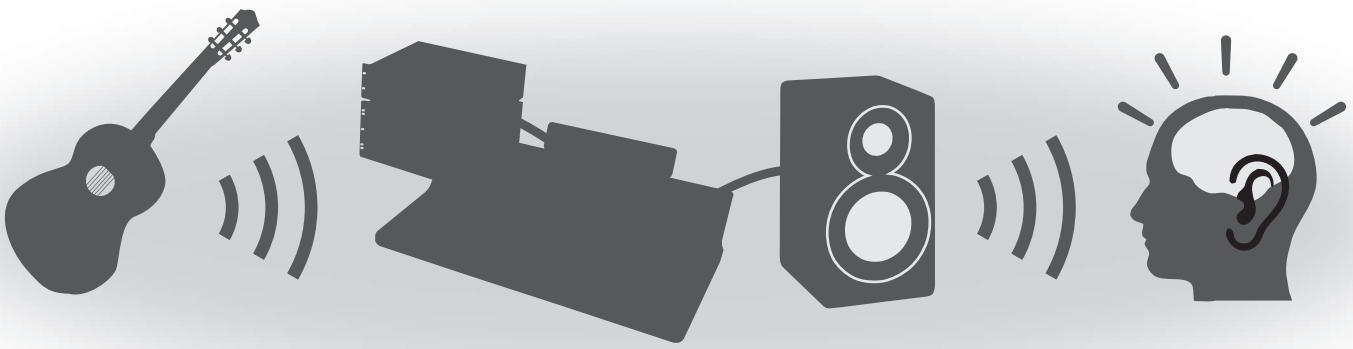


# Audio quality in networked systems

This white paper presents a summary of the most important quality issues in networked audio systems to support system designers and sound engineers in maximizing their system's *Performance* and *Response*, achieving the maximum *Audio Quality* and *Sound Quality*.

# Table of Contents

# Preface

Since the first professional digital audio systems were launched in the 1980s, the transition from analogue to digital in the audio industry has been close to completed: the majority of professional audio systems in the world use digital components for mixing, processing and distribution. The coming 10 years will show a follow-up transition with a similar impact: from closed systems  based on point-to-point distribution topologies to open, integrated networked systems. Just as the analogue to digital transition posed new challenges for audio quality and sound quality, the transition to networked systems poses new issues - particularly concerning level and time coherency.

This white paper presents an integral analysis of quality issues in networked audio systems. It is assumed that the reader is an audio professional, familiar with audio technology, components and standards. Also, it is assumed that the reader is familiar with interpretation of level diagrams, and also with representations of signals in the time domain and frequency domain. Knowledge of digital systems and network technologies helps, but is not necessary to read the document - the authors tried to present the subjects without too much mathematical and IT jargon.

This white paper's subject is a technical one: *networked audio systems*, an area that is complex by nature because it comprises systems rather than single components - increasing not only creative possibilities but also complexity. The introduction of networking technologies in the live sound field further makes designing and operating systems more complex. (but also more exciting of course).

Discussing audio quality and sound quality however is not only a technical discussion - it is also a philosophical one. Before discussing anything technical, we first have to agree on the philosophical meaning of audio quality and sound quality - and share the same definitions and wording. These are presented in chapters 1 and 3 of this white paper.

Reading chapters 1 and 3 at first might be a bit odd for an audio professional, as the language used to describe audio quality and sound quality issues is a bit different from the language normally used in the professional audio field. Before reading the other - more technical - chapters, we do ask you to read these two chapters carefully because the definition and wording of quality issues is very delicate. For example: *audio quality* and *sound quality* are presented as completely different concepts, where in every day discussions the two often are mixed up. Without a thorough understanding of the difference, the technical chapters in this white paper - often referring to the concepts presented in chapters 1 and 3 - might not always make sense...

This white paper is conceived to provide an overview of quality issues in networked audio systems. By no means it is a complete or even accurate overview - a detailed presentation of every issue is far beyond the scope of this white paper. We encourage readers to study the materials suggested in the 'further reading' section in appendix 2.

With the launch of the CL series digital mixing systems in 2012, advanced *networked audio* technology with the *natural sound* philosophy have been introduced to the professional audio field, marking a land sliding change in the way live audio systems are designed and used. To support system designers and sound engineers to utilise these new possibilities optimally, this white paper introduces the concepts of '*Performance*' and '*Response*' to clarify the natural sound philosophy and its implications for *Audio Quality* and *Sound Quality* issues in networked audio systems. All with the goal to achieve fantastic sounding audio systems, in-line with the Yamaha corporate philosophy: *creating Kando together*.

The Yamaha Commercial Audio team.

# Summary

This white paper is structured in 9 chapters, each presenting a quality issue.

Chapter 1 - **Audio Quality** - presents a set of definitions and requirements. To support meaningful discussions on audio quality, the concepts ‘quality’, ‘audio’ and ‘sound’ are defined in detail.

Chapter 2 - **Networked audio systems** - presents a description of a typical networked (and therefore digital) audio system. The described system is modular, supported by networking technologies that have become common practise in the professional audio field.

Chapter 3 - **Performance & Response** - presents the Performance / Response concept - identifying system process parameters and requirements to help assessing the quality of audio systems. Two design philosophies are presented: ‘*natural sound*’ - where the focus lies on preserving the artistic quality of the audio event and offering Response tools to the sound engineer as variable parameters, and ‘*coloured sound*’ where a fixed sound-changing Response is designed into products and systems.

Chapter 4 - **The human auditory system** - briefly presents a description of the human auditory system, including the mechanics of the outer and middle ear, the bio-mechanical coding to the frequency domain by the inner ear, and the transport of the coded firing patterns to the brain through auditory nerves. Using this description, a ‘human audio universe’ is defined to possess three dimensions: level, frequency and time. Also some auditory functions such as localization and masking are presented.

Chapter 5 - **Sampling issues** - presents the audio digitalization (sampling) concept in relation to level, frequency and timing. Dynamic range and frequency range are more or less common concepts, developed to a mature state by the manufacturers of digital audio equipment in the past 25 years. Compared with the 1985 digital (16-bit) technologies, modern 24-bit A/D, D/A and distribution technologies and 32-bit or higher DSP architecture have caused noise floors and distortion levels to move close the boundaries of the audio universe. On timing however, the use of networked audio systems pose new challenges to system designers and sound engineers. This chapter presents the digitalization concept in relation to timing, including latency, jitter and clock phase.

Chapter 6 - **Distribution & DSP issues** - presents a description of the transport and DSP infrastructure in a digital audio system. Transport and DSP architecture - eg. bit depth, fixed/floating point processing - are described to have an effect on a system’s *audio* quality, with only the algorithm (plug-in) design to affect the system’s *sound* quality.

Chapter 7 - **Signal chain level issues** - focuses on audio levels in a system, proposing a ‘0dBFS’ level standard as the optimal design paradigm that allows easy identification of quality problems in a signal chain. Several practical quality issues in system design are presented, such as head amps, gain compensation, clip level mismatch, double pass signal chains. Also, audio compression in speaker processing stage (unbalanced output modes) is discussed, placing the responsibility in the Response (sound quality) domain rather than the Performance (audio quality) domain.

Chapter 8 - **Operational quality** - presents operational quality issues in a networked audio environment, including topology and protocol and their effect on logistics, reliability and redundancy. The use of Ethernet - either as protocol or as embedded service - is posed to be of essential importance to comply with operational quality requirements on design freedom and user interfacing.

Chapter 9 - **Quality assessment methods** - presents methods for subjective and objective quality assessments of audio systems. Conditions for controlled listening tests are proposed for audio quality assessment. Full control over the experiments with careful adjustment of test equipment and environment, and proper statistical analysis are crucial to obtain meaningful results that justify statements on product and system audio quality and Response characteristics.

Appendix 1 - **Subject index & glossary** - lists all topics in this white paper written in *italic* fonts, including all definitions.

Appendix 2 - **Information sources & further reading** - lists information sources and further reading suggestions.

# 1. Audio Quality

The subject of this white paper is audio quality in networked systems. One might think that everybody in the audio industry knows what the words ‘audio’, ‘quality’ and ‘system’ mean. The length of this first chapter proves otherwise; the words can be - and very often are - perceived in different ways by different individuals, often causing discussions about a system’s audio quality and sound quality to end up in endless repetitions of the words ‘is’ and ‘is not’.

## 1.1 Audio

In the field of neurosciences, the human hearing system is named ‘human auditory system’. It’s a bio-mechanical system that converts acoustic audio waves reaching the human ears through the air and the scull’s bone structure into coded neural firing patterns. Auditory nerve strings transfer them to the central auditory nervous system in the human brain, with the latter interpreting the firing patterns to produce a hearing sensation. The hearing sensation is invoked by the heard audio signal, but it is influenced by all other sensations in the brain - from memory, but also from other real-time sensory organs such as vision, smell and touch.

In this white paper, all phenomenon, processes, systems and characteristics pertaining to generating, processing and transporting signals in the audible range of the human auditory system are referred to by the adjective ‘***audio***’.

> *Audio*
>
> *(adjective) designates objects (eg. signals, processes, devices, systems) or characteristics (eg. frequency, level, time) to pertain to signals in the audible range of the human auditory system.*

*figure 101: audio system diagram*

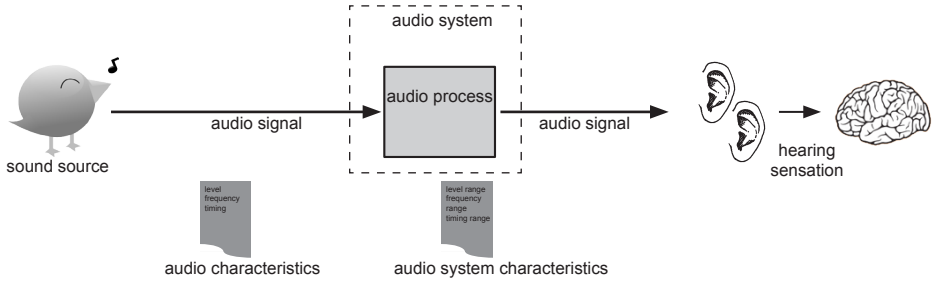



*table 101: audio designation examples*

| term | meaning |
|---|---|
| audio signal | the portion of any time-variant signal that falls in the audible range of the human auditory system, capable of invoking a hearing sensation. The signal can be acoustic, electronic or digital. |
| audio process | generation, transport, change and/or storage of an audio signal. |
| audio system | a system that processes audio signals. |
| audio characteristic | a (physical) feature of an audio signal (eg. level, frequency, time). |
| audio system characteristic | a (physical) feature of an audio system (eg. dynamic range, frequency range, time range). |
| sound source | a human voice, musical instrument or any phenomenon that generates an audio signal. |

## 1.2 Sound

Sound is an adjective, a noun or a verb, used to identify or describe the perceptional characteristic of an audio signal - most often by describing the hearing sensation it invokes. The four other main sensory inputs vision, touch, smell and taste influence the hearing sensation in real time. But the most powerful sound influencer is memory. Already starting in the embryonal phase, the human brain 'learns' how to listen to audio signals, developing preferences for timbres, rhythm, patterns, sound colour, word recognition. As the brain actively controls the bio-mechanic processes in the middle and inner ear, we also literally train our auditory system to be as effective as possible. This means that the way we hear is greatly influenced by our hearing experience - including the developing of preferences for musical styles. Because no individual is the same, the same audio signal will invoke different hearing sensations ('sound') for different individuals.

Where audio characteristics describe physical characteristics of an audio signal, the adjective, noun and verb '*sound*' most often describe perceptional features such as 'warmth', 'transparency', 'definition'. However, sometimes also physical characteristics are used in conjunction with the word 'sound' - eg. 'speed of sound'.

In this white paper we will use 'audio' as *adjective* to pertain to physical characteristics, and 'sound' as an *adjective*, *noun* and *verb* to pertain to perceptional characteristics. If exceptions of the use of the words 'audio' and 'sound' occur, the context will be clarified in the accompanying text.

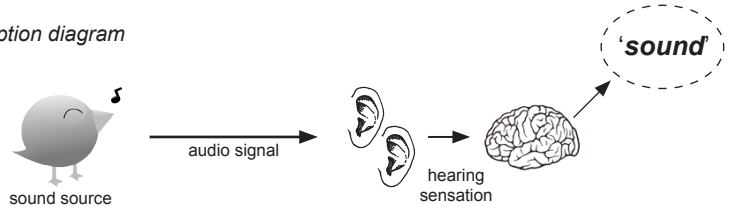
*Sound*

*(adjective, noun, verb) describes the subjective hearing sensation produced by stimulation of the human auditory system of an individual listener by an audio signal, transmitted through the air or other medium.*

*Sound source*

*(noun) designates the origin of an audio signal.*

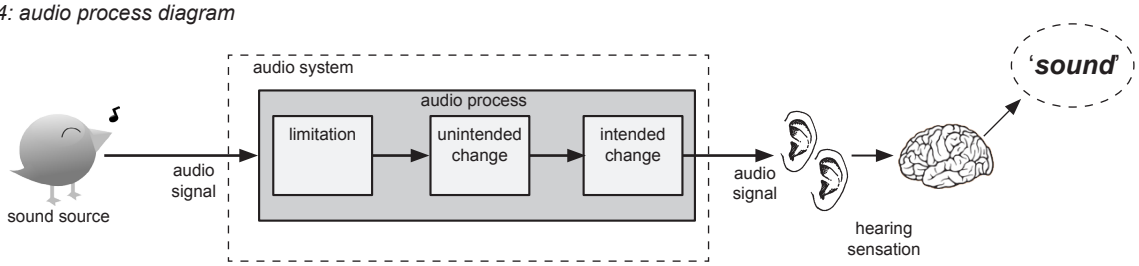
*figure 103: sound perception diagram*



## 1.3 Audio Processes: limitation, unintended change and intended change

An audio system changes the characteristics of an audio signal by applying its audio process. The audio process is divided into three sub-processes: limitation, unintended change and intended change.

*figure 104: audio process diagram*




*Limitation*

*a system's limits in representing signals in level, frequency, and time.*

*Unintended change*

*the change of an audio signal caused by unintended processes in an audio system.*

*Intended change*

*the change of an audio signal caused by intended processes in an audio system.*

An audio system's *limitation* poses physical limits to level, frequency and time. For example, the high-end limitation in an audio system's level range is any incapability to reach 120dB$_{SPL}$ at a listeners position, while an audio system's low-end level limitation often presents itself as a constant level error signal higher than 0dB$_{SPL}$ at the listeners position such as a noise floor. Frequency limitation includes any low and high frequency bandwidth limits within the 20Hz-20kHz frequency range, while timing limitation includes any response time or time coherence incapability of more than 6 microseconds (eg. network latency), or any time coherence problem generating audible level errors (eg. jitter). Chapter 4 presents details on the limits of the human auditory system.

An audio system's *unintended change* poses changes to the audio signal such as equalising, distortion, compression. These changes are not intended by the designers or operators of the audio system - they are included in the audio process because they could not be avoided due to technological, financial, time and/or expertise constraints of the designer and/or operator. Unintended changes can be represented as error signals that are (partly) linear with the audio signal, sometimes summarized by a percentage (eg. %THD) or level ratio (eg. dB gain of a filter). Most commonly, unintended changes are regarded as having a negative impact on sound quality. But in some cases, if a system's initially unintended change is perceived to have a positive effect on sound quality, the product manufacturer or system designer can actively decide to not take countermeasures - thus turning the unintended change into an intended change.

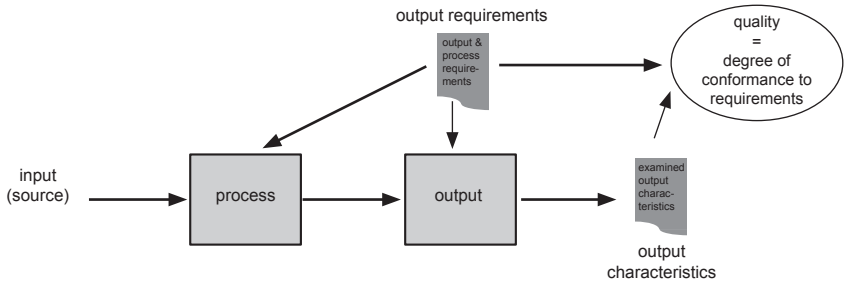
An audio system's *intended change* poses changes to the audio signal by intention of the designer and/or operator of the system - most commonly to improve sound quality to the opinion or expectation of the designer/operator (on behalf of an audience), or to change the sound to match an external context (eg. video postproduction). An intended change can be designed into products and systems by manufacturers and system designers as a fixed process, or offered to system operators (sound engineers) to apply as a variable parameter. More on fixed and variable intended change ('coloured sound vs natural sound') in chapter 3: Performance & Response.

## 1.4 'Quality'

*Quality is conformance to requirements*

This definition of quality comes from the renowned quality management guru Philip B. Crosby[*1A]. His idea is that quality management should focus on setting well defined and realistic requirements - and then design clever management processes to make sure that an organisation's output is meeting up to these requirements.

*figure 105: quality management diagram*




Crosby's definition states that quality is always related to requirements set for the output of the process. To enable the organisation to achieve the desired output quality, process requirements are set. And this area is where quality discussions in many of the debates in the audio industry go wrong: two individuals seldom agree on the requirements of both the process and the output - not even on the definition of the parameters that represent the requirements.

In this white paper, the term 'audio quality' refers to the physical characteristics of an audio signal, the term 'sound quality' refers to the perceptional characteristics of the invoked hearing sensation.

Using Crosby's definition of quality, stating (system) audio quality means stating to what degree the audio signal (system) conforms to set requirements. Audio quality requirements can be stated as physical characteristics, for example in the form of electrical system specifications based on international standards (ISO, AES, IEC). If not otherwise specified,100% accurate representation can be assumed as audio (system) quality requirement.

Stating sound quality means stating to what degree a hearing sensation conforms to an individual listener's requirements - which can be either a preferred hearing sensation, or an expected hearing sensation if the individual is assessing the hearing sensation on behalf of an audience, or for use in an external context. Sound characteristics are often discussed using terminology such as 'warmth', 'transparency', 'definition' - which are not always standardized terms. Assumed that a group of persons agree on the definition of these terms, the degree of conformance will still differ from person to person, depending on individual hearing abilities and preferences.

## 1.5 Audio quality

In this white paper we propose the following requirement for audio signals:

> *Requirement for an Audio signal*
>
> *An examined audio signal should represent the originally generated audio signal accurately, disregarding the intended changes of an audio system.*

If there is no audio system between generating and examining (eg. hearing) an audio signal, the examined signal is exactly the generated signal. The closest we can get is listening to an acoustic signal at very close distance - think millimetres - without any audio disturbances eg. other signals, wind, movement.

In real life there is always a system between the generation and the hearing - even a short distance already constitutes a system as the turbulence in the air between audio source and listener changes the audio signal. Nearby objects or walls, and of course a networked audio system, add further changes.

In this white paper, we propose the following definition for audio quality:

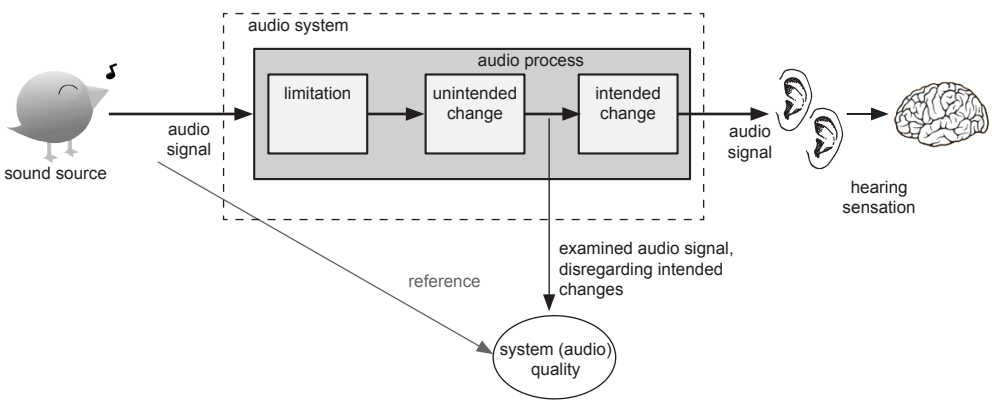
> *Audio quality*
>
> *The degree of representation accuracy of an examined audio signal, disregarding the intended changes of an audio system.*

'Audio quality' describes how accurate an examined audio signal (at the output of a system) resembles the original audio signal generated by the sound source, disregarding the changes applied intentionally by product manufacturers, sound system designers and engineers.

The audio quality of a system between the input audio signal and an examined output signal is called the 'system audio quality'. It can be described using the same requirement as for audio quality: an audio system should accurately transport and process the audio signal - without limitations or unintended changes. In common speech, 'audio quality' is often used to describe a system's audio quality.

