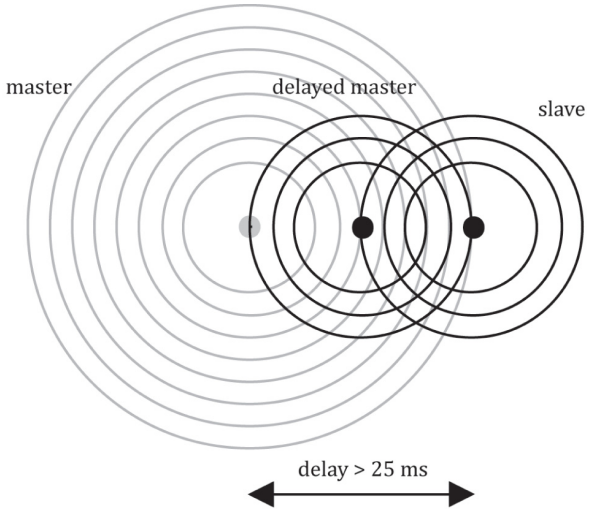


12 | Laid back approach

Laid-Back Approach is similar to the Master-Slave Approach and is mainly determined by the number of participating instruments and their role. As previously mentioned, two rhythm based instruments separated by delays beyond 25 ms have to play with MSA but in case one of the instruments being a solo instrument, the situation changes. Exchanging the drums with a saxophone in the example scenario results in a remote rhythm solo constellation in which the bass represents the rhythm instrument and the saxophone the solo instrument.
Since the bass now has no rhythmic counter part anymore, it alone takes the responsibility for the groove while the saxophone plays its solo part on it. Equal to MSA to saxophone has a perfect sync on its side and is transmitted back with the roundtrip time but in comparison to MSA this has no disturbing effect for the rhythm instrument in LBA. The saxophone is delayed by the roundtrip delay time, which adds an artificial laid back style on it and hence this playing constellation is not to be considered as problematic anymore. LBA of course does not work for unison music parts in which both parties have to play exactly on the same beat at the same time. The perceived roundtrip delay on the master's end ranges between 50 ms up to a maximum of 100 ms but still depends on the musician's subjective perception and the bpm (beats per minute) of the actual song [8]. Figure 12 equals the MSA principle but due a maximal one-way delay of 50 ms and the determination of a rhythm and a solo section this situation leads to an artificial “laid-back” effect.
LBA is used when the delay ranges in areas slightly beyond the 25 ms RIA threshold. Again SoundWIRE and Soundjack represent potential candidates, beside the Musigy [2] software as one of the fist commercial products. It provides audio delays, which range at the edge between RIA and LBA.

4.4 | Delayed Feedback Approach (DFA)

In case the 25 ms delay threshold is exceeded, DFA tries to make musicians feel like playing with the RIA by delaying player's own signal artificially.
By principle delays beyond 25 ms lead to either LBA or MSA styles in which the master hears the slave with a delay equal to the roundtrip time while the slave plays in perfect sync. When delaying the playback of the master's signal, both sounds finally have a closer proximity at the master's ear, which improves the problematic delay gap in MSA or reduces the laid back effect in LBA. The larger the self-delay the better the synchronization of both signals. The best synchronization can be reached with a self-delay equal to the roundtrip-time. This principle is illustrated in figure 13. Anyhow, we have to mention that a spontaneous switch in the master-slave's role



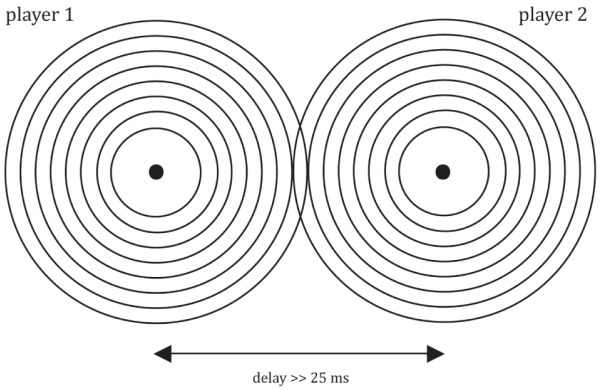
13 | Delayed feedback approach

will lead to a worse situation for both players than it would appear without an artificial delay. In this particular scenario the none-delayed master performs like in a normal not delayed MSA or LBK scenario but the now delayed slave will not to be able to play in sync with its master track due to a possible confusion by its own delay [8]. Though DFA improves the delay situation between two musicians, it is no doubt that a delay of one's own signal typically can be considered as inconvenient and not natural. The larger the delay gets and the louder the instrument's direct noise, theworse the realistic instrument feel and playing conditions. This is especially valid for any acoustic instrument such as a violin or drums. On the other hand DFA can be a suitable approach for the synchronization of remote playback sound sources. In case of e.g. two DJ's turntables are connected with each other, a delay of the turntable's output would not lead to timing-problems. Unlike human beings a machine's playback behavior does not depend on an inner time or feel and hence can of course reproduce delayed sounds without loosing any kind of rhythm. Systems based on DFA are eJamming [1]and the NMcP project of a research group at the University of Braunschweig, Germany[12].

4.5 | Latency accepting approach (LAA)

While all previous approaches try to find alternative ways for realistic network music performances, the latency accepting approach steps back from latency optimized or compromised solutions and simply accepts delays beyond 25 ms. In principle LAA has no motivation to create conventional music and thus can allow any delay, which is consciously taken it into account. In this scenario musicians play with the delay and use it as an artistic way of expression.