> *System audio quality*
>
> *The degree of representation accuracy of an examined audio system, disregarding the intended changes of the audio system*

*figure 106: system (audio) quality diagram*



## 1.6 Sound quality

In this white paper, the term ‘sound quality’ refers to the perceptional characteristics of the hearing sensation invoked by an audio signal. Using Crosby’s definition of quality, stating sound quality means stating in what degree the hearing sensation conforms to what we specified as requirements. And here things start to become tricky: every individual has different requirements.

In this white paper we propose the following requirement for sound:

> *Requirement for sound*
>
> *An audio signal should satisfy either the expected or the preferred hearing sensation of an individual listener.*

For the definition of a system’s sound quality, the sound quality of the original signal has to be considered as well. We will name the sound quality of the original signal ‘*source quality*’. The requirements for the sound source then read as follows:

> *Requirement for a sound source*
>
> *An audio signal generated by a sound source should satisfy either the expected or the preferred hearing sensation of an individual listener without limitation or change by an audio system*

The satisfaction of listening to a sound source - without a system in between ears and source - depends on individual hearing abilities and preferences, and also on the sound characteristics - or the ‘sound’ - of the source. For example, when listening to a solo violin performance, the hearing sensation is influenced by the composition played, the proficiency and virtuosity of the player, the characteristics of the violin. All these parameters together constitute the sound characteristics of the source. Although a statistical average appreciation can be found, for example by assessing the popularity of the solo violin performance by counting the number of persons who bought a concert ticket, every individual will assess source quality in a different way.

Knowing the requirements for the sound source, the sound requirements for the audio system can be defined:

> *Requirement for an audio system’s sound*
>
> *The intended change of an audio signal by an audio system should satisfy either the expected or the preferred change in the hearing sensation of an individual listener with a given source sound.*

In the case of the solo violin performance, the acoustics of the concert hall constitutes an audio system. If the performance needs amplification, then the PA system constitutes an audio system. In both cases, the audio system intentionally changes the audio signal produced by the sound source, contributing positively to the hearing experiences of the audience: the concert hall adds reverberation, the PA system adds loudness.

Sound requirements for sources and systems use perceptional characteristics ‘warmth’, ‘transparency’, ‘definition’. Note that in real life, multiple sound sources as well as multiple audio systems are involved.

In this white paper we propose the following definition for sound quality:

*Sound quality*

*The degree of satisfaction of the expected or the preferred hearing sensation of an individual listener as a result of hearing an audio signal.*

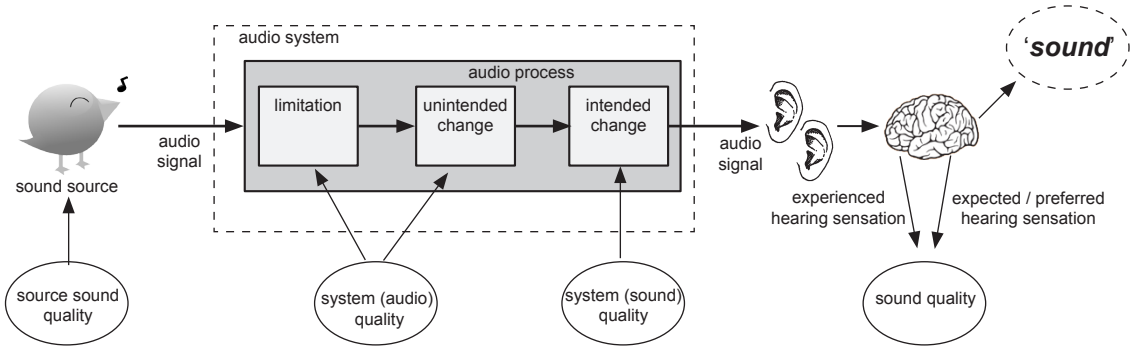
*Source sound quality*

*The degree of satisfaction of the expected or the preferred hearing sensation of an individual listener as a result of hearing an audio signal from a sound source disregarding the limitation or change by an audio system.*

*System sound quality*

*The degree of satisfaction of the expected or preferred hearing sensation of an individual listener as a result of the intended change of an audio signal by an audio system with a given source sound.*

*figure 107: sound quality diagram*




In words: if sound quality and audio quality are summarized as a percentage 'Q' - with 0% as 'minimum quality' and 100% as 'maximum quality', the sound quality of an audio signal experienced by a listener is the product of the source's sound quality, the system's audio quality and the system's sound quality:

*formula 101: experienced sound quality*

$$Q_{experienced/expected\ (sound)} = Q_{source\ (sound)} \times Q_{system\ (audio)} \times Q_{system\ (sound)}$$

## 1.7 Discussing audio quality

The objective of using an audio system to process one or more sound sources is to achieve a better sound quality compared to not using an audio system. The main issue in the minds of product manufacturers, system designers and sound engineers is therefore *sound*. Assuming the sound source as a fixed parameter, the main tools to achieve a better sound quality are the audio system's intended changes - either built into the audio system as fixed characteristics, or available to the sound engineer as variable parameters.

However, a system's sound quality is significantly influenced by the system's audio quality. By definition, the more limitations and unintended changes the system imposes on the processed audio signals, the lower the sound quality will be. This white paper aims to provide insight in audio quality issues in networked systems to allow system designers and sound engineers to achieve the highest possible audio quality, allowing them to appreciate the system's intended changes as a basis for investments or rentals, and apply the available variable intended changes at their will to achieve the best possible sound quality.

Audio quality discussions can be conducted based on physical measurements of the audio signal and audio systems. Basing discussions on listening sessions however brings up the issue of disregarding the intended changes of the audio process - to leave only the limitations and unintended changes to discuss. Disregarding intended changes is easy if an audio system is built according to the '*natural sound*' philosophy - such a system passes audio signals as transparent (natural) as possible, and offers the system's intended changes to the sound engineer as variable components (*colouring options*, eg. equalizers, compressors), including the possibility to switch them off to allow audio quality assessment. Systems designed with a '*coloured sound*' philosophy have fixed intended changes, making it more difficult to assess audio quality issues because the intended changes can never be switched off.

As every system includes some amount of fixed intended changes, most prominently in the loudspeakers, listening session scripts can be used to focus on single parameters when comparing systems - the equivalent of the *ceteris paribus* approach in the economic sciences. If the compared systems all possess the same fixed intended changes, the listener can decide to concentrate on comparing a selected single audio quality parameter. More detailed information on this topic can be found in chapter 9: Quality assessment methods.

To facilitate system audio quality discussions, system audio quality characteristics can be represented by the characteristics of the difference between input and output of a system - the error signal. This error signal can be constant, linear with the level of a signal's frequency components, partially linear or nonlinear:

*table 103: error signal types*

| difference | type | examples |
|---|---|---|
| constant | limitation | audio: *HA noise, A/D quantization noise (unintended)*<br>sound: *masking noise (intended)* |
| linear | change<br>linear with signal level | audio: *jitter noise, equalising (unintended)*<br>sound: *equalising (intended)* |
| non linear | change<br>partially or not linear with<br>signal level | audio: *amplifier clipping, zero-crossing distortion,*<br>*compression (unintended)*<br>sound: *guitar amp distortion, compression (intended)* |

Figure 108A and 108B on the next pages presents a listing of audio quality and sound quality issues in a networked system. In figure 108A (audio quality issues in a networked audio system), a selection of *limitation* and *unintended change* error signals and their causes are presented as grey bars:

> Constant error signals (eg. *limitations*) are shown with one arrow pointing to the average error level in $dB_{FS}$.
>
> Linear error signals (eg. *unintended changes*) are shown with two arrows connected with a dotted line - one arrow pointing to the audio signal level at 0 $dB_{FS}$, the other to the error level to indicate that the error level depends on the signal level.

In figure 108B (sound quality issues in a networked audio system), a selection of available *intended change* processes are presented.

*figure 108A: audio quality issues in a networked system*

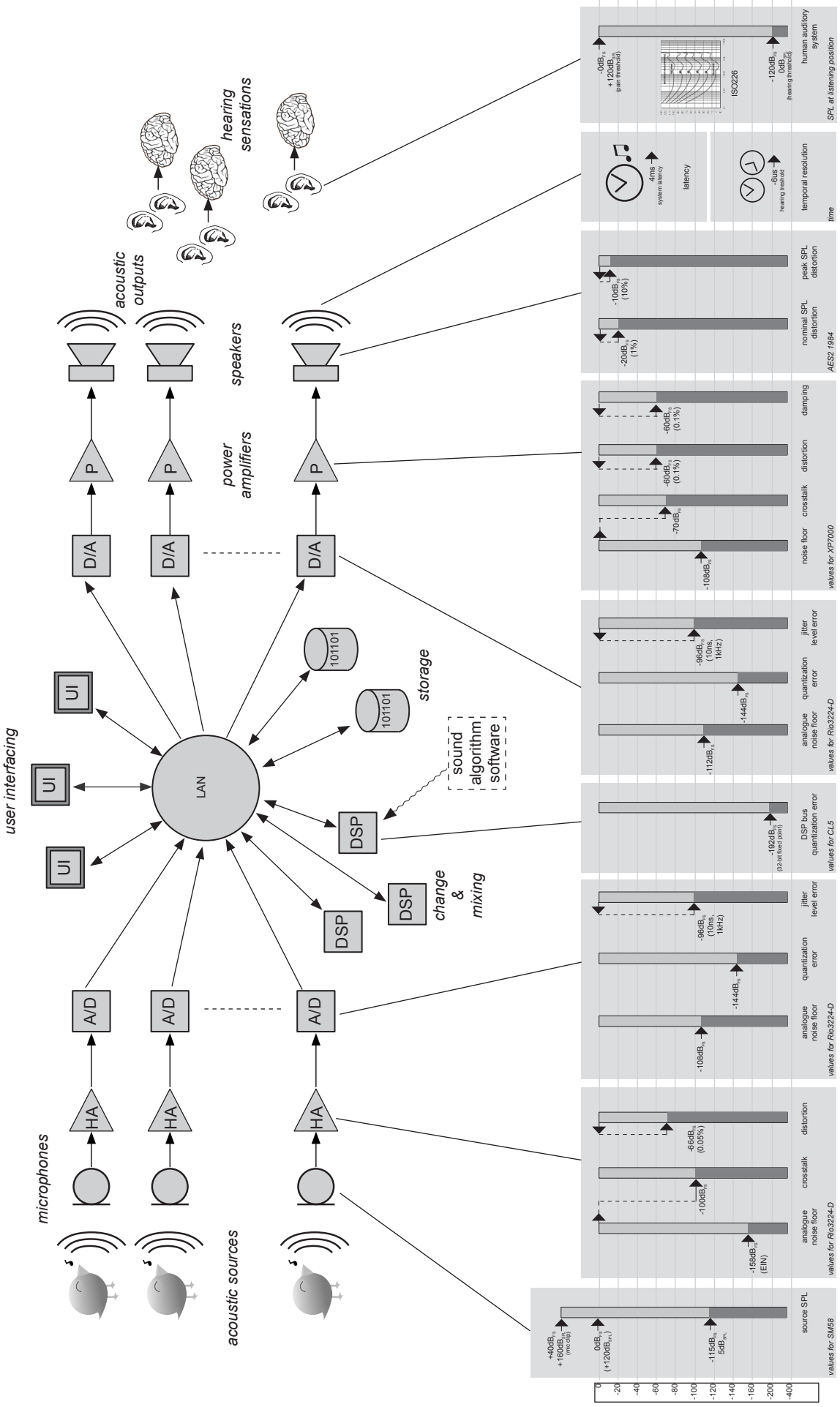
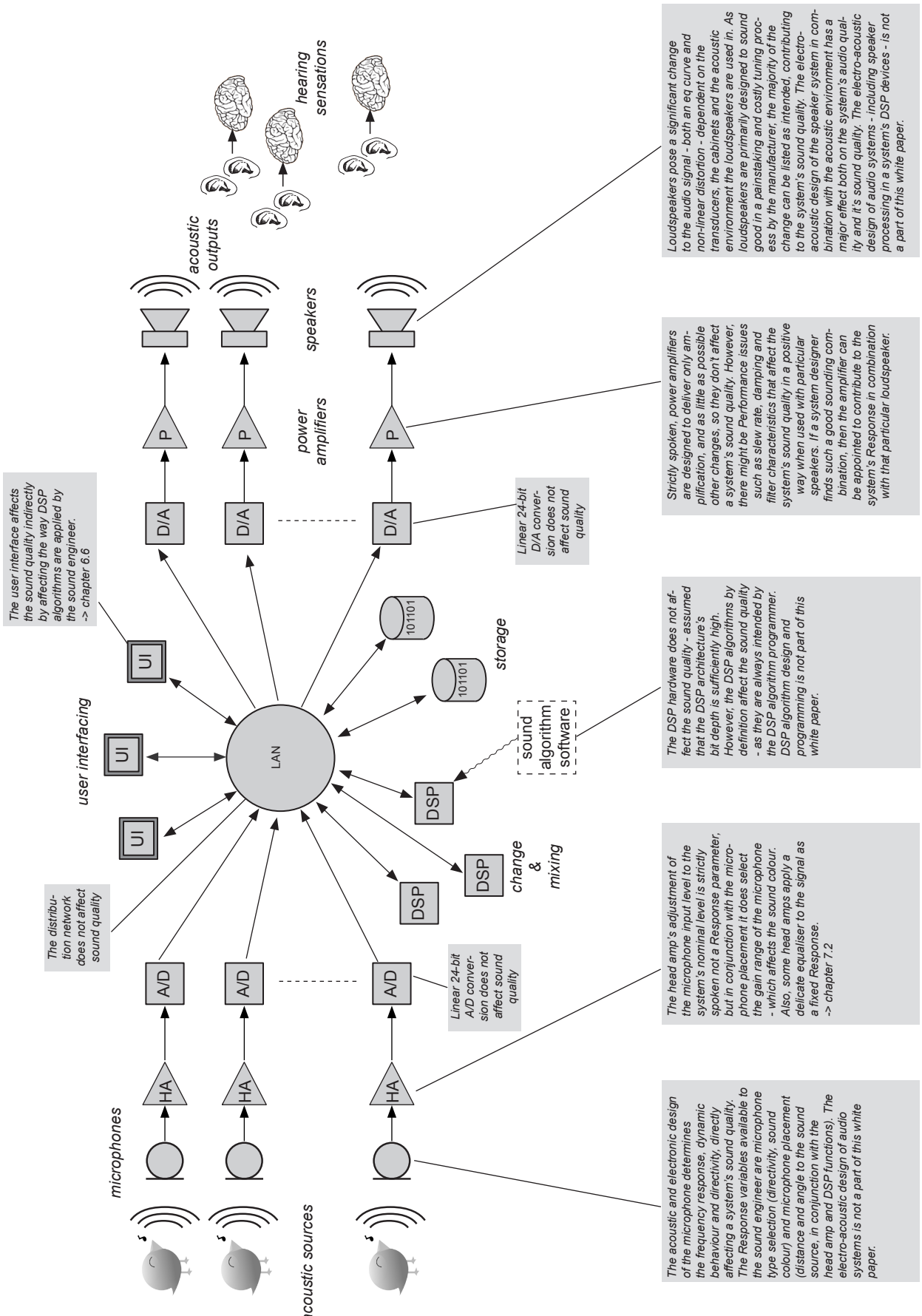


*figure 108B: sound quality issues in a networked system*

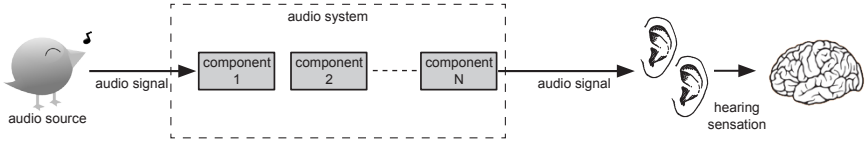


# 2. Networked audio systems

This chapter presents a modular networked audio system as a reference for the rest of this white paper. A collection of 12 modules are introduced as building blocks of a system. The described system processes audio in acoustic, analogue and networked formats.

> *Audio System*
>
> *A collection of components connected together to process audio signals in order to increase a system's sound quality.*

*figure 201: audio system components*




The following pages will elaborate further on audio processes, formats and components.

## 2.1 Audio processes

A system's audio processing can include:

*table 201: audio processing types*

| function | description |
|---|---|
| conversion | format conversion of audio signals |
| transport | transport of signals, eg. through cables |
| storage | storage for editing, transport and playback using audio media, eg. tape, hard disk, CD |
| mixing | mixing multiple inputs to multiple outputs |
| change | equalising, compression, amplification etc |

The audio system can be mechanical - eg. two empty cans with a tensioned string in-between, or a mechanical gramophone player. But since the invention of the carbon microphone in 1887-1888 individually by Edison and Berliner, most audio systems use electrical circuits. Since the early 1980's many parts of audio systems gradually became digital, leaving only head amps, A/D and D/A conversion and power amplification remaining as electronic circuits, and microphones and loudspeakers as electroacoustic components. At this moment, digital point-to-point audio protocols such as AES10 (MADI) are being replaced by network protocols such as Dante, EtherSound.

In this white paper, the terms 'networked audio system' and 'digital audio system' are applied loosely, as many of the concepts presented concern both. When an issue is presented to apply to *networked audio systems*, the issue does not apply to digital audio systems. When an issue is presented to apply to *digital audio systems*, it also applies for networked audio systems.

## 2.2 Audio formats

Although with the introduction of electronic instruments the input can also be an electrical analogue or digitally synthesised signal, in this white paper we will assume all inputs and outputs of an audio system to be acoustic signals. In the field of professional audio, the following identification is used for different formats of audio:

*table 202: audio formats*

| format | description |
|---|---|
| acoustic | audio signals as pressure waves in air |
| analogue | audio signals as a non-discrete electrical voltage |
| digital | audio signals as data (eg. 16 or 24 bit - 44.1, 48, 88.2 or 96kHz) |
| networked | audio data as streaming or switching packets (eg. Ethernet) |

A networked audio system includes these audio formats simultaneously - using specialised components to convert from one to another:

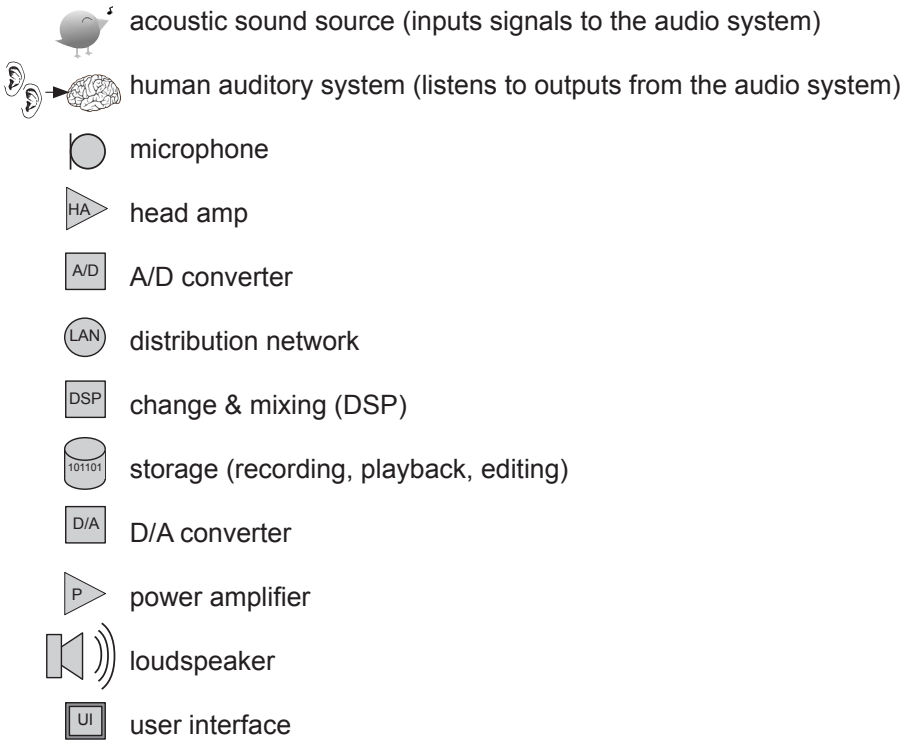
*table 203: audio format conversion components*

| source format | destination format | component |
|---|---|---|
| acoustic -> | analogue | microphone |
| analogue -> | digital/networked | A/D converter |
| digital/networked -> | analogue | D/A converter |
| analogue -> | acoustic | loudspeaker |

## 2.3 Audio system components.

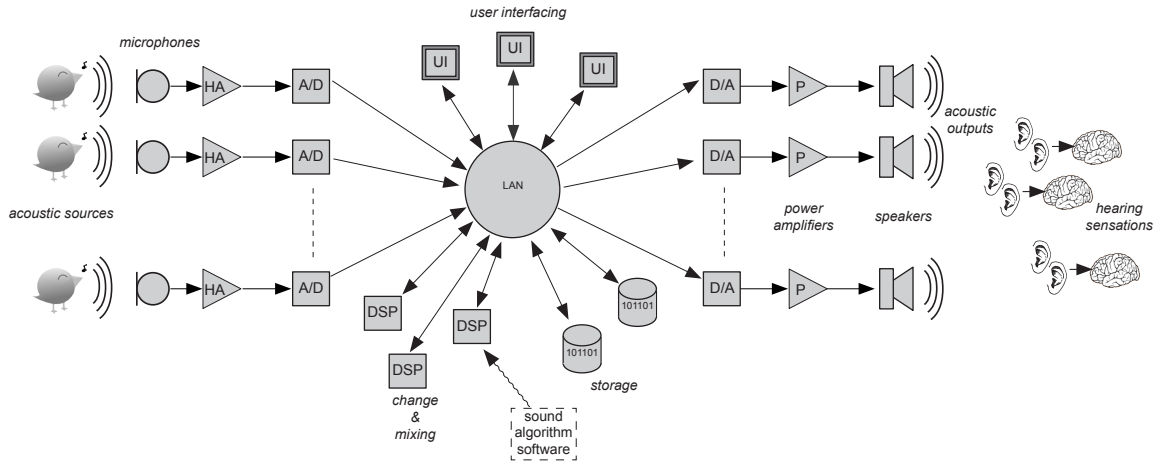
In this white paper we assume an audio system to be modular, using digital signal processing and networked audio and control distribution. An audio system's inputs and outputs are assumed to be acoustic audio signals - with the inputs coming from one or multiple acoustic sound sources, and the output or outputs being picked up by one or more listeners. A selection of functional modules constitutes the audio system in between sound sources and listeners.

*table 204: audio system components*



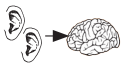
A typical networked audio system is presented in the diagram below. Note that this diagram presents the audio functions as separate functional blocks. Physical audio components can include more than one functional block - eg. a digital mixing console including head amps, A/D and D/A converters, DSP and a user interface. The distribution network in this diagram can be any topology - including ring, star or any combination.

*figure 202: networked audio system example*




Acoustic source 

An acoustic sound source generates vibrations and radiates them to the air. Sound sources can be omni-directional - radiating to all directions, or directional, concentrating energy in one or more directions. Musical instruments use strings (eg. guitar, piano, violin), surfaces (eg. drums, marimba) or wind (eg. flute, trombone) to generate sound. In nature, sound often is generated by wind shearing past objects (eg. trees, buildings). The output of an audio system is also an acoustic sound source. Finally, almost all human activities (including singing) - and man-made machinery (including car engines and bomb detonations) generate sound. The lowest sound pressure level in dB generated by acoustic sound sources closes in to minus infinity - eg. resting bodies at absolute zero temperature . The maximum undistorted sound pressure level is said to be above 160dB$_{spl}$ before vacuum pockets start to form in the air. The lowest frequency an acoustic sound source can generate closes in to zero Hertz ('subsonic'), where the maximum wave pressure frequency in air without distortion is said to be above 1GHz.

Human auditory system 

The human auditory system constitutes the combination of two ears and one brain, creating a hearing sensation invoked by audio signals generated by acoustic sources. The inner-ear codes a level range of appr. 120dB and a frequency range of appr. 20kHz into neural firing patterns, and sends them to a specialised part of the brain called 'auditory nervous system'. The brain interprets the coded signals and invokes a hearing sensation. The hearing sensation is most significantly influenced by changes in level and frequency over time, with the lowest detectable time slot being as low as 6 microseconds. Basic parameters of hearing sensations are loudness, pitch, timbre, localisation.

Microphone 

Microphones convert acoustic signals into electric signals - the analogue domain. Dynamic microphones use a coil and a magnet to generate the electrical signal, condenser microphones reach a higher accuracy using a variable capacitor construction that is much lighter than a coil. Further varieties are Piezo microphones and electromagnetic elements to directly pick up guitar strings.

head amp 


The professional audio market adopted a nominal analogue signal level of $0.775V_{rms}$ as $0dB_u$ reference for line level audio signals, optimally supporting electronic circuit designs with 9V to 15V balanced power supplies used in many audio products. As microphones generally output a much lower signal level - typically around 0.3mV ($-68dB_u$) for the average sound level of conversational speech at 1 meter from the microphone ($60dB_{SPL}$), these signals are amplified to a nominal level before entering further electronic circuits using a microphone preamplifier, or 'head amp', abbreviated HA. Head amps most commonly have an amplification range of around 70dB, and are designed to have a very low noise floor. The most common Equivalent Input Noise (EIN) of a head-amp is $-128dB_u$ (0.3 µVrms), with a maximum input level before clipping of up to $+30dB_u$ (24V). But as the switching, balancing and buffering circuits of the HA block also add noise, the maximum dynamic range a typical HA can deliver to the A/D block is 112 dB. Of course, whenever the HA gain is increased to match a microphone's signal level, the HA noise floor will increase as well, lowering the dynamic range. More details on head-amp quality issues are presented in chapter 7.

A/D converter 


An A/D converter converts electrical (analogue) signals to digital data for further processing in digital audio systems. This process is called 'sampling', with most modern A/D converters using a 24-bit data width to represent audio signals. This allows a theoretical dynamic range of approximately 144dB to be registered accurately, with the inaccuracies in the A/D process accumulating in a digital noise floor at -144dB. Most modern digital audio equipment use 48kHz or 96kHz sampling rates, supporting 20kHz or 40kHz frequency ranges. More details on sampling are presented in chapter 5.

distribution network 


A distribution network is a collection of components used to transfer data from and to all physical locations in the audio system. The distribution of course includes audio, but it can also include data to control audio components, and other media data such as video and lighting control. A distribution network can consist of multiple point-to-point connections, separately for audio, control and other data. Such a network needs hardware routers or patch panels at every location to patch sources and destinations. This is not only expensive, but it also limits design freedom as functional connections are restricted by physical connections - and for every change in a system's functional design, the physical design has to change with it. Also, distribution systems based on point-to-point connections have very limited redundancy options. This is why networked systems have become a standard for audio distribution systems - supporting the functional and physical designs to be fully independent and also fully redundant. The audio protocol can be based on Ethernet, or it can include an embedded Ethernet tunnel. As most control systems use Ethernet, and protocol converters are available for other protocols (eg. USB, MIDI, RS232), the use of Ethernet allows virtually any digital data format to be transported over the same distribution network. If the audio system is Ethernet based - using Dante, EtherSound and/or CobraNet, the distribution network will typically be a collection of Ethernet *switches* and cables. More details on operational (non-audio) quality issues in networks is presented in chapter 8.

change & mixing (DSP) 


Digital Signal Processors are used to perform real-time change and mixing of audio signals. Some LSI manufacturers, including Yamaha, Analog Devices, Motorola and Texas Instruments, offer dedicated DSP hardware architectures. Combined with general purpose Field Programmable Gate Arrays (FPGA) chips, the processing power of digital systems has evolved to a level way beyond the capabilities of previously used analogue systems. High data widths - eg. 32 bit or higher - ensure that error residuals of DSP calculations stay well under the head-amp and A/D converter's noise floors, leaving algorithm design and the user interfacing as main quality parameters for DSP functionality.
In the past, dedicated DSP was normally built into mixing consoles, effect units or speaker processors. But since networks started to support high channel counts, DSP units - including 'plug-in servers', 'mixing engines', effect units, speaker processing and user-programmable DSP units - can be located anywhere in the system in any quantity. More details on DSP quality issues are presented in chapter 6.

storage (recording, playback, editing) 


A digital audio system can process audio in real time, but it also can store audio streams on media such as hard disks, memory cards, CD, DVD for later processing or playback. Through storage, an audio process can flow through multiple audio systems at different time slots - eg. a multitrack live recording being stored on a hard disk, then edited on a second system to an authoring DVD, then mixed down on a third system to CD, then transferred to a customer by post and then played back on a fourth system: the stereo system at the customer's home. Multitrack recording, editing and authoring is most commonly done with Digital Audio Workstation (DAW) software running on Personal Computers - using Ethernet connectivity to connect to networked audio systems.

D/A converter 


D/A converters convert digital audio data to electrical (analogue) signals to be sent to power amplifiers, accepting the same data width and rate as the A/D converters and the distribution network of the audio system.

power amplifier 


A power amplifier increases an audio signal voltage to a higher level at a low impedance to drive energy into loudspeakers. Modern power amplifiers use high frequency switching output stages to directly drive loudspeakers (class-D), sometimes combined with AB class circuits (class TD, EEEngine*2A). Some power amplifiers have distribution interfaces, DSP (for speaker processing) and D/A converters built-in.

loudspeaker 

Loudspeakers convert electric signals into acoustic signals. High quality loudspeaker systems use multiple transducers to generate a combined acoustic output, each delivering a separate frequency range. Multiple time-aligned transducers - 'line arrays' - can be used to generate coupled acoustic coverage. High frequency transducers (tweeters, compression drivers, ribbon drivers) are available in sizes varying from 0.5" to 3", mid frequency transducers from 5" to 15", and low frequency transducers ('woofers, sub woofers') from 8" to 21". Loudspeakers and individual transducers have an efficiency (sensitivity) and a maximum SPL output (peak SPL), standardized through the AES1984 norm. In a networked audio system, the loudspeakers are the most prominent sources of distortion - depending on the build quality of the transducer, but also the enclosure. Fortunately, the kind of harmonic distortion generated in loudspeakers often positively contributes to sound quality.

User interface 


To allow sound engineers to operate audio systems, manufacturers of components provide some form of user interface. Conventional (mostly analogue) audio components use hardware 'tactile' user interfaces such as knobs and faders as an integral part of the analogue electronic circuitry. The use of digital technology introduced remote and graphic interfaces such as mouse/trackpad, display and touch screens, while the introduction of networking technology allowed multiple user interfaces to coexist in one system, sharing physical connections through the network protocol, and also functionality through common control protocols. Examples are the many available online graphic user interfaces on personal computers and tablets for digital mixing consoles.

# 3. Performance and Response

## 3.1 Unintended and Intended changes

The definitions and requirements for audio, sound and quality described in chapter 1 present two classes of processes in an audio system: those that unintentionally decrease the *audio* quality of the system (and with it the sound quality), and those that intentionally increase the *sound* quality of the system.

An ideal audio system will pass any signal in the audible range of the human auditory system in full, without limitation or unintentional change, offering fixed and variable processes that intentionally change the audio signal in order to increase the system's sound quality. In real life however, audio systems always limit and unintentionally change signals within the audio universe - which means that the limitation and unintended changes can be heard by individuals listening to the system. By definition, limitation and unintended changes of audio signals decrease the sound quality because if such a limitation or change could be avoided, the product manufacturer / system designer would have done so. Also, if a limitation or unintended change would be found to increase the sound quality, we assume that the product manufacturer / system designer would redefine it as an intended change.

Intended changes posed to audio signals by a system's processes by definition increase the sound quality - because if they would not increase the sound quality, the product or system designer would not have applied them in the system.

This arrangement seems logical, but it poses a philosophical problem: 'Sound quality' is a subjective parameter - every individual will assess sound quality differently. Therefore, the decision to designate a system process as intended or unintended change is subjective - different listeners might designate processes - eg. a signal path with a fixed EQ curve or a compression - differently: one listener might prefer the process, and the other might not.

To make the designation of processes in an audio system 'semi-objective', or at least properly defined, we propose to view the matter strictly from a manufacturer / system designer's perspective, and define the decision method to designate system processes as unintended or intended change as follows:

*audio system processes designated as unintended change*

*[] all fixed processes that are listed by the product manufacturer / system designer without reference to a positive contribution to the product / system's sound quality*
*[] all fixed processes not mentioned in the product manufacturer / system designer's promotion or specification sheets.*

*audio system processes designated as intended change*

*[] all variable processes*
*[] all fixed processes that are promoted by the product manufacturer / system designer in promotion or specification sheets as a positive contribution to the system's sound quality*

Of course individual listeners can disagree with the decisions made by product manufacturers and system designers, differing in their preferences for individual processes in the audio system. Even when preferences are averaged, audiences for different genres of music (eg. pop, classical, jazz) might have different average preferences for an audio system's process. To accommodate these differences, there are two design philosophies that product manufacturers and system designers can apply: '*natural sound*' and '*coloured sound*'. To define these philosophies, first the concept of *Performance* and *Response* is presented.

## 3.2 Performance and Response

To identify quality issues in audio systems, two concepts are proposed to designate limitation and unintentional change as affecting a system's audio quality, and fixed and variable intended change to affect a system's sound quality:

*Performance*

*A collection of system processes that **limit** and **unintentionally** change audio signals, in reference to an ideal audio system, representing how accurate the system passes audio*

*Response*

*A collection of fixed and variable system processes that **intentionally** change audio signals, posing a positive contribution to the system's sound quality.*

Summarizing: a hearing sensation is invoked when an audio signal - generated by a ***sound source*** - reaches a listener through an audio system. The ***sound quality*** experienced by a listener depends on the ***source sound quality*** of the generated audio signal, limited and unintentionally changed by the ***Performance*** and intentionally changed by the ***Response*** of the audio system. This arrangement is equivalent to the formula presented in chapter 1.6.

*figure 301: Performance & Response processes in an audio system*

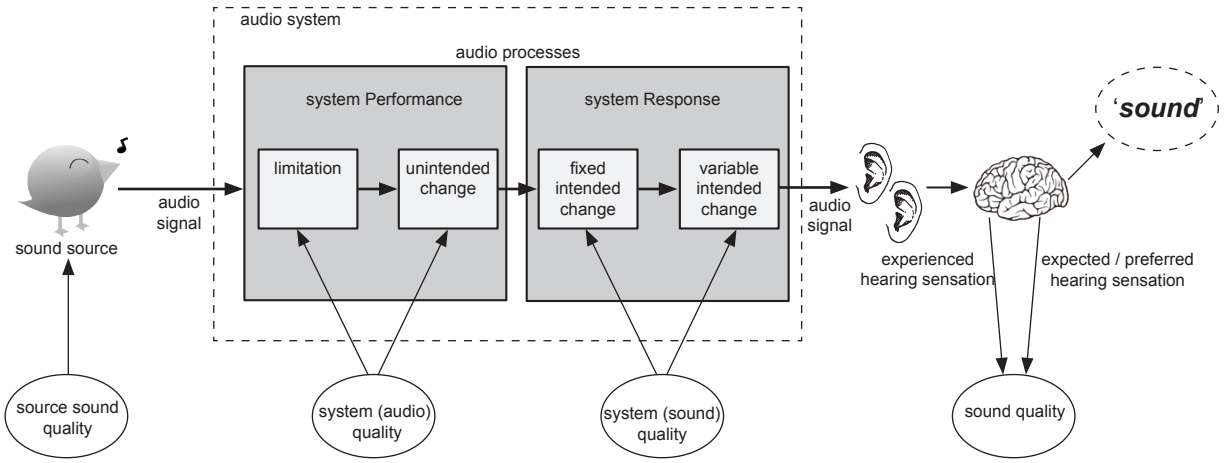


*table 301: examples of Performance and Response processes in audio systems*

*Performance processes*

| | |
|---|---|
| limitations | noise floor in a signal chain |
| | inability of a speaker system to reach 120dB at a listeners position |
| | frequency bandwidth limitation of a speaker system |
| unintended changes (always fixed) | distortion in a signal chain |
| | damping factor in amp/speaker combinations |
| | jitter level errors |

*Response processes*

| | |
|---|---|
| intended fixed changes | EQ curve of a head amp, promoted as 'warm sound' |
| | passive crossover of a loudspeaker cabinet |
| | fixed compressor and equalising in a speaker processor |
| intended variable changes | a mixing console's channel strip with gain, delay, EQ, compression |
| | plug-in and outboard effects |
| | graphic equalizers in speaker system processors |

## 3.3 Natural Sound and Coloured Sound

Audiences normally have no knowledge of the audio system in between the sound sources and their ears. For their assessment of sound quality - the degree of satisfaction of their individual hearing experience - they see the sound source and the audio system as one entity. For performances stored on media carriers (eg. CD, DVD, USB stick, hard disk), the play-out ('HiFi') system is recognised by consumers as a seperate system affecting sound quality, but the system used for producing the content (eg. music studio equipment) is not recognised separately.

Normally only 'audio professionals' distinguish the full audio system from the sound source(s) as a separate entity affecting audio quality. Audio professionals include product manufacturers engineers, system designers and sound engineers operating audio systems.

Most audio professionals will agree that the Performance of a system should be as high as possible, offering low noise floors, low distortion, high output, high bandwidth etc.

On the Response of systems however, the opinion of audio professionals differ. For products specially designed for a genre of applications, manufacturers and system designers sometimes apply fixed Response processes in systems because they are generally required for that particular application genre - eg. 'warm' sounding head amps for pop music. For products and systems designed to serve a variety of applications, the Response processes are offered as variable parameters, transferring the decision to use them - and in what degree - to the sound engineer operating the audio system.

With the increasing complexity of projects, and also the increased focus of investors and artists on a system's sound quality and the sound engineer's creativity, there is an increasing demand for a systems with a *natural sound* default response - maximally respecting the 'natural' sound characteristics of the sound source through a high system Performance to retain the sonic qualities of the audio signal as it was generated by the sound source as much as possible. All system Response processes are then available as *colouring tools* - allowing sound engineers to freely 'shape' the system's sound to suit the project's creative goals.

In general, for manufacturers and system designers it is more costly to offer variable Response processes than fixed Response processes because of the extra connectivity and user interface facilities required for variable processes. If the 'warm sounding pre-amp' is offered to the sound engineer as a variable process, it needs additional switching and control circuitry (either analogue or digital), and a user interface (eg. knobs or a touch screen GUI) to control them - which is more costly than offering the process in a fixed form. With the increased processing power of digital mixing consoles, many Response processes are now offered to the sound engineer as 'plug-in' units, available to be inserted in any signal chain in the system. In general, digital (networked) systems offer much more variability of the process parameters compared to analogue systems, but there are still many differences in variability between digital systems as a result of design philosophy and application genre target.

Using the Performance and Response concept, sound systems can be categorized as 'coloured sound' and 'natural sound' systems. In the below figures, the amount of fixed and variable processing is represented by the size of the Response process blocks:

*figure 302: Performance & Response in a 'Coloured Sound' system*

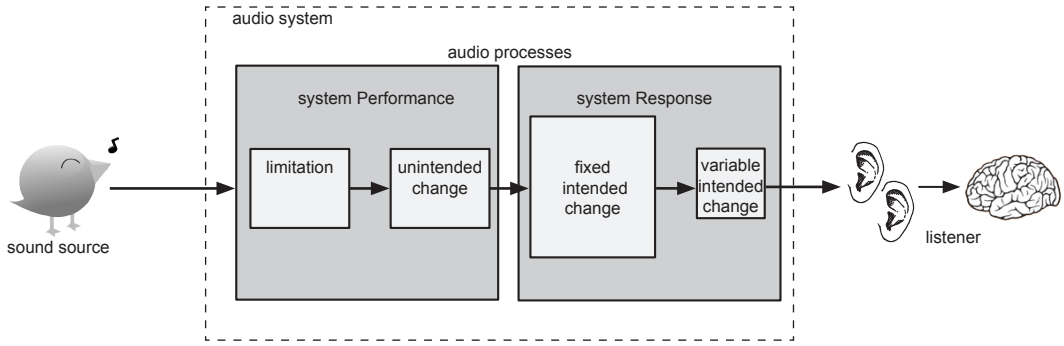
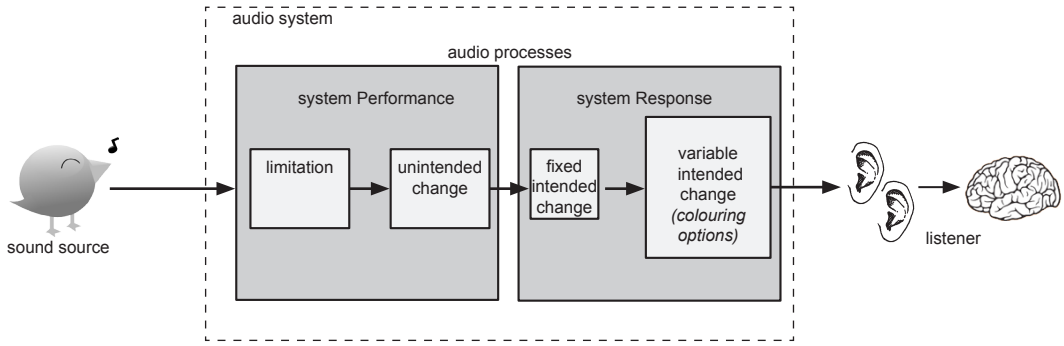



*figure 303: Performance & Response in a 'Natural Sound' system with colouring options*




*Coloured Sound* systems are easy to operate as they already provide a default Response matching the application genre. Coloured Sound systems can be set up quickly because many Response processes have been set already by the product manufacturer or system designer. However, the fixed processes are designed to support a particular application genre. The downside is that if a fixed process mismatches with the application, it can not be turned off - the sound engineer and his audience are stuck with it.

*Natural Sound* systems can be applied for any genre, giving the sound engineer more control over processes to influence sound quality in more detail compared to coloured sound systems. All variable intended processes are available as colouring tools. The downside is that the sound engineer has to do more work to control all variable Response processes. Also, the sound engineer has to have additional knowledge and experience to be able to control these processes properly.

Both design philosophies can offer the same Performance (audio quality) and Response (sound quality), the choice for an investment or hire depends on the requirements of the application genres, time constraints, and the abilities of the available sound engineers to operate the audio system.

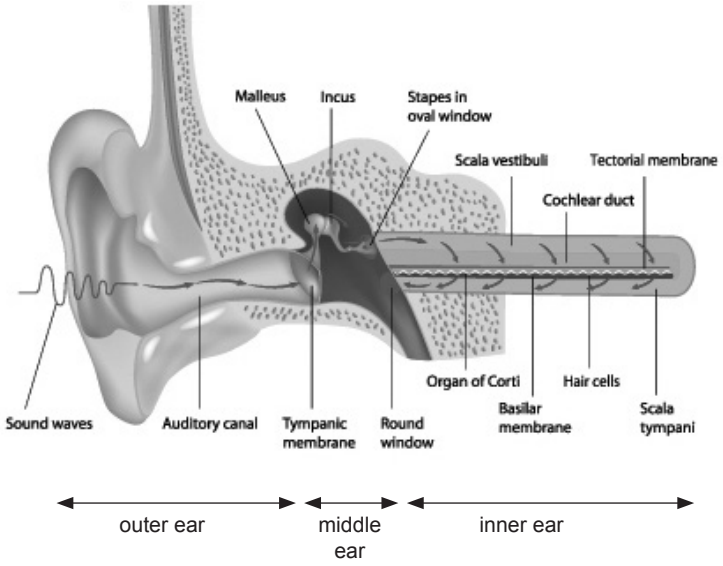
# 4. The human auditory system

The human auditory system is a real wonder - with two compact auditive sensor organs - the ears - at either side of the head, connected with a string of high speed nerve fibres to the brain stem. The brain stem is the central connecting point of the brain - connecting the human body's nervous system inputs (such as audio, vision) and outputs (such as muscle control) to the other areas of the brain. The brain stem redirects the audio information from the ears to a section of the brain that is specialised in audio processing: the *auditory cortex*.

This chapter consists of three parts: the *ear anatomy* - describing in particular the inner ear structure, *the audio universe* - presenting the limitations of the human auditory system in three dimensions, and *auditory functions* - describing how our auditory cortex interprets the audio information coming from the ears. The ear anatomy and the auditory functions are presented using very simplified models - only describing the rough basics of the human auditory system. For more detailed and accurate information, a list of information sources for further reading is suggested in appendix 2.

## 4.1 Ear anatomy

*figure 401: anatomy of the human ear* [*4A]




The ear can be seen as three parts: the outer ear, the middle ear and the inner ear - each with a dedicated function.

The outer ear consists of the ear shell (*pinna*) and the auditory canal. Its function is to guide air pressure waves to the middle ear- with the ear shell increasing the sensitivity of the ear to the front side of the head, supporting front/rear localisation of audio signals.

The middle ear consists of the ear drum (*tympanic membrane*), attached to the inner ear through a delicate bone structure (*malleus, incus and stapes*). The middle ear bones (*ossicles*) and the muscles keeping them in place are the smallest in the human body. One of the major functions of the middle ear is to ensure the efficient transfer of sound from the air to the fluids in the inner ear. If the sound were to impinge directly onto the inner ear, most of it would simply be reflected back because acoustical impedance of the air is different from that of the fluids. The middle ear acts as an impedance-matching device that improves sound transmission, reduces the amount of reflect sound and protects the inner ear from excessive sound pressure levels. This protection is actively regulated by the brain using the middle ear's muscles to tense and un-tense the bone structure with a reaction speed as fast as 10 milliseconds. The middle ear's connection to the inner ear uses the smallest bone in the human body: the stapes (or stirrup bone), approximately 3 millimetres long, weighing approximately 3 milligrams.

The inner ear consists of the *cochlea* - basically a rolled-up tube. The middle ear's stapes connects to the cochlea's 'oval window'. The rolled-up tube contains a tuned membrane populated with approximately 15,500 hair cells and is dedicated to hearing. The structure also has a set of semi-circular canals that is dedicated to the sense of balance. Although the semi-circular canal system also uses hair cells to provide the brain with body balance information, it has nothing to do with audio.

For audio professionals, the cochlea is one of the most amazing organs in the human body, as it basically constitutes a very sensitive 3,500-band frequency analyser with digital outputs - more or less equivalent to a high resolution FFT analyzer.

Unrolling and stretching-out the cochlea would give a thin tube of about 3.4 centimetres length, with three cavities (scala vestibuli, and scala timpani, filled with perilymph fluid, and scala media, filled with endolymph fluid), separated by two membranes: the Reissner's membrane and the basilar membrane. The basilar membrane has a crucial function: it is thin and stiff at the beginning, and wide and sloppy at the end - populated with approximately 3,500 sections of 4 hair cells.

*figure 402: side view of a stretched out cochlea*

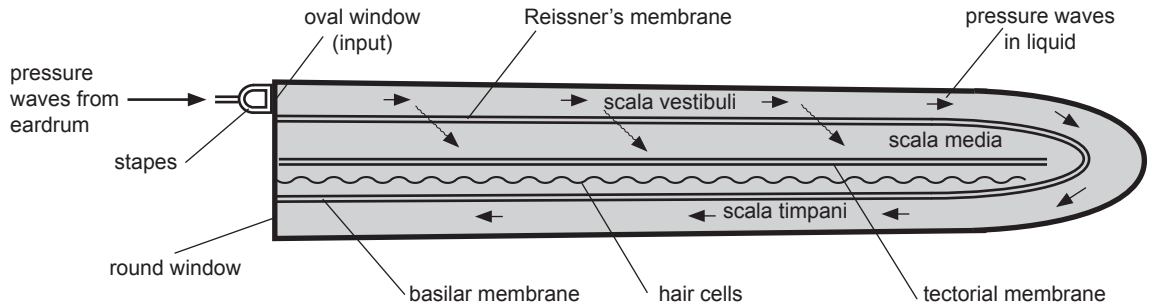
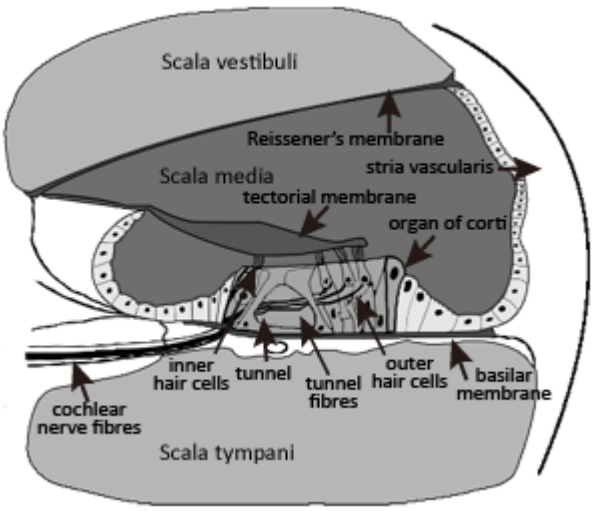
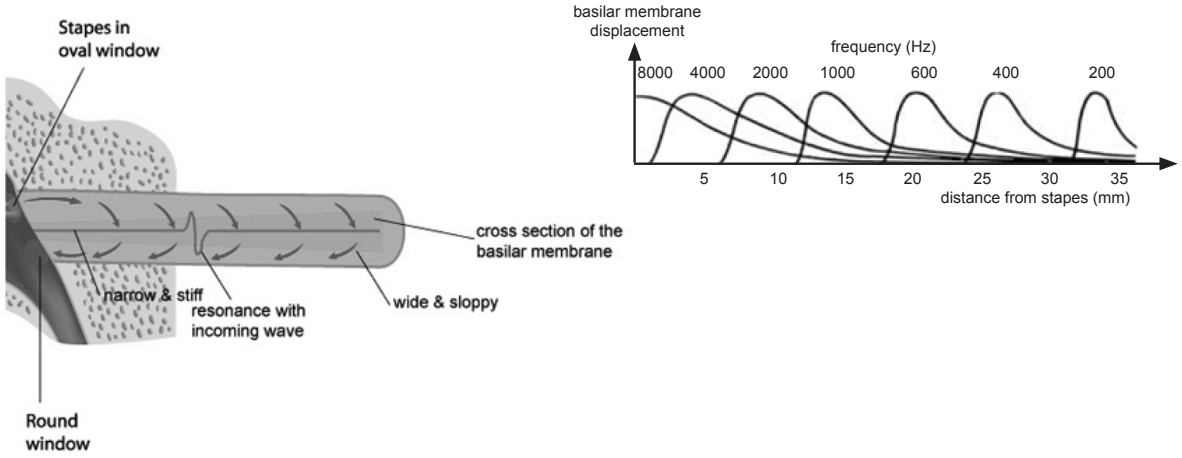



*figure 403: cross section of the cochlea tube* [*4B]



Incoming pressure waves - delivered to the cochlear fluid by the stapes - cause mechanical resonances on different locations along the basilar membrane, tuned to 20 kHz at the beginning near the oval window - where the membrane is narrow and stiff, down to 20 Hz at the end where the membrane is wide and sloppy.

*figure 404: resonance distribution of the basilar membrane*

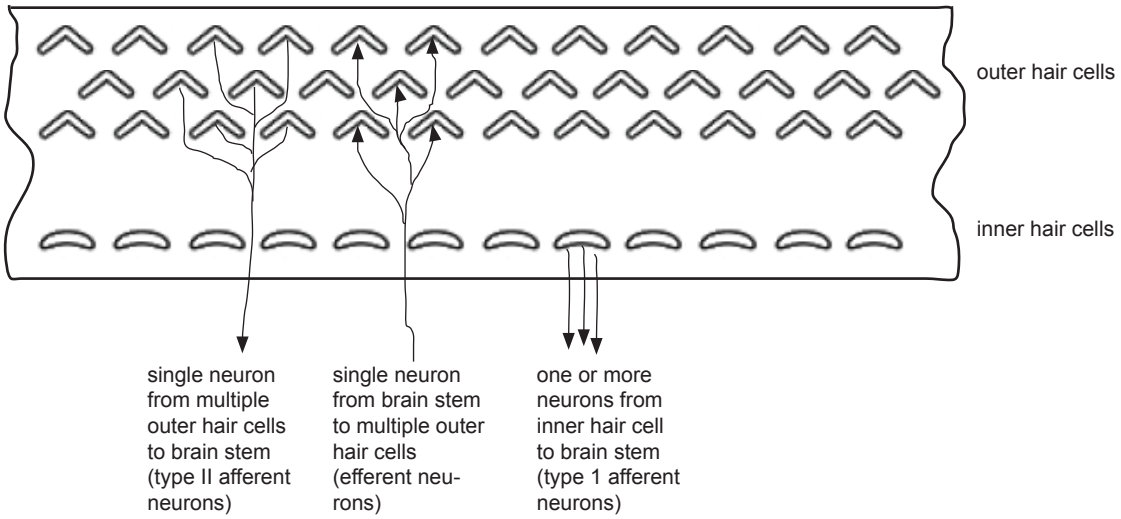



A row of approximately 3,500 *inner hair cells* (IHC's) are situated along the basilar membrane, picking up the resonances generated by the incoming waves. The inner hair cells are spread out exponentially over the 3.4 centimetre length of the tube - with many more hair cells at the beginning (high frequencies) than at the end (low frequencies). Each inner hair cell picks up the vibrations of the membrane at a particular point - thus tuned to a particular frequency. The 'highest' hair cell is at 20 kHz, the 'lowest' at 20 Hz - with a very steep tuning curve at high frequencies, rejecting any frequency above 20 kHz. (more on hair cells on the next page)

Roughly in parallel with the row of 3,500 inner hair cells, three rows of *outer hair cells* (OHC's) are situated along the same membrane. The main function of the inner hair cells is to pick up the membrane's vibrations (like a microphone). The main function of the outer hair cells is to feed back mechanical energy to the membrane in order to amplify the resonance peaks, actively increasing the system's sensitivity by up to 60dB[*4C].

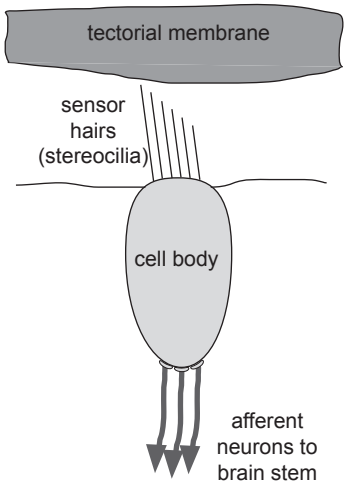
Hair cells are connected to the brain's central connection point - the brain stem - with a nerve string containing approximately 30,000 neurons (*axons*)[*4D]. Neurons that transport information from a hair cell to the brain stem are called afferent neurons (or sensory neurons). Neurons that transport information from the brain stem back to (outer) hair cells are called efferent neurons (or motor neurons). Afferent and efferent connections to outer hair cells use a one-to-many topology, connecting many hair cells to the brain stem with one neuron. Afferent connections to inner hair cells (the 'microphones') use a many-to-one topology for hair cells tuned to high frequencies, connecting one hair cell to the brain stem with many neurons.

*figure 405: inner and outer hair cell distribution and neurons arrangement on the basilar membrane*



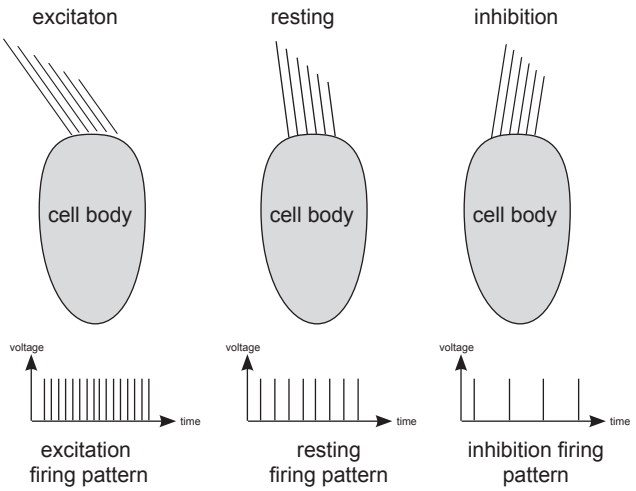
Zooming in to the hair cell brings us deeper in the field of neurosciences. Inner hair cells have approximately 40 hairs (*stereocilia*) arranged in a U shape, while outer hair cells have approximately 150 hairs arranged in a V or W shape[*4E] . The hair cells hairs float free in the cochlea liquid (endolymph) just below the tectorial membrane hovering in the liquid above them, with the tips of the largest outer hair cells' hairs just touching it.

*figure 406: inner hair cell detail*



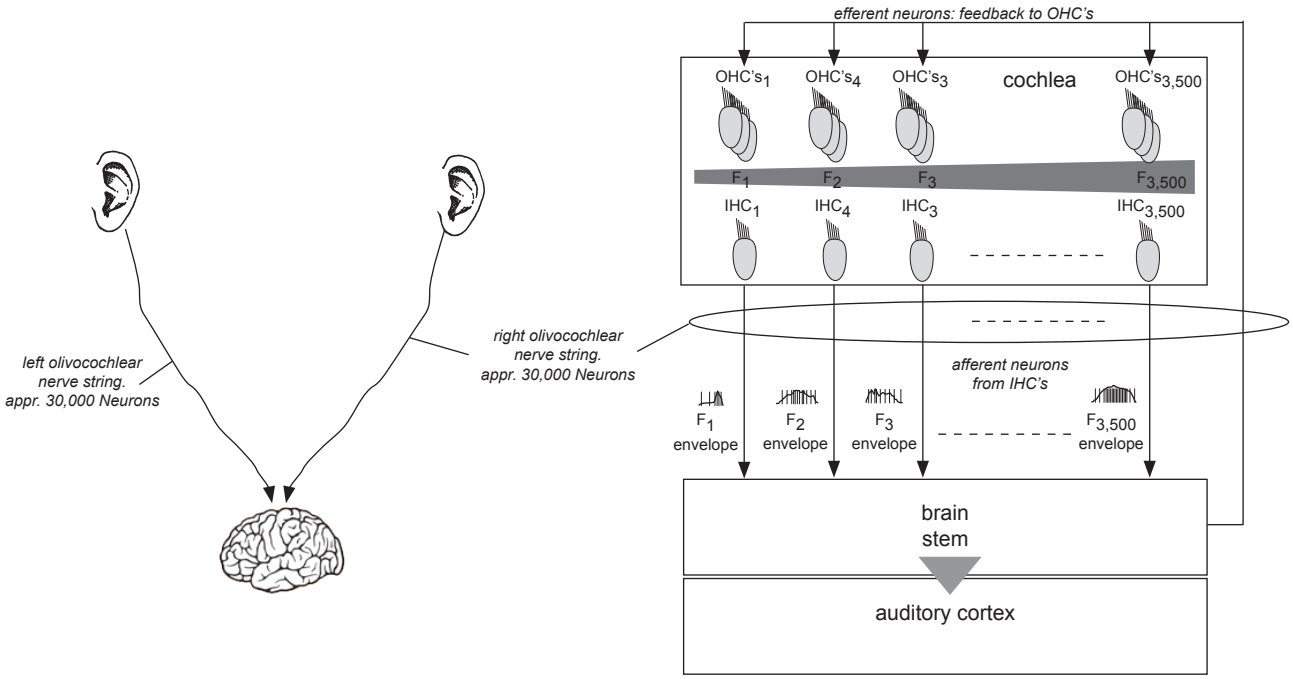

When the basilar membrane below the hair cell moves, the hairs brush against the tectorial membrane and bend - causing a bio-electrical process in the hair cell. This process causes neural nerve impulses (*action potentials*, a voltage peak of approximately 150 millivolts) to be emitted through the afferent neuron(s) that connect the hair cell to the brain stem. The nerve impulse density (the amount of nerve impulses per second) depends on how much the hairs are agitated. When there's no vibration on a particular position on the basilar membrane, the corresponding hair cell transmits a certain amount of nerve impulses per second. When the vibration of the basilar membrane causes the hairs to bend back (excitation) and forth (inhibition), the density of nerve impulses will increase and decrease depending on the amplitude of the vibration.

*figure 407: inner hair cell response to bending of the hairs*




The maximum firing rate that can be transported by the neurons attached to the hair cell is reported to be up to 600 nerve impulses (*spikes*) per second. The nerve string from the cochlea to the brain contains approximately 30,000 auditory nerve fibres - of which more than 95% are thought to be afferent - carrying information from hair cells to the brain stem[*4F]. The brainstem thus receives up to 18 million nerve impulses per second, with the spike density carrying information about the amplitude of each individual hair cell's frequency band, and the spike timing pattern carrying information about time and phase coherency. As the human auditory system has two ears, it is the brain's task to interpret two of these information streams in real time.

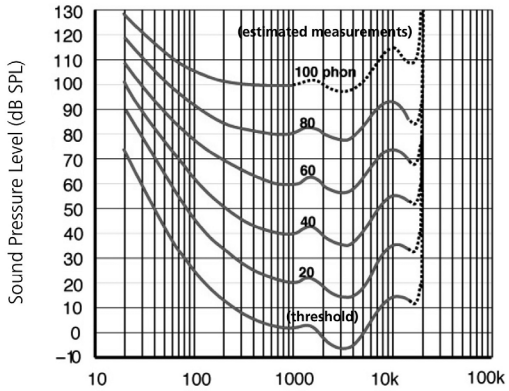
*figure 408: coding of frequency domain envelopes by firing patterns from cochlea to the brain stem (very simplified model)*



## 4.2 The audio universe

The sensitivity of the human auditory system has been measured for individual frequencies in many research projects, accumulating in the ISO226:2003 'loudness contour' standard[*4G]. This sensitivity measurement includes the outer/middle/inner ear and the route from the cochlea via the auditive cortex to the higher brain functions that allow us to report the heard signal to the researcher. The ISO226:2003 graph shows equal perceived loudness ('phon') in the frequency domain for different sound pressure levels for the average human - the actual values can vary by several dB's. The lowest line in the graph is the hearing threshold - with the approximate sound pressure level of 20 micro Pascal at 1kHz to be used as the SPL reference point of $0dB_{SPL}$.

*figure 409: ISO226:2003 loudness contour*




The ISO226 contour graph shows a maximum sensitivity around 3 kHz - covering the most important frequencies used in human speech. Above 20 kHz the basilar membrane doesn't resonate, and there are no hair cells to pick up any energy - limiting human hearing to 20 kHz with a very steep slope. Below 20 Hz, hair cells at the end of the basilar membrane can still pick up energy - but the sensitivity is very low.

At a certain sound pressure level, the hair cells become agitated above their maximum range, causing pain sensations. If hair cells are exposed to high sound pressure levels for long times, or to extremely high sound pressure level for a short-period (so the middle ear's sensitivity can not be adjusted in time), hairs will damage or even break off. This causes inabilities to hear certain frequencies - often in the most sensitive range around 3 kHz. With increasing age, hair cells at the high-end of the audio spectrum - having endured the most energy exposure because they are at the beginning of the basilar membrane - will die, causing age-related high frequency hearing loss. Sometimes, when hair cells are damaged by excessive sound pressure levels, the disturbed feedback system causes energy detection even without any audio signal - often a single band of noise at constant volume (tinitis)[*4H].

Damaged hair cells can not grow back, so it is very important to protect the ears from excessive sound pressure levels. The European Parliament health and safety directive 2003/10/EC and the ISO 1999:1990 standard state an exposure limit of $140dB_{SPL}$(A) peak level exposure and a maximum of $87dB_{SPL}$(A) for a daily 8 hours average exposure[*4I].

In many research projects, a pain sensation is reported around $120dB_{SPL}$ exposure, to be almost constant along the frequency spectrum[*4J]. Although sound quality requirements differ from individual to individual, in this white paper we will assume that 'not inflicting pain' is a general requirement for audio signals shared by the majority of sound engineers and audiences. Therefore we will arbitrarily set the upper limit of sound pressure level exposure at $120dB_{SPL}$. (note that this is the SPL at the listeners position, not the SPL at 1 meter from a loudspeaker's cone - this level needs to be much higher to deliver the required SPL over long distances).

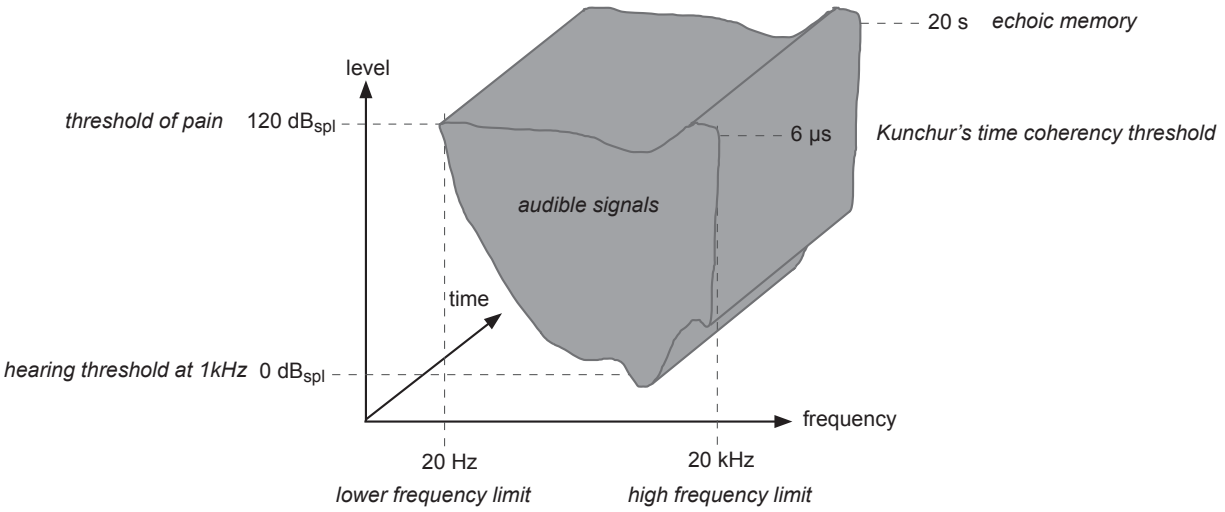
For continuous audio signals, the described level and frequency limits apply in full. But most audio signals are not continuous - when examined in the frequency domain, each frequency component in the audio signal changes over time. For frequencies under 1500 Hz, the hair cells on the membrane can fire nerve impulses fast enough to follow the positive half of the waveform of the vibration of the basilar membrane - providing continuous information of the frequency component's level envelope and relative phase. For higher frequencies, the vibrations go too fast for the hair cell to follow the waveform continuously - explaining that for continuous signals humans can hardly detect relative phase for high frequencies.

If there would be only one hair cell connected to the brain with only one neuron, the maximum time/phase detection would be the reciprocal of the neuron's thought maximum firing rate of 600 Hz, which is 1667 microseconds. But the cochlear nerve string includes as much as 30,000 afferent neurons, their combined firing rate theoretically could reach up to 18 MHz - with a corresponding theoretical time/phase detection threshold of 0.055 microseconds. Based on this thought, the human auditory system's time/phase sensitivity could be anywhere between 0.055 and 1667 microseconds. To find out exactly, Dr. Milind N. Kunchur from the department of physics and astronomy of the university of South Carolina performed a clinical experiment in 2007, playing a 7kHz square wave signal simultaneously through two identical high quality loudspeakers[*4K]. The frequency of 7kHz was selected to rule out any audible comb filtering: the first harmonic of a square wave is at 3 times the fundamental frequency, in this case at 21kHz - above the frequency limit, so only the 7 kHz fundamental could be heard with minimum comb filtering attenuation. First the loudspeakers were placed at the same distance from the listener, and then one of the loudspeakers was positioned an exact amount of millimetres closer to the listener - asking the listener if he or she could detect the difference (without telling the distance - it was a blind test). The outcome of the experiment indicated that the threshold of the perception of timing difference between the two signals was 6 microseconds. A later experiment in 2008 confirmed this value to be even a little lower. In this white paper we propose 6 microseconds to be the timing limitation of the human auditory system. Note that the reciprocal of 6 microseconds is 166kHz - indicating that an audio system should be able to process this frequency to satisfy this timing perception - a frequency higher than the frequency limit of the cochlea. Kunchur identified the loudspeaker's high frequency driver as the bottleneck in his system, having to make modifications to the loudspeakers to avoid 'time smearing'. More on the timing demands for audio systems is presented in chapter 6.

The maximum time that humans can remember detailed audio signals in their short term *aural activity image memory* (*echoic memory*) is reported to be 20 seconds by Georg Sperling[*4L].

Using the ISO226 loudness contour hearing thresholds, the 120dB$_{SPL}$ pain threshold, Kunchur's 6 microsecond time coherence threshold and Sperling's echoic memory limit of 20 seconds, we propose to define the level, frequency and time limits of the human auditory system to lie within the gray area in figure 410:

*figure 410: the audio universe: level, frequency and time limits*



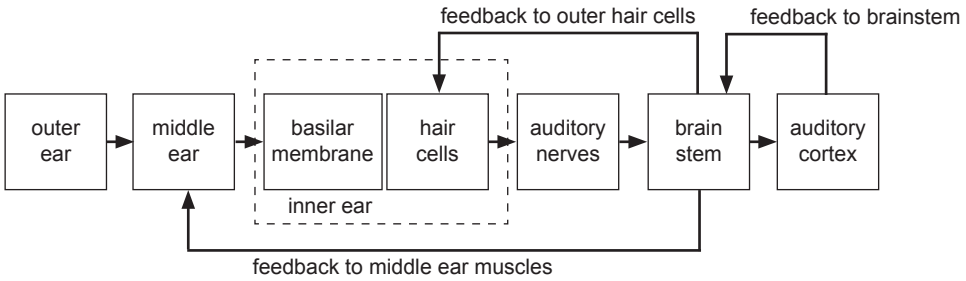
## 4.3 Auditory functions

The adult human brain consists of approximately one hundred billion ($10^{11}$) brain cells called *neurons*, capable of transmitting and receiving bio-electrical signals to each other[*4M]. Each neuron has connections with thousands of other nearby neurons through short connections (*dendrites*). Some neurons connect to other neurons or cells through an additional long connection (*axon*). With an estimated 5 trillion ($5*10^{14}$) connections, the human brain contains massive bio-electrical processing power to manage the human body processes, memory and 'thinking'.

All connections between sensory organs (eg. ears, eyes) and motor organs (eg. muscles) use axons from a dedicated 'multiconnector' section of the brain: the *brain stem*. After processing incoming information, the brain stem sends the information further to other sections of the brain dedicated to specific processes. Audio information is sent to the '*auditory cortex*'.

The brainstem receives data streams from both ears in the form of firing patterns that include information about the incoming audio signals. First, the brainstem feeds back commands to the middle ear muscles and the inner ear's outer hair cells to optimise hearing in real time. Although this is a subconscious process in the brainstem, it is assumed that the feedback is optimized through learning; individual listeners can train the physical part of their hearing capabilities.

*figure 411: schematic diagram of the human auditory system's feedback processes*



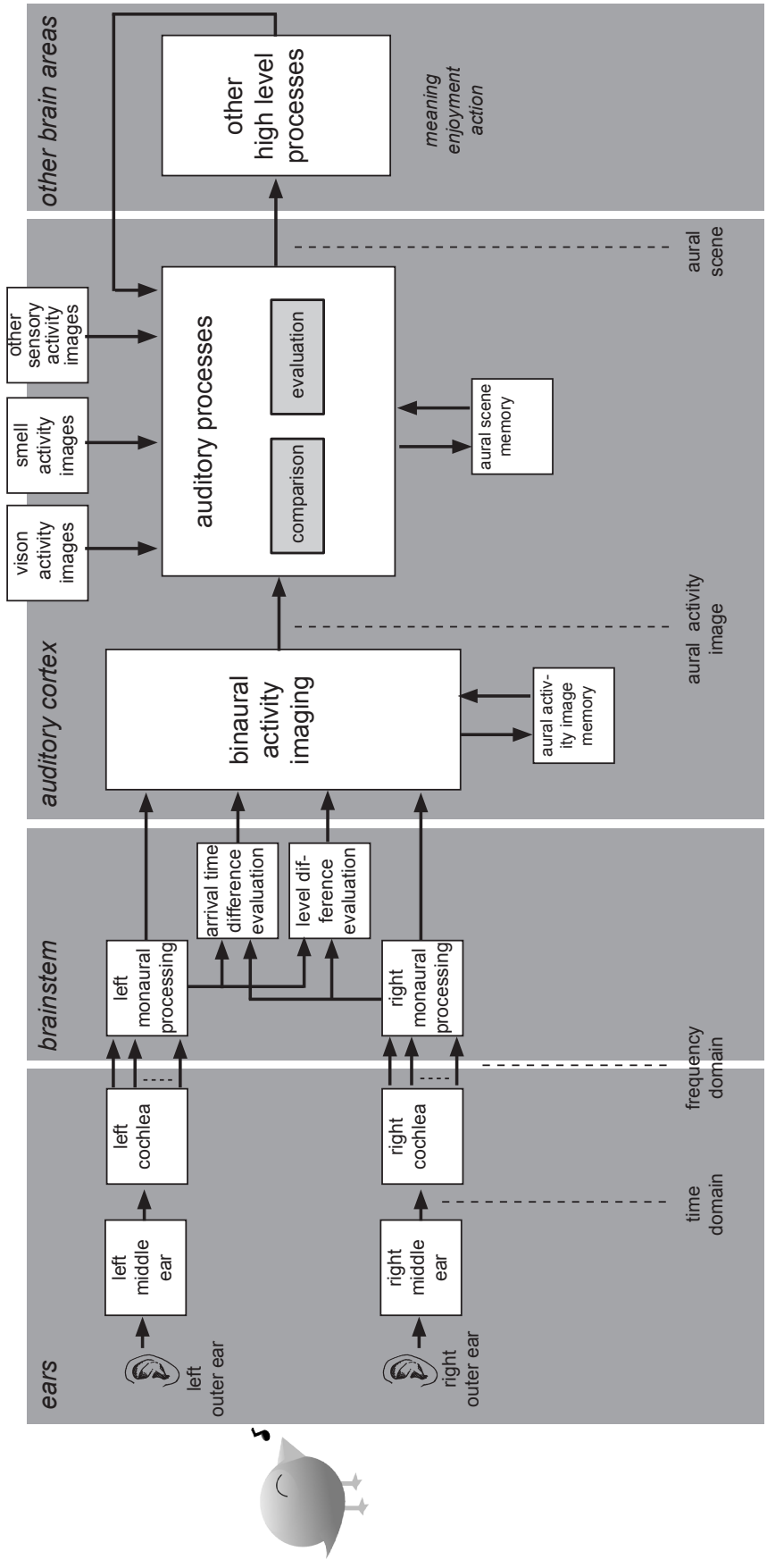

But then the real processing power of the brain kicks in. Rather than interpreting, storing and recalling each of the billions of nerve impulses transmitted by our ears to our brain every day, dedicated parts of the brain - the brain stem and the auditory cortex - interpret the incoming information and convert it into hearing sensations that can be enjoyed and stored as information units with a higher abstraction than the original incoming information: *aural activity images* and *aural scenes*. To illustrate this, we propose to use a simplified *auditory processing model* as shown in figure 412 on the next page[*4N].

The model describes how the incoming audio signal is sent by the ears to the brainstem in two information streams: one from the left ear and one from the right ear. The information streams are received by the brainstem in the frequency domain with individual level envelopes for each of the 3,500 incoming frequency bands - represented by nerve impulse timing and density. The information is sent to the auditory cortex, grouping the spectrum into 24 'critical band rates' called *Barks* (after Barkhausen[*4O]) to be analysed to generate *aural activity images* including level, pitch, timbre and localisation. The localisation information in the aural activity image is extracted from both level differences (for high frequencies) and arrival time differences (for low frequencies) between the two information streams. The aural activity image represents the information contained in the nerve impulses as a more aggregated and compressed package of real-time information that is compact enough to be stored in short term memory (*echoic memory*), which can be up to 20 seconds long.

Comparing the aural activity images with previously stored images, other sensory images (eg. vision, smell, taste, touch, balance) and overall context, a further aggregated and compressed *aural scene* is created to represent the meaning of the hearing sensation. The aural scene is constructed using efficiency mechanisms - selecting only the relevant information in the auditory action image, and correction mechanisms - filling in gaps and repairing distortions in the auditory activity images.

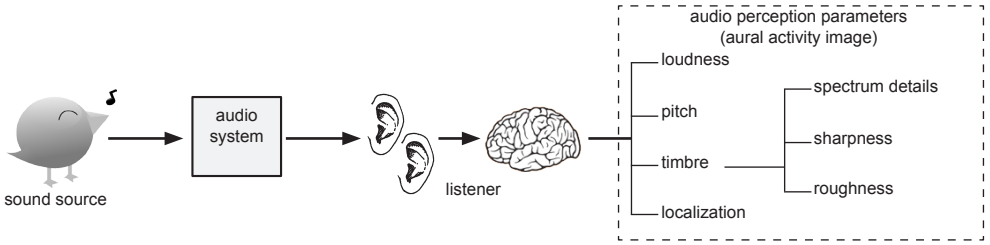
The aural scene is made available to the other processes in the brain - including thought processes such audio quality assessment.

*figure 412: auditory processing model*



The processing in the auditory cortex converts the raw audio data into more aggregated images: short term *aural activity images* containing the characteristics of the hearing sensation in detail, and more aggregated *aural scenes* representing the meaning of the hearing sensation. The science of describing, measuring and classifying the creation of hearing sensations by the human auditory cortex is the area of *psycho-acoustics*. In this paragraph we will very briefly describe the four main psycho-acoustic parameters of audio characteristics perception: loudness, pitch, timbre and localization. Also some particular issues such as masking, acoustic environment and visual environment are presented. Note that this chapter is only a very rough and by no means complete summary of the subject, for further details we recommend further reading of the literature listed in appendix 2.

*figure 413: psycho acoustics: main parameters of the perception of audio characteristics*



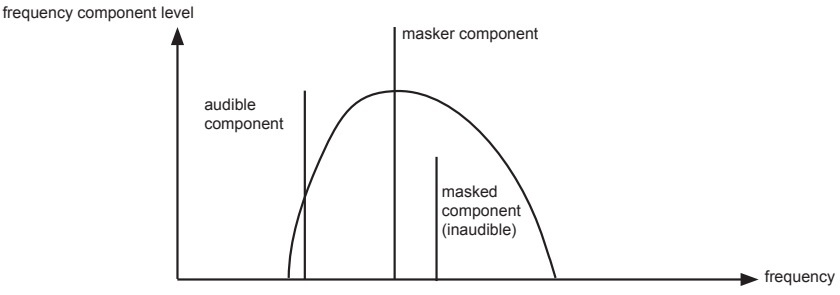

## Loudness

In psycho-acoustics, loudness is not the acoustic sound pressure level of an audio signal, but the individually perceived level of the hearing sensation. To allow comparison and analysis of loudness in the pshysical world (sound pressure level) and the psycho-acoustic world, Barkhausen defined *loudness level* as the sound pressure level of a 1kHz tone that is perceived just as loud as the audio signal. The unit is called '*phon*'. The best known visualization is the ISO226:2003 graph presented in chapter 4.2 which represents average human loudness perception in quiet for single tones.

## Masking

The loudness level of individual frequency components in an audio signal however is also strongly influenced by the shape (duration) of the frequency component's level envelope, and by other frequency components in the audio signal. The auditory cortex processing leads to an as efficient as possible result, picking up only the most relevant characteristics of the incoming signal. This means that some characteristics of the incoming signal will be aggregated or ignored - this is called *masking*. Temporal masking occurs when audio signals within a certain time frame are aggregated or ignored. Frequency masking occurs when an audio signal includes low level frequency components within a certain frequency range of a high level frequency component[*4P]. Clinical tests have shown that the detection threshold of lower level frequency components can be reduced by up to 50dB, with the masking area narrowing with higher frequencies of the masker component. Masking is used by audio compression algorithms such as MP3 with the same goal as the auditory cortex: to use memory as efficient as possible.

*figure 414: masking of nearby frequency components in the same signal*




## Pitch

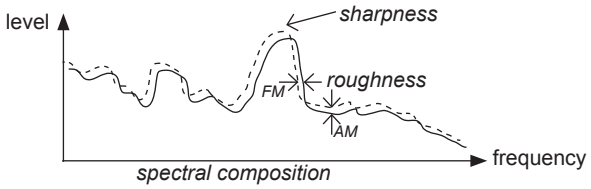
In psycho-acoustics, pitch is the perceived frequency of the content of an audio signal. If an audio signal is a summary of multiple audio signals from sound sources with individual pitches, the auditory cortex has the unique ability to decompose them to individual aural images, each with their own pitch (and loudness, timbre and localization). Psycho-acoustic pitch is not the same as the frequency of a signal component, as pitch perception is influenced by frequency in a non-linear way. Pitch is also influenced by the signal level and by other signal components in the signal. The unit *mel* (as in 'melody') was introduced to represent pitch ratio perception invoked by frequency ratios in the physical world. Of course in music, pitch is often represented by notes with the 'central A' at 440Hz.

**Timbre**

'Timbre' is basically a basket of phenomenon that are not part of the three other main parameters (loudness, pitch, localization). It includes the spectral composition details of the audio signal, often named 'sound colour', eg. 'warm sound' to describe high energy in a signal's low-mid frequency content. Apart from the spectral composition, a *sharpness*[*4Q] sensation is invoked if spectral energy concentrates in a spectral envelope (bandwidth) within one critical band. The effect is independent from the spectral fine structure of the audio signal. The unit of sharpness is *acum,* which is latin for 'sharp'.

Apart from the spectral composition, the auditory cortex processes are very sensitive to modulation of frequency components - either in frequency (FM) or amplitude (AM). For modulation frequencies below 15 Hz the sensation is called fluctuation, with a maximum effect at 4Hz. Fluctuation can be a positive attribute of an audio signal - eg 'tremolo' and 'vibrato' in music. Above 300Hz, multiple frequencies are perceived - in case of amplitude modulation three: the original, the sum and the difference frequencies. In the area between 15Hz and 300Hz the effect is called *roughness*[*4R], with the unit *asper* - Latin for rough. The amount of roughness is determined by the modulation depth, becoming audible only at relatively high depths.

*figure 415: timbre: spectral composition of an audio signal*



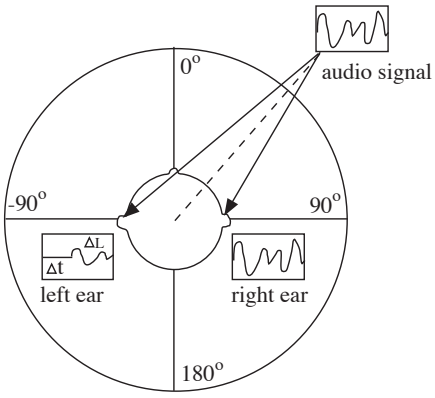

**Localisation**

For an average human being, the ears are situated on either side of the head, with the outside of the ear shells (*pinnae*) approximately 21 centimetres apart. With a speed of sound of 340 meter per second, this distance constitutes a time delay difference for signals arriving from positions at the far left or right of the head ($90^o$ or $-90^o$ in figure 416) of plus and minus 618 microseconds - well above the Kunchur limit of 6 microseconds. Signals arriving from sources located in front of the head ($0^o$ angle) arrive perfectly in time. The brain uses the time difference between the left ear and the right ear information to evaluate the horizontal position of the sound source.

The detection of *Interaural Time Differences* (or ITD's) uses frequency components up to 1,500 Hz - as for higher frequencies the phase between continuous waveforms becomes ambiguous. For high frequencies, another clue is used by the auditory cortex: the acoustic shadow of the head, causing attenuation of the high frequency components of signals coming from the other side (*Interaural Level Difference* or ILD).

Because the two ears provide two references in the horizontal plane, auditory localisation detects the horizontal position of the sound source from $90^o$ or $-90^o$ with a maximum accuracy of approximately $1^o$ (which corresponds to approximately 10 µs - close to the Kunchur limit). For vertical localisation and for front/rear detection, both ears provide almost the same clue, making it difficult to detect differences without prior knowledge of the sound source characteristics. To provide a clear second reference for vertical localisation and front/rear detection, the head has to be moved a little from time to time[*4S].

*figure 416: horizontal localisation through arrival time and spectral difference evaluation*

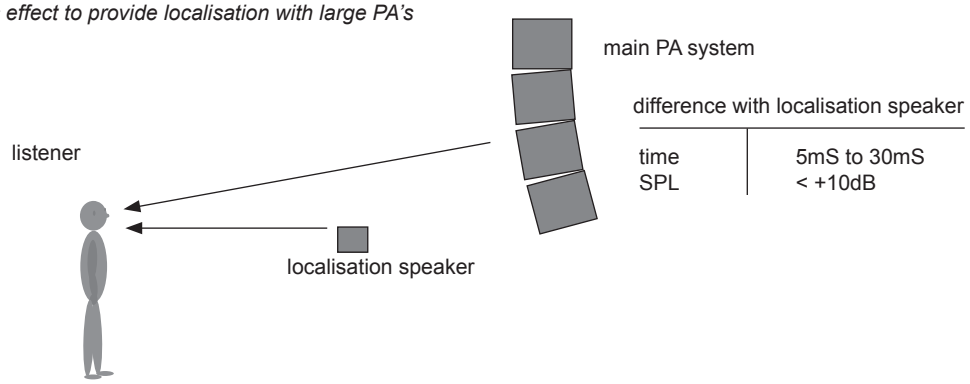


**Temporal masking**

An example of temporal masking is the *Haas effect*[*4T]. The brain spends significant processing power to the evaluation of time arrival differences between the two ears. This focus is so strong that identical audio signals following shortly after an already localised signal are perceived as one audio event - even if the following signal has an amplitude of up to 10dB more than the first signal. With the second signal delayed up to 30 milliseconds, the two signals are perceived as one event, localized at the position of the first signal. The perceived width however increases with the relative position, delay and level of the second signal. The second signal is perceived as a separate event if the delay is more than 30 milliseconds.

For performances where localisation plays an important role, this effect can be used to offer a better localisation when large scale PA systems are used. The main PA system is then used to provide a high sound pressure level to the audience, with smaller loudspeakers spread over the stage providing the localisation information. For such a system to function properly, the localisation loudspeakers wave fronts have to arrive at the audience between 5 and 30 milliseconds before the main PA system's wave front.

*figure 417: Utilising the Haas effect to provide localisation with large PA's*




**Acoustic environment**

The hearing sensation invoked by an audio signal emitted by a sound source is significantly influenced by the acoustic environment, in case of music and speech most often a hall or a room. First, after a few milliseconds, the room's first early reflections reach the ear, amplifying the perceived volume but not disturbing the localization too much (Haas effect). The reflections arriving at the ear between 20 ms and 150 ms mostly come from the side surfaces of the room, creating an additional 'envelopment' sound field that is perceived as the representation of the acoustic environment of the sound source. Reflections later than 150 ms have been reflected many times in a room, causing them to lose localization information, but still containing spectral information correlating with the original signal for a long time after the signal stops. The reverberation signal is perceived as a separate phenomenon, filling in gaps between signals. Long reverberation sounds pleasant with the appropriate music, but at the same time deteriorates the intelligibility of speech. A new development in electro-acoustics is the introduction of digital Acoustic Enhancement Systems such as Yamaha AFC, E-Acoustics LARES and Meyer Constellation to enhance or introduce variability of the reverberation in theatres and multi-purpose concert halls[*4U].

**Visual environment**

Visual inputs are known to affect the human auditory system's processing of aural inputs - the interpretation of aural information is often adjusted to match with visual information. Sometimes visual information even replaces aural information - for instance when speech recognition processes are involved. An example is the McGurk-Mcdonald effect, describing how the word 'Ba' can be preceived as 'Da' or even 'Ga' when the sound is dubbed to a film of a person's head pronouncing the word 'Ga'[*4V]. With live music, audio and visual content are of equal importance to entertain the audience - with similar amounts of money spent on audio and light and video equipment. For sound engineers, the way devices and user interfaces look have a significant influence on the appreciation - even if the provided DSP algorithms are identical. Listening sessions conducted 'sighted' instead of 'blind' have been proven to produce incorrect results.

# 5. Sampling issues

Until the late 19th century, audio systems for live and recording applications were designed using acoustical and mechanical tools - such as the walls of a concert hall (amplification, colouring), copper tubes in shipping (short distance transport), grooves on a wax roll picked up by a needle and amplified by a big horn (long distance transport). In 1872 everything changed when, independently from each other, Emil Berliner and Thomas Edison invented the carbon microphone[*5A], introducing the possibility of transforming acoustic waves to electronic signals. This was the beginning of the 'analogue' era - using electrical circuits to mix, amplify, modify, store and transport audio signals. This era saw the introduction of the electrical Gramophone in 1925 (Victor Orthophonic Victrola[*5B]), the reel to reel tape recorder in 1935 (AEG's Magnetophon[*5C]) and the compact cassette in 1962 (Philips[*5D]). In the years to follow, large scale live audio systems became available built around mixing consoles from Midas, Soundcraft, Yamaha and others.

## 5.1 Digital Audio

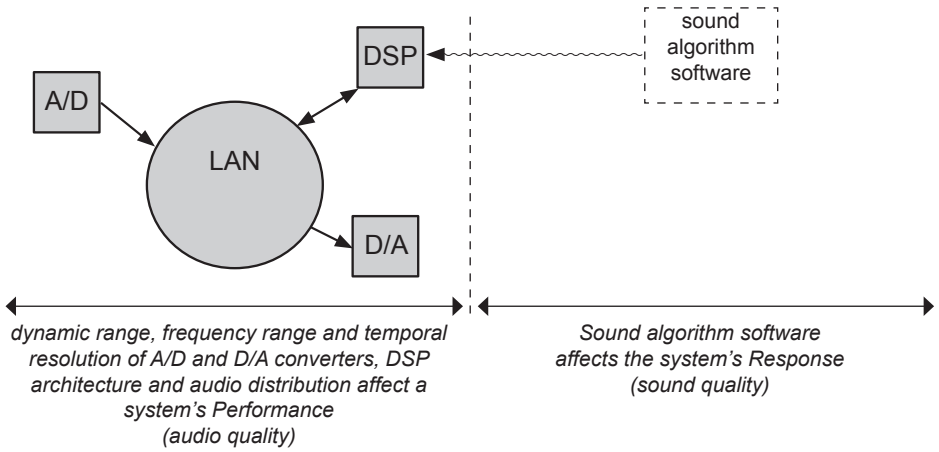
In 1938, Alec Reeves from ITT patented Pulse Code Modulation (PCM)[*5E], marking the start of the development of digital systems to process and transport audio signals - with the goal to make systems less susceptible for noise and distortion. In 1982, the first album on Compact Disc, developed by Philips and Sony, was released (Billy Joel's 52nd street) - introducing 16-bit 44.1 kHz PCM digital audio to a broad public[*5F].

The first mass produced Yamaha digital mixing system was the DMP7, launched in 1987 - introducing 16-bit A/D and D/A conversion and Digital Signal Processing (DSP) architecture to small scale music recording and line mixing applications[*5G]. But where 16 bits are (only just) enough for CD quality playback, the 90dB typical dynamic range it can reproduce is insufficient for large scale live mixing systems. In 1995, the Yamaha 02R digital mixing console introduced 20-bit A/D and D/A conversion, supporting 105 dB dynamic range[*5H]. Matching - or in many cases surpassing - the audio quality of analogue systems at that time, the 02R was the trigger for both the recording and the live audio market to start a massive migration from analogue to digital. Since then, A/D and D/A conversion technologies have matured to 24-bit A/D and D/A conversion and transport, and 32-bit or more DSP architecture, supporting system dynamic ranges of up to 110 dB - far beyond the capabilities of analogue systems[*5I]. DSP power has reached levels that support massive functionality that would have been simply not possible with analogue systems.

At the time of writing of this white paper, the majority of investments in professional live audio systems involve a digital mixer. Also, networked distribution systems to connect inputs and outputs to mixing consoles are now replacing digital and analogue multicore connection systems.

All networked audio systems use A/D converters to transform analogue audio signals to the digital domain through the *sampling* process, and D/A converters to transform samples back to the analogue domain. Assuming that all A/D and D/A converters, and the distribution of the samples through the digital audio system, are *linear*, sampling does not affect the Response of a digital audio system - all A/D and D/A processes and distribution methods *sound* the same. (only intended changes - eg. sound algorithm software installed on a system's DSP's - affect the sound). Instead, sampling and distribution can be seen as a potential limitation as it affects audio signals within the three audio dimensions: level, frequency and time. This chapter presents the sampling process limitations to a system's Performance: *dynamic range*, *frequency range* and *temporal resolution*.

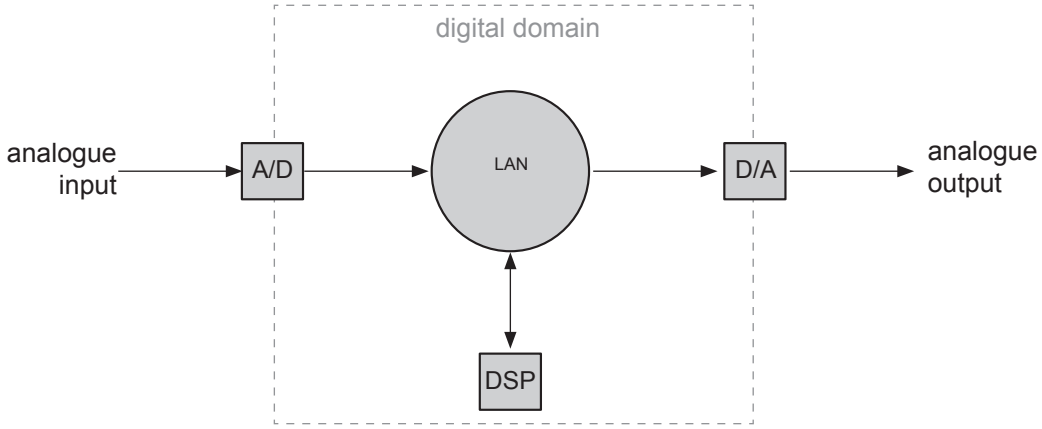
*figure 501: the sampling process in a networked audio system*



## 5.2 Dynamic range

A basic signal chain from analogue line input to line output in a networked audio system consists of an A/D converter, a distribution network (Local Area Network or LAN), a DSP unit and a D/A converter:

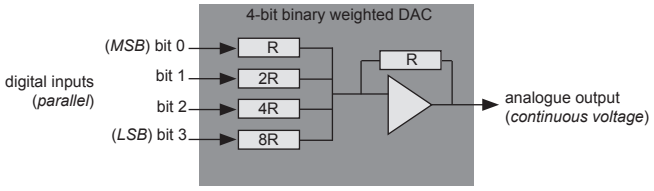
*figure 501: basic signal chain in an audio network*




A/D converter circuits are used to sample a continuous analogue (electronic) audio signal at a constant time-interval, in broadcast and live sound applications normally 20.8 microseconds, corresponding with a *sample rate* of 48,000 samples per second. Each time interval, the sampling process produces a number to represent the analogue signal using a certain amount of binary digits: the *bit depth*. For linear A/D converters with a bit depth of 24 bits, the sample can cover 16,777,216 different values. After transport through the audio network and DSP processing, a D/A converter converts the digital samples back to a continuous analogue signal, in most cases with the same sample rate and bit depth.

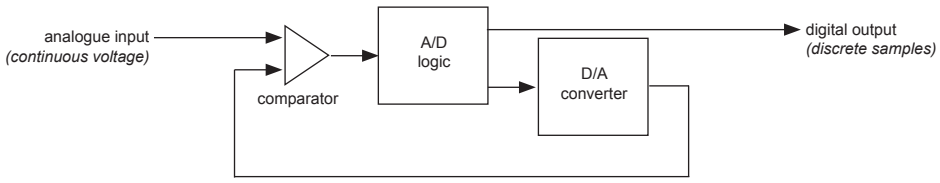
D/A converter circuits basically consist of one or more highly accurate switched current sources producing an analogue output current as a result of a digital input code. To illustrate the basic concept of a D/A converter, an example of a simple 4-bit binary weighed D/A converter is presented in figure 502. The circuit uses individual resistors for each bit, connected to a summing point. Commonly available D/A converters use more complex high speed techniques such as *delta-sigma modulation*[*5J].

*figure 502: simplified circuit diagam of a 4-bit binary weighted D/A converter*




A/D converters are slightly more complex, using an AD logic component driving a D/A converter. The output of the D/A converter is compared to an analogue input, with the result (higher or lower) driving the AD logic component.

*figure 503: simplified schematic diagram of an A/D converter*



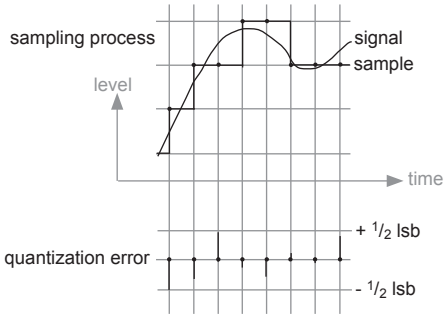
**Quantization error**

In a digital system's A/D converter, the sampled waveform can never be represented accurately by the digital samples because the value representation of the *least significant bit (LSB)* is always a little off by up to plus or minus $^1/_2$ LSB. Similar with A/D conversion, DSP MAC operations (Multiply and Accumulate[*5K]) and transformations from a high bit depth to a lower bit depth have to be rounded to up to plus or minus $^1/_2$ LSB after each processing step. These errors are called *quantization errors*.

The level of the quantization error in an A/D converter depends on the sample rate, the bit depth and the sampled waveform. The quantization error in a DSP process depends also on the number of steps. As D/A converters are assumed to reproduce the digital representation of the signal - and not the original signal, they do not generate quantization errors. However, a D/A converter can never reach a higher resolution than 1 LSB. Figure 504 illustrates an A/D converter's quantization error.

In binary numbers, each bit represents a factor of 2, which corresponds to a ratio of 6.02 dB. Although in reality it is much more complex[*5L], as a rule of thumb for A/D converters and DSP operations, a worst case quantization error noise floor of -6.02 dB times the bit depth is often used.

*figure 504: quantization error as a result of a sampling process*




**Linearity error**

Ideal A/D and D/A converters use 100% accurate components - but in real life, electronic components always possess a certain tolerance causing nonlinearity. It is assumed that professional A/D and D/A converters possess a linearity error of less than 1 LSB. Digital transport, networked distribution, storage and DSP processes are all digital and therefore don't generate linearity errors (unless compression is used eg. MP3).

**Dynamic range**

The internal resolution of dedicated audio DSP chips - the bit depth at which MAC operations are performed in the DSP core - is usually 40 bits or higher, providing enough internal resolution to keep the output of high DSP power algorithms (with many calculations) to stay well above 32 bits. If the audio network protocol used in the system is also 32 bits - eg. Dante in 32-bit mode, the dynamic range of the digital part of a networked audio system can be estimated using the rule of thumb of 6 dB per bit times 32 bits = 192 dB.

As this range is far greater than the theoretical 144 dB dynamic range limit of a system's A/D and D/A converters, it can be assumed that a system's dynamic range is mainly limited by just these components. Still, systems available on the market will not reach this value. Electronic circuits such as the power supplies, buffer amps and head amps before the A/D converter, as well as buffer amps and balancing circuits after the D/A converter, add electronic noise to the signal chain. Also, in most designs noise is added intentionally in the digital domain to improve the Performance at very low signal levels (*dither*)[*5M]. A typical dynamic range of a signal chain in a networked audio system - excluding the power amplifier - is 108dB[*5N].

## 5.3 Frequency range

Where the dynamic range of a signal chain in a networked audio system is mainly determined by the bit depth, the frequency range is mainly determined by the *sampling rate*.

The process of sampling an audio signal every sample interval is visualised in figure 505: the waveform of the audio signal is chopped into discrete samples with a high sample rate (figure 505a) and a low sample rate (figure 505b). It can be assumed that for high frequencies to be reproduced by a sampling system, a high sample rate has to be used. However, a high sample rate implicates that electronic circuits have to be designed to operate with high frequencies, the data transfer and DSP involve a high data rate (bandwidth), and storage requires a high storage capacity. For efficiency reasons, it is necessary to determine exactly how high a sampling rate must be to capture a certain frequency bandwidth of an audio signal, without too much costs for the design of electronic circuits, digital transport, DSP and storage.

*figure 505: the sampling process*

*figure 505a: high sample rates can reproduce high frequencies*

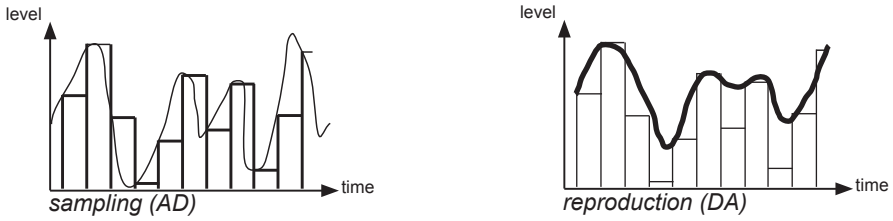
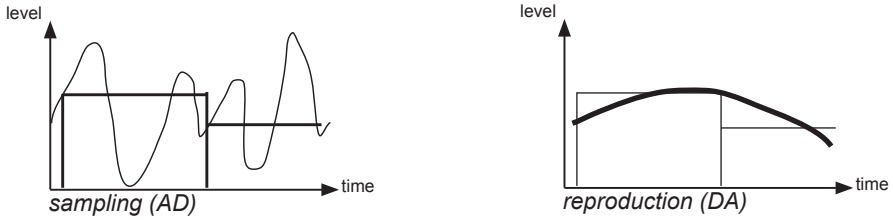



*figure 505b: low sample rates can not reproduce high frequencies*



To determine the minimum sampling rate to capture an audio signal's full bandwidth, figure 506 presents the sampling process both in the time domain and in the frequency domain. The most important observation when looking at the process in the frequency domain is that the sampling signal is a pulse train with a fundamental harmonic at the sampling frequency, and further odd harmonics at 3, 5, 7 etc. times the fundamental frequency. Multiplying the audio waveform with the sample pulse train results in the original waveform, plus sum and difference artefacts around every harmonic of the sample waveform. (Figure 506 only shows the 1st and 3rd harmonic).

*figure 506: a simplified representation of the sampling process*

*figure 506a: the waveform to be sampled*

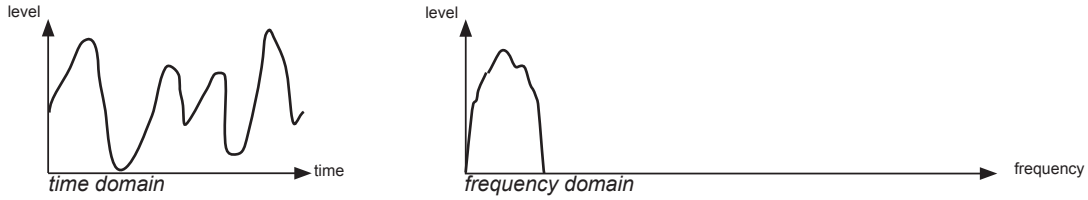



*figure 506b: the sample pulse train with frequency $f_s$*

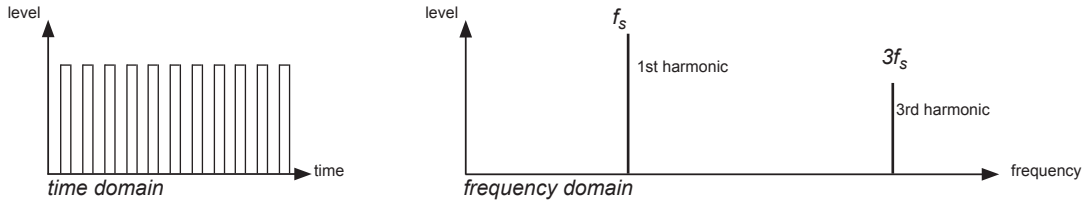
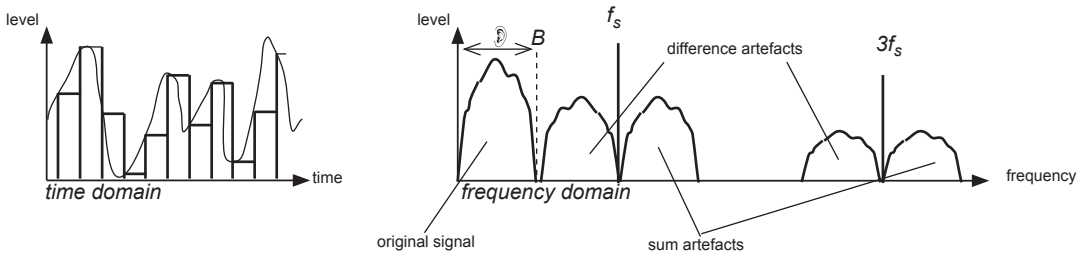



*figure 506c: the result after sampling*




In figure 506c, the human auditory system's frequency limit is indicated with the dotted line *B*. If the difference artefact of the 1st harmonic of the sample pulse train at the sample rate $f_S$ lies completely above B, then all artefacts fall outside of the hearing range - so the reproduction of the audio signal is accurate inside the boundaries of the audio universe. However, if either the audio signal has a full bandwidth (B) *and* $f_S$ falls below 2B (figure 507a), or if $f_S$ is twice the full bandwith B *and* the audio signal includes frequency components higher than B (figure 507b), the reproduction is no longer accurate because the difference artefacts of the first harmonic fall in the audible range. This phenomenon is called *aliasing*.

*figure 507a: $f_s$ falls below 2B*

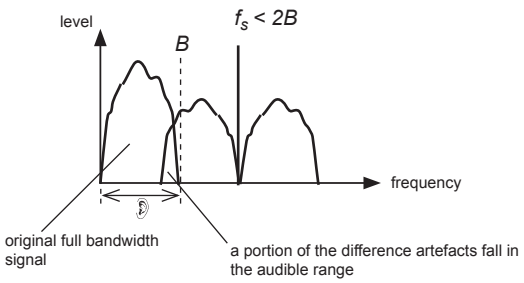
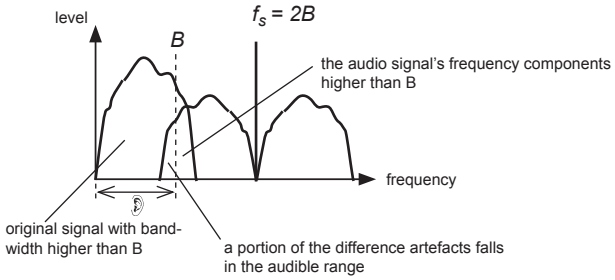


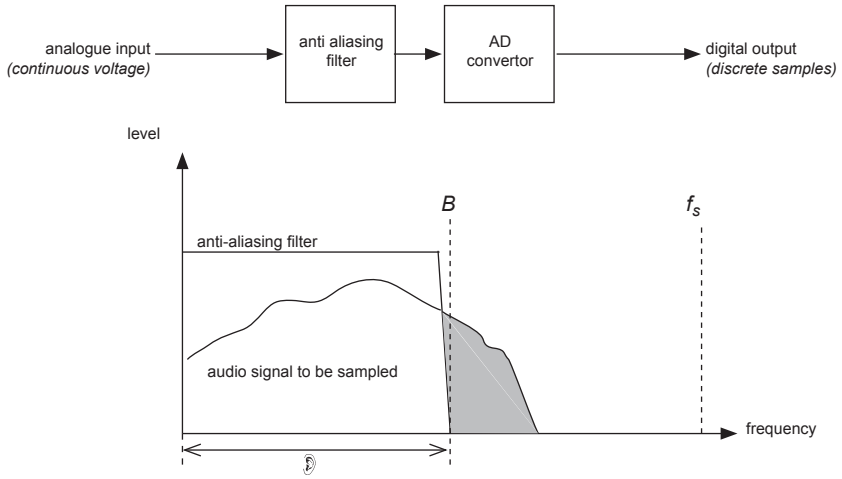

*figure 507b: $f_s$ = 2B but the audio signal has frequency components above B*



The conclusion is that 1) $f_s$ must be at least twice the frequency limit of the human auditory system (20 kHz) to be able to reproduce an accurate audio signal, and 2) the bandwidth of the audio signal must be less than half of $f_s$ . The second conclusion is called the Nyquist-Shannon sampling theorem[*5O].

The first conclusion leads to the system design parameter of selecting a sample frequency of at least 40 kHz to allow an accurate representation of signals up to the hearing limit of 20 kHz. But the second conclusion makes it very difficult to do so, because most audio signals that come into a digital audio system - eg. the output of a microphone - possess frequency components above 20 kHz. With a sampling frequency of 40 kHz, the frequencies above 20 kHz must be attenuated by an analogue *anti aliasing* (low pass) filter with a brickwall slope to prevent them from entering the audible hearing range.

*figure 508: anti aliasing filter*




In the analogue world, its impossible to design a brickwall filter with infinitive roll-off. The solution is to select a slightly higher sampling frequency to allow a less steep roll-off of the low pass filter. But still, a high roll-off slope creates phase shifts in the high frequency range - so to ensure phase linearity in the analogue anti aliasing filter, the sampling rate should be as high as possible. With the conception of the CD standard, Philips and Sony settled on a 44.1 kHz sampling rate as the optimal compromise between analogue filter design and CD storage capacity - leaving only 2 kHz for the roll-off, so the analogue filters on the first CD players where very tight designs. For broadcast and live applications, 48 kHz became the standard, allowing for slightly less tight anti aliasing filter designs. Some years later, with the availability of faster digital processing, the sampling rate could be set to a much higher level to allow much simpler analogue filters to be used - the processing to prevent aliasing could then be performed in the digital domain. This concept is called ‘oversampling’, now used by virtually all manufacturers of A/D and D/A converters[*5P].

*figure 509a: sampling without anti aliasing filter*

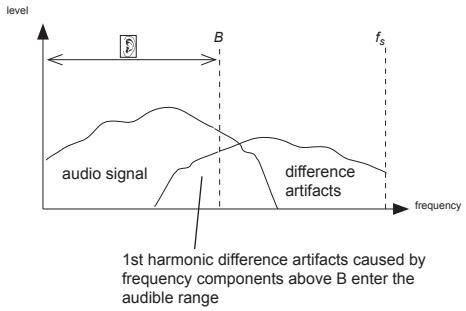
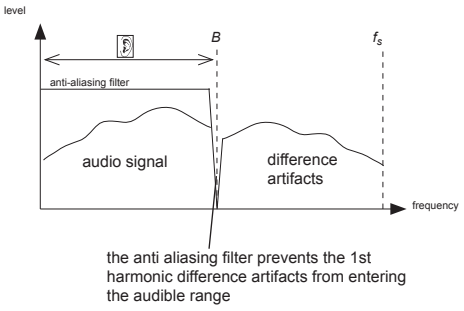



*figure 509b: sampling with anti aliasing filter*




All professional digital audio systems on the market today are capable of using a sample rate of at least 44.1 kHz with oversampling A/D and D/A converters, ensuring accurate full bandwidth conversion and processing of audio signals.

## 5.4 Timing issues

Signal propagation through cables happens at close to light speed - 299.792.458 meters per second[5Q]. The speed is limited by the cable's insulation - an unshielded cable will transfer audio signals at approximately 96% of the light speed, a coaxial cable at appr. 66%. Because of this limitation, a 100 meter coaxial cable will induce approximately 0,5 microseconds delay. Semiconductors also induce delays - a 74HC4053 analogue switch (an integrated circuit used in many remote-controlled head amps) typically adds 4 nanoseconds delay to the signal path[*5R].

Analogue signal distribution and processing circuits operate with this limitation - which normally does not affect audio quality unless for example the cabling spans more than 1,8 kilometers, lifting the delay above the 6 microseconds hearing limit.

Digital systems however operate at much lower speeds because signals are chopped in samples and then distributed and processed sample-by-sample - which requires a multiple of the sample time involved. For digital audio systems with a 48kHz sampling rate, the sample time is 20.8 microseconds - well above the 6 microseconds hearing limit. Where dynamic range and frequency range of digital audio systems have developed to span almost outside the reach of the human auditory system, time performance is a bottleneck that can not be solved. Instead, system designers and operators will have to study the time performance of their digital systems in detail, and take countermeasures to limit the consequences for audio quality as much as they can for every individual system design.

With analogue systems, engineers only had to concern themselves with the acoustic timing issues of microphone and speaker placement. With the introduction of digital (networked) audio systems, this task is expanded with the concern for digital timing or *latency.* The acoustic and digital timing issues basically cause the same effects, requiring a combined approach to achieve a high audio quality.

The following issues concerning time performance of digital audio systems will be covered in the following chapters:

*table 501: timing issues in digital systems*

| issue | typical delay (48kHz) | cause |
|---|---|---|
| network latency | 1 ms | one pass through a networked distribution system |
| processing latency | 1 ms | a digital mixer's signal processing |
| conversion latency | 0.5 ms | AD or D/A conversion |
| clock phase | 0.01 ms | synchronisation and receiver circuits |
| jitter | 0.000005 ms | PLL, electronic circuits, cabling, network |

## 5.5 Absolute latency

The latency of a signal chain in a typical networked audio system at a sampling frequency of 48 kHz is around 4 milliseconds. The main factors are distribution latency, DSP latency and AD/DA latency. In recording studio's with isolated control rooms - where the listener never hears the sound source directly - latency is not a problem at all. If the listener can see the sound source (but not hear it), a latency of up to 20 milliseconds (corresponding with the PAL video frame rate of 50 Hz) is allowed before video and audio synchronisation mismatch can be detected - mainly due to the slow reaction time of the eyes and the visual cortex in the human brain.

In live audio systems however, the audio signals radiated by the sound sources are often mixed with the processed output of the audio system - causing problems if the two signals have similar amplitude. With high latencies, a 'delay' is perceived by listeners and performers, which is specially inconvenient for musicians because it disturbs musical timing. Low latencies cause comb filtering, disturbing timbre and pitch. This problem is most prominent for vocal performances where a singer hears his or her own voice acoustically and through bone conduction, as well as through a monitor system.

For in-ear monitoring, latencies of more than 5 milliseconds might cause single sounds to be heard as two separate audio signals, disturbing musical timing. Below 5 milliseconds, latency causes comb filtering - which is also a problem, but can be adapted to by experienced performers. As a rule of thumb, high quality vocalist in-ear monitoring ideally requires a system latency (from microphone input to headphone output) of less than 5 milliseconds, preferably lower.

For monitoring systems using loudspeakers, the issue is less significant because the acoustic path from the loudspeaker to the ear is included in the signal chain, adding about 4 milliseconds or more acoustic delay. Although the latency is noticeable, because of the acoustic reflections the comb filter effect is less audible.

For Front Of House (FOH) Public Address (PA) speaker systems, the absolute latency of a networked system is much less significant because the distance of the listener to the sound source normally is several meters, adding tens of milliseconds acoustic delay - making the system's latency less significant. Also, FOH PA speaker systems normally are situated in front of the performers to prevent feedback, in some cases compensating the latency difference between system PA output and the source's direct sound for the audience. Figure 510 and table 502 show a typical networked live audio system and the latencies and acoustic delays that occur. Table 503 shows an average typical latency perception of musicians (pianist, vocalist, guitarist) analysed in numerous test sessions by Yamaha.

*figure 510: latency and acoustic delay in a typical networked live audio system*

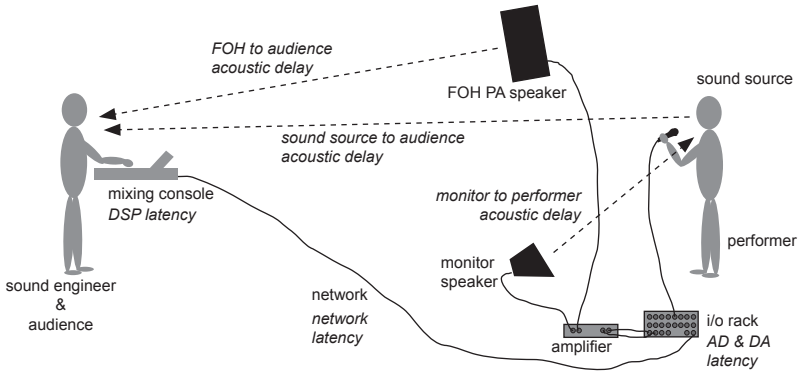



*table 502: latency and acoustic delay in a typical networked live audio system*

| signal path latency & acoustic delay (ms) | in-ear to performer | monitor speaker to performer | PA speaker to audience | sound source to audience |
|---|---|---|---|---|
| A/D conversion | 0.5 | 0.5 | 0.5 | n/a |
| network i/o rack -> FOH | 1 | 1 | 1 | n/a |
| DSP | 1 | 1 | 1 | n/a |
| network FOH -> i/o rack | 1 | 1 | 1 | n/a |
| D/A conversion | 0.5 | 0.5 | 0.5 | n/a |
| monitor speaker @ 2m* | n/a | 6 | n/a | n/a |
| PA speaker @ 20m* | n/a | n/a | 58 | n/a |
| sound source @ 23m* | n/a | n/a | n/a | 67 |
| | ----- | ----- | ----- | ----- |
| total latency (ms) | 4 | 10 | 62 | 67 |

** speed of sound = 343 m/s*

*table 503: average monitor system's absolute latency perception by musicians (Pianist, Vocalist, Guitarist)*

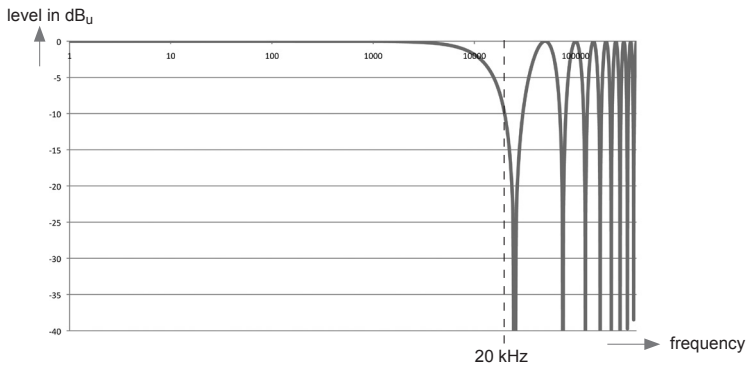
| signal path latency | for in-ear monitor systems | for floor monitor systems |
|---|---|---|
| 1.15 - 2 ms | Playable without any big problem. | Playable. |
| 2 - 5 ms | Playable, however tone colour is changed. | Playable. |
| 5 - 10 ms | Playing starts to become Difficult. Latency is noticeable. | Playable. Although latency is noticeable, it is perceived as ambience. |
| >10 ms | Impossible to play, the delay is too obvious. | Impossible to play, the delay is too obvious. |

## 5.6 Relative latency

Similar to the design challenge of coping with acoustic delays in speaker systems, latency in networked systems cause timing and comb filtering problems if multiple signal paths exist from sound source to listener. An example is a stereo overhead of a drum kit, where a certain portion of the outputs of the left and right microphones contain identical - *correlated* - signals. When these signals arrive at a listener with different latencies, the timing difference as well as the comb filtering can be detected by listeners - even for very small latencies down to 6 microseconds.

Figure 511 shows the comb filter effect caused by mixing two identical signals with a latency difference of 20 microseconds (roughly one sample at 48 kHz). Although the cancellation frequency of the comb filter effect lies at 24 kHz - above the hearing limit of 20 kHz, the slope already starts to attenuate frequencies within the hearing range, so the effect can be detected by listeners independent from the detection of the time difference itself. At 20 kHz, the attenuation is -10 dB. For latency differences of two samples or more, the cancellation frequency moves inside the hearing range causing clearly audible interference patterns - similar to the interference patterns created by the acoustic delay when using multiple loudspeakers.

*figure 511: comb filter effect for correlated signals mixed with a latency difference of 20 microseconds*



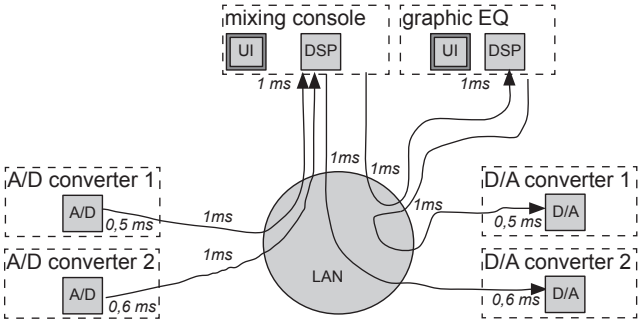

Many digital mixing consoles and digital signal processors have delays built in to manipulate latency - for example to correct relative latency problems.

Networks (timing protocol) and some DSP processes (latency compensation) have built-in latency correction mechanisms that align latencies of all signal paths in the system. However, it is important to remember that this is only valid for the default path - eg. a default signal chain in a mixer without any insertion, or a single pass through a network. If an external process is inserted in the signal path, eg. a graphic equalizer, the latency of that signal path increases with the latency of the graphic equalizer relative to all other signal paths in the mixer. If the equalizer is connected through a network, the network latency has to be added as well.

Another possible cause for latency differences is the use of different digital components, including A/D and D/A converters, which may add different latencies to the signal paths. In general, it is advisable to route correlated signals through the same type of components - preferably only one.

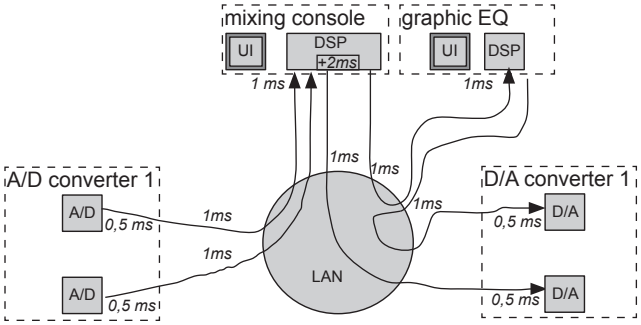
Figure 512A shows a system where 0,2 milliseconds (10 samples) latency difference is caused by using different A/D and D/A converters, and another 2 milliseconds (100 samples) is added by the inserted graphic equalizer and the extra pass through the network. Figure 512B shows the same application using the same A/D and D/A converter devices for both signal paths, and applying a manual compensation delay in the mixing console output for signal path 2, resulting in the same latency for both signal paths.

*figure 512A: signal paths with different latencies*




| signal path 1 | latency | signal path 2 | latency |
|---|---|---|---|
| A/D converter 1 | 0,6 | A/D converter 2 | 0,5 |
| LAN | 1 | LAN | 1 |
| mixer | 1 | mixer | 1 |
| LAN | 1 | LAN | 1 |
| graphic EQ | 1 | | |
| LAN | 1 | | |
| D/A converter 1 | 0,6 | D/A converter 2 | 0,5 |
| | ------ | | ------ |
| total latency | 6,2 | total latency | 4,0 |

*figure 512B: signal paths with matching latency*




| signal path 1 | latency | signal path 2 | latency |
|---|---|---|---|
| A/D converter 1 | 0,5 | A/D converter 1 | 0,5 |
| LAN | 1 | LAN | 1 |
| mixer | 1 | mixer | 1 |
| LAN | 1 | compensation delay | 2 |
| graphic EQ | 1 | LAN | 1 |
| LAN | 1 | | |
| D/A converter 1 | 0,5 | D/A converter 1 | 0,5 |
| | ------ | | ------ |
| total latency | 6,0 | total latency | 6,0 |

## 5.7 Word clock

All digital audio devices need a *word clock* to trigger all the device's digital audio processes. If a device is used as a stand-alone component - eg. a CD player, or a digital mixing console used with only analogue connections - then the internal clock is most often a crystal oscillator, providing a stable and accurate clock signal. However, as soon as two or more digital audio devices are used in a system, their internal word clocks need to use the same frequency to assure that all samples sent by one device are actually received by the other. If for example one device's internal word clock is running at 48,000 Hz, and it sends its data to a second device running at 47,999 Hz, each second one sample will go missing, causing a distortion in the audio signal. To prevent this, all devices in a digital audio system have to synchronize to a single, common '*master'* word clock source - which can be any of the devices in the network, or a separate *external* word clock generator.

Digital audio devices can not simply be connected directly to a master word clock - any disruption in the master word clock (eg. induced by electronic circuit noise, power supply noise) or its cabling (time delays, electromagnetic interference) can spread around the system to all other devices, with the possibility to cause unstable situations. To ensure stability in a digital system, all devices in it are synchronised to the master word clock through *Phase Lock Loop* (PLL) circuits, following only slow changes in the master clock's frequency, but ignoring fast disruptions. PLL's use a *Voltage Controlled Oscillator* (VCO) or a more accurate Crystal VCO (VCXO) to generate the device's internal sample clock to drive all its processes, with the frequency kept in phase with the master word clock by a phase comparator circuit that only follows the slow phase changes. VCXO PLL designs are suited for stable studio environments because they are very accurate. The downside of VCXO's is that they can only support a limited range of external frequencies - losing synchronisation if the master clock's frequency runs out of the specified range. Also, some VCXO based PLL designs can take a long time to synchronise. In broadcast and live systems, a broad range of sample rates have to be supported to enable the use of a variety of external synchronisation sources (eg. time code regenerated word clocks, digital tape recorders), with a fast synchronisation time to ensure a quick recovery in redundant systems. For this reason, VCO based PLL's are often used. With the introduction of the *Precision Time Protocol* (PTP) in AVB networks, also used by Dante, a part of the synchronisation is taken care of by the network interfaces.

From a functionality point of view, synchronisation signals are distributed separately from the audio data. Packet switching network protocols, such as CobraNet and Dante, distribute the synchronisation signal physically on the same cabling as the audio packets, but logically as separate data packets (figure 513a). Serial digital protocols such as AES3 (AES/EBU), AES10 (MADI), and packet streaming network protocols such as EtherSound, include the synchronisation signal in the audio stream (figure 513b). At any time, the designer of the system can decide to use an alternative word clock distribution using an external master word clock generator in order to synchronise to an external source, for example a video signal. An external word clock can be connected to just one device in the networked audio system - distributed to all other devices through the original distribution method (figure 513c), or to all devices individually (figure 513d). Special care should be taken for live systems with long cable runs as the word clock signals transported over coaxial cables are prone to degeneration, potentially causing synchronisation instabilities in a system.

*figure 513a: word clock distribution in a star topology network (eg. Dante)*

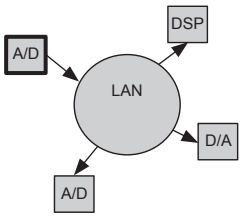



*figure 513b: word clock distribution in a daisy chain topology network (eg. EtherSound)*

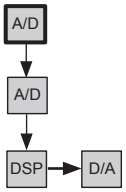



*figure 513c: word clock distribution in a star topology network with external word clock*

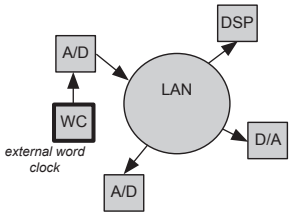
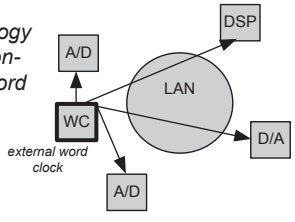



*figure 513d: word clock distribution in a star topology network with individual connections to an external word clock*



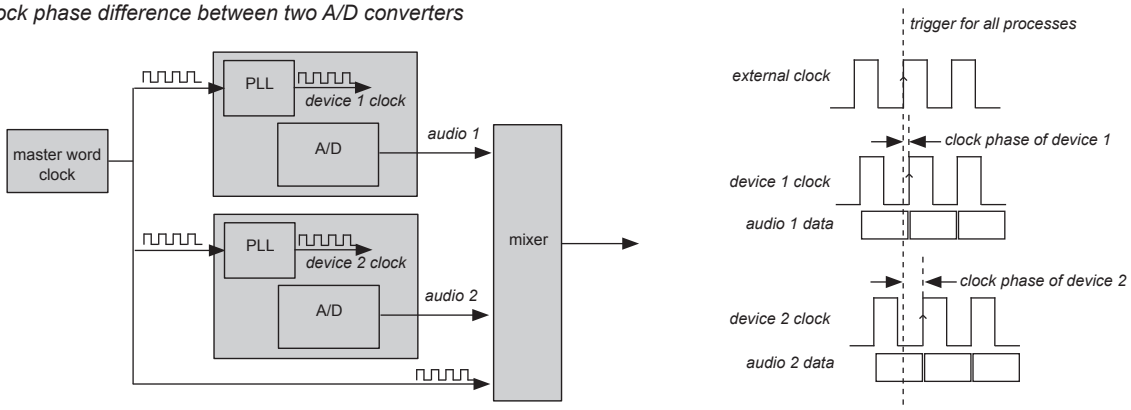
## 5.8 Clock phase

All processes in a digital audio system are triggered by a common synchronisation signal: the *master word clock*. The rising edge of the word clock signal triggers all processes in the system to happen simultaneously: A/D and D/A conversion, DSP processing and the transmission and receiving of audio data (transport).

For packet switching audio protocols using star topology networks (such as Cobranet, Dante, AVB), the distribution of the synchronisation information uses separately transmitted timing packets, ensuring that all devices in the network receive their clock within a fraction of the sample time (20.8 microseconds at 48 kHz). This assures for example that all AD samples in the system end up in the same sample slot. But for streaming audio protocols using daisy chain or ring topologies it is different: the clock information is represented by the audio data packets, so the length of the daisy chain or ring determines the clock phase caused by the network. For example, EtherSound adds 1,4 microseconds latency per node - accumulating to higher values when many nodes are used.

Additionally, a device that synchronises to an incoming synchronisation signal needs some time to do so. Digital circuit designers focus primarily on matching the incoming synchronisation signal's frequency as stable, flexible and fast as possible, rather than on the clock phase. As a result, digital audio devices on the professional audio market all have a different *clock phase*, which is almost never documented in the product specifications. Figure 514 presents a system that uses two different A/D converters that have a different clock phase: the samples are taken and sent to a digital mixer at a different time. As soon as the samples arrive in a digital mixer, the receiver circuit in the mixer will buffer and align the samples, mixing them as if they were taken at the same time, introducing a latency difference between the samples that is smaller than one sample. Because the latency difference caused by differences in clock phase are smaller than one sample, digital delays can not be used to compensate them - the minimum delay time available in a digital signal processor or mixer is one sample.

*figure 514: clock phase difference between two A/D converters*



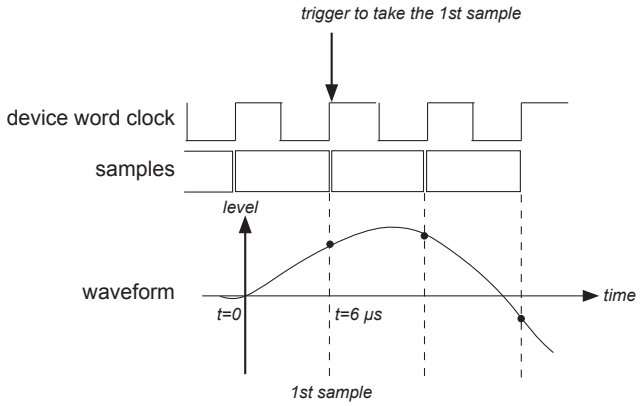

The compensation of clock phase differences can not be included in the system engineering phase of a project because almost all digital audio devices on the market do not specify clock phase delay. Conveniently, in practise, acoustically tuning a system (placing the microphones and speakers) already includes the compensation of all acoustical delays as well as clock phase; aligning all signal paths to produce a satisfactory audio quality and sound quality according to the system technician. As it is important not to move microphones and loudspeakers after a system has been acoustically tuned, it is of equal importance not to change routing and word clock distribution. Some 'internal/external clock' comparison listening sessions change the word clock distribution topology manually. The resulting differences in sound are often attributed to jitter performance, while in some cases clock phase changes might have been more significant: a clock phase difference close to half a sample can already be detected both in timing and comb filter effect (see also figure 511 in chapter 5.5).

## 5.9 Temporal resolution

Chapter 5.2 describes a digital audio system with a sample frequency of 48 kHz to be able to accurately represent frequencies up to 20 kHz. For continuous signals, this frequency is the limit of the human hearing system. But most audio signals are discontinuous, with constantly changing level and frequency spectrum - with the human auditory system being capable of detecting changes down to 6 microseconds.

To also accurately reproduce changes in a signal's frequency spectrum with a *temporal resolution* down to 6 microseconds, the sampling rate of a digital audio system must operate at a minimum of the reciprocal of 6 microseconds = 166 kHz. Figure 515 presents the sampling of an audio signal that starts at t = 0, and reaches a detectable level at t = 6 microseconds. To capture the onset of the waveform, the sample time must be at least 6 microseconds.

*figure 515: sampling of the onset of a waveform*




In the professional live audio field, a 48 kHz sampling rate is adopted as standard, with some devices supporting multiples of this rate: 96khz and 192kHz. (Some devices also support 44.1 khz and 88.2 kHz for compatibility with the music recording field, eg. the Compact Disk). However, apart from the temporal resolution of a digital part of an audio system, the temporal characteristics of the electro-acoustic components of a system also have to be considered. In general, only very high quality speaker systems specially designed for use in a music studio are capable of reproducing temporal resolutions down to 6 microseconds assumed that the listener is situated on-axis of the loudspeakers (*the sweet spot*). For the average high quality studio speaker systems, a temporal resolution of 10 microseconds might be the maximum possible. Live sound reinforcement speaker systems in general can not support such high temporal resolutions for several reasons.

Firstly, high power loudspeakers use large cones, membranes and coils in the transducers - possessing an increasing inertia at higher power ratings. A high inertia causes '*time smear*' - it takes some time for the transducer to follow the changes posed to the system by the power amplifier's output voltage. Some loudspeaker manufacturers publish 'waterfall' diagrams of the high frequency drivers, providing information about the driver's response to an impulse - often spanning several milliseconds. The inertia of a driver prevents it from reacting accurately to fast changes.

Secondly, live systems often use multiple loudspeakers to create a wide coverage area, contradictory to the concept of creating a sweet spot. The electro-acoustic designer of such a system will do what ever is possible to minimize the interference patterns of such a system, but the result will always have interference on all listening positions that is more significant than the temporal resolution of the digital part of the system.

Aside from audio quality parameters, the choice of a sampling rate can also affect the bandwidth - and with it the costs - of a networked audio system. Table 504 on the next page presents the main decision parameters.

As a rule of thumb, 48 kHz is a reasonable choice for most high quality live audio systems. For studio environments and for live systems using very high quality loudspeaker systems with the audience in a carefully designed sweet spot, 96 kHz might be an appropriate choice. Regarding speaker performance, 192 kHz might make sense for demanding studio environments with very high quality speaker systems - with single persons listening exclusively in the system's sweet spot.

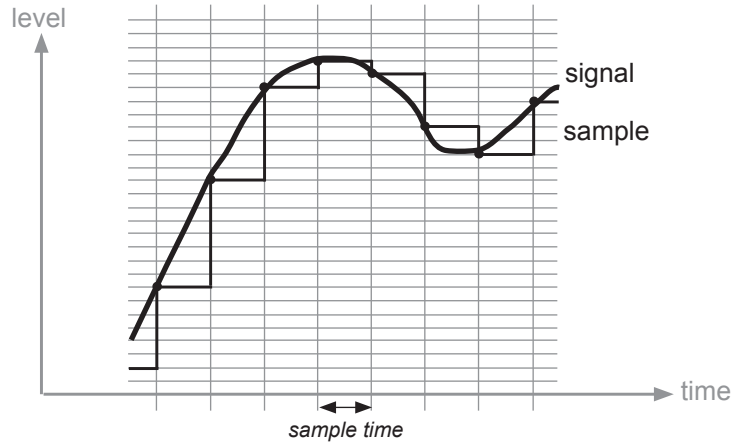
*table 504: Main decision parameters for the selection of a digital audio system's sample rate*

| | | |
|---|---|---|
| Audio quality issues | | |
| desired temporal resolution | | |
| 48 kHz | 20 µS - high quality | |
| 96 kHz | 10 µS - very high quality | |
| 192 kHz | 5 µS - beyond human threshold | |
| typical latency | | |
| 48 kHz | 4 ms default signal chain | |
| 96 kHz | 2 ms default signal chain | |
| 192 kHz | 1 ms default signal chain | |
| application type | | |
| sweet spot | supports high temporal resolutions | |
| wide coverage | difficult to achieve high temporal resolutions | |
| speakers | | |
| low power | might support 96kHz and 192 kHz at sweet spot | |
| high power | supports 48kHz | |
| cost & logistics issues | | |
| DSP power | | |
| 48 kHz | default DSP power rating | |
| 96 kHz | requires double DSP power | |
| 192 kHz | requires quadruple DSP power | |
| cable bandwidth (channels) | | |
| 48 kHz | default channel count | eg. Dante supports 512 channels |
| 96 kHz | reduced to 50% | eg. Dante supports 256 channels |
| 192 kHz | reduced to 25% | eg. Dante supports 128 channels |
| storage | | |
| 48 kHz | default storage | eg. 24GB for 1 hour 48ch (24-bit) |
| 96 kHz | requires double storage | eg. 48GB for 1 hour 48ch (24-bit) |
| 192 kHz | requires quadruple storage | eg. 96GB for 1 hour 48ch (24-bit) |

## 5.10 Jitter

In any digital audio system, time information is ignored. It is not registered by A/D converters, and it is not passed through the distribution protocol - the packets in AES/EBU bit streams or CobraNet bundles only include level information, not time information. Instead, the sample time is assumed to be reciprocal of the system's sample frequency - generated by system's master word clock. Furthermore, it is assumed that all samples are sent - and received - sequentially, and that there are no missing samples. Even the DSP algorithm programmer just assumes that his software will be installed on a system that runs on a certain sampling frequency - eg. filter coefficients are programmed to function correctly at 48 kHz. If the system's master word clock runs at 47 kHz, the system will probably be perfectly stable, but all filter parameters will be a little off.

*figure 516: sampling process with a constant sampling time*




All devices in a digital audio system have an internal word clock - most often synchronised to a common external word clock through a PLL circuit. Both the internal word clock signal and the external word clock are pulse trains at a frequency equal to the system's sample rate that provide a rising edge to trigger all processes in the system.

An ideal word clock will produce a rising edge in constant intervals. But in reality, noise in related electronic circuits (eg. oscillators, buffer amps, PLLs, power supplies) and electromagnetic interference and filtering in cabling will distort the word clock's waveform - causing the rising edge to come too early or too late, triggering the processes in the digital audio system at the wrong time. The signal that represents the deviation from the ideal time is called *jitter*. In digital audio products it is normally a noise-shaped signal with an amplitude of several nanoseconds.

*figure 517A: ideal word clock waveform*

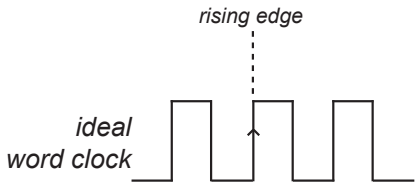



*figure 517B: word clock waveform with jitter*

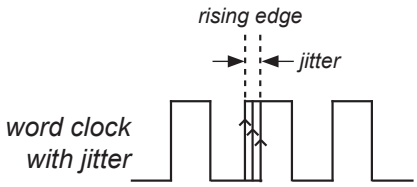
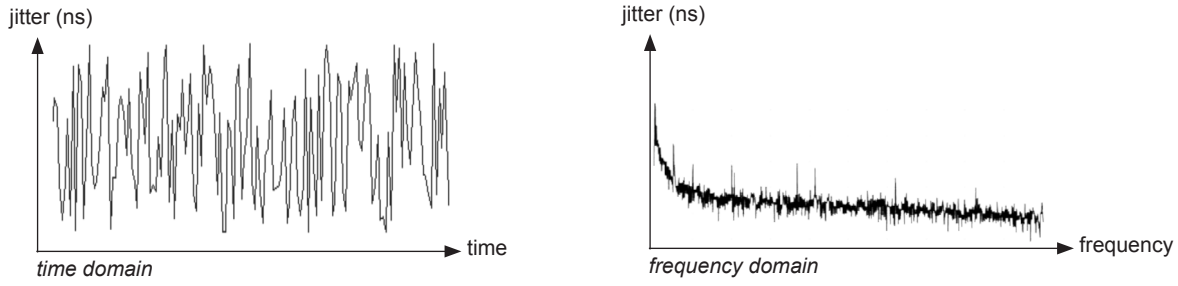
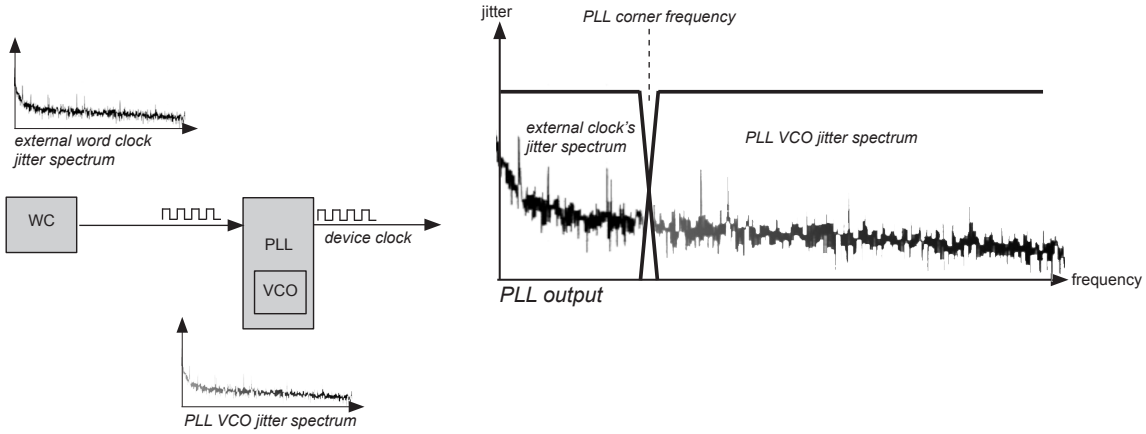



*figure 518: digital audio device's typical jitter in the time and frequency domain*



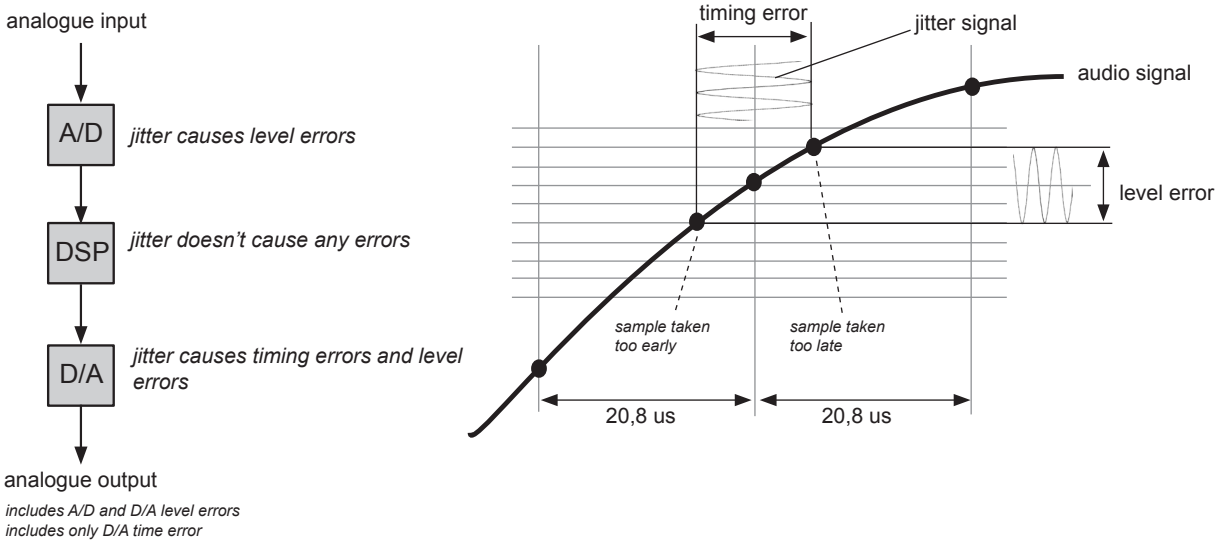
In digital audio systems, all devices synchronise to a common master clock through their PLL circuits. The PLL will follow only the slow changes in phase (low frequencies in the external word clock's jitter spectrum), and ignore the fast changes - keeping the PLL's own VCO's jitter spectrum. The jitter spectrum of the PLL's output (the device clock) is a mix of the low frequencies of the external word clock's jitter spectrum, and the high frequencies of the PLL's VCO jitter spectrum. The frequency where the external jitter starts to get attenuated is called the PLL's *corner frequency*. In digital audio devices for live applications, this frequency is normally between 10 Hz and 1 kHz with a relatively short synchronisation time (the time it takes for the PLL to synchronise to a new word clock). For studio equipment this frequency can be much lower - offering a higher immunity for external word clock quality, but with a longer synchronisation time.

*figure 518: PLL corner frequency*




Jitter in a device's word clock is not a problem if the device only performs digital processes - a MAC operation in a DSP core performed a little earlier or later than scheduled gives exactly the same result. In digital audio systems, jitter