LAA is the most avantgardistic approach resulting in a total dissociation of musical conventions and function with the Internet as the core technology. The latency between the players in figure 14 has such a strong dimension that their rhythmical interaction has no relation anymore. In terms of new avantgardistic music in LAA, the Quintet.net framework by Georg Hajdu [14] fulfills relevant requirements and can be applied under any kind of network condition. Quintet.net transmits MIDI control data and does not necessarily require the user to play a musical instrument. The user can play with an electronic input device for the sound generation instead. Apart from that, various worldwide network sessions with SoundWIRE have taken place, in which modern, new music is the dominating style of performance [8].



14 | Rhythmically unrelated sound sources in LAA mode

5 | Conclusion and Future Work

Apart from audio engineering, network and music skills, the awareness of delay dimensions and their musical consequences is the main basic requirement for a successful network music performance. In order to give the potential user this ability, this article firstly describes the main technical fundamentals, which determine the latency between separated musicians and secondly explains the related categories of delay in influenced musical interaction. Depending on the actual network connection and the respective delays, the user can now consciously apply the suitable category of musical interplay, which allows him to perform under any given network situation. In parallel this gives him awareness of actual possibilities and limitations in his current situation. However, due to the high amount of interdisciplinary knowledge, distributed music has so far mainly been used by a small community of experts in IT as well as in music so that it cannot be considered as a major technology for musical interaction yet. Despite the existence of first commercial products, musicians and sound engineers remain passively in terms of accepting and applying this new approach. As the technical facts clearly prove the feasibility of network music performances, with this article we hope to motivate musicians and engineers to take advantage of the musical possibilities distributed music can offer.
In the future we will further investigate in the realistic interaction approach for the Internet in order to overcome the high challenges of asynchronous low latency network engineering and finally increase the radius, in which RIA can be applied.
As a new field of interest we will examine the delay and interaction restrictions for conducted orchestrated music with the final goal of developing a decent and versatile low delay audio and video streaming solution.

REFERENCES

[1] ejamming website : www.ejamming.com, August 2008.

[2] Musigywebsite : www.musigy.com, August 2008.

[3] Skypewebsite : www.skype.com, August 2008.

[4] RME Audio. Fireface 400 datasheet, 2007.

[5] Alexander Carôt. Livemusic on the internet. In Diplomarbeit, Fachhochschule Lübeck, 2004.

[6] Alexander Carôt, Ulrich Krämer, and Gerald Schuller. Network music performance in narrow band networks. In Proceedings of the 120th AES convention, Paris, France, May 2006.

[7] Alexander Carôt, Alain Renaud, and Bruno Verbrugghe. Network music performance with Soundjack. In Proceedings of the 6th NIME Conference, Paris,France, June 2006.

[8] Alexander Carôt and Christian Werner. Network music performance - problems, approaches and perspectives. In Proceedings of the “Music in the Global Village” - Conference, Budapest,Hungary, September 2007.

[9] C. Chafe, M. Gurevich, G. Leslie, and S. Tyan. Effect of time delay on ensemble
accuracy. In Proceedings of the International Symposium on Musical Acoustics,
Nara,Japan, March 2004.

[10] C. Chafe, S. Wilson, R. Leistikow, D. Chisholm, and G. Scavone. A simplified approach to high quality music and sound over ip. In Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFXñ00), Verona,Italy, December

2000.
[11] Martin P. Clark. ATM Networks ñ Principles and Use. Wile-Teubnerl, 1996.

[12] Xiaoyuan G, Matthias Dick, Ulf Noyer, and Lars Wolf. Nmp - a new networked music performance system. In Proceedings of the 4th NIME Conference, June 2004.

[13] Richard T. Griths. History of the internet, internet for historians :
www.let.leidenuniv.nl, August 2008.

[14] Georg Hajdu. Quintet.net ñ A quintet on the internet. In Proceedings of the International Computer Music Conference, Singapore, 2003.

[15] International Telecommunication Union (ITU). The status of voice over internet protocol (voip) worldwide, 2006. In The Future of Voice, 2007.

[16] Gary C. Kessler and Peter Southwick. ISDN - Concepts, Facilities, and Services.
McGraw-Hill, third edition, 1990.

[17] A. Kos, B. Klepec, and S. Tomazic. Techniques for performance improvement of VoIP applications. In Proceedings of the 11th Electrotechnical Conference MELECON,
2002.

[18] Ulrich Krämer, Jens Hirschfeld, Gerald Schuller, Stefan Wabnik, Alexander Carôt, and Christian Werner. Network music performance with ultra-low-delay audio coding under unreliable network conditions. In Proceedings of the 123rd AES - Convention, New York, USA, October 2007.

[19] Levine and Shefner. Fundamentals of Sensation and Perception. Oxford University
Press, "third edition" edition, 2001.

[20] Marshal Mc Luhan, Quentin Fiore, and Jerome Agel. The Medium is the Massage: An Inventory of Effects.Gingko Press, 1996.

[21] Dave Philips. Computer music and the linux operating system : A report from the front. In BT Technology Journal, 2003.

[22] Ken C. Pohlmann.Principles of Digital Audio. The Mcgraw-Hill Companies, fifth edition, 2005.

[23] J. Postel. RFC 791: Internet Protocol, September 1981.
[24] Lars Arne Roenningen.Adaptive video resolution and traffic control in packet networks.
In Guir, UC Berkeley, 2004.

[25] Nathan Schuett. The effect of latency on ensemble performance. In Bachelor Thesis, CCRMA Department of Music, Stanford University, 2002.

[26] William Stallings. ISDN and Broadband ISDN with Frame Relay and ATM. Prentice-Hall, third edition, 1995.

[27] Andrew S. Tanenbaum. Computer Networks. Pearson Studium, fourth edition, 2003.

[28] James R. Wilcox. Videoc onferencing - The whole picture. Telecom Books, third edition, 2000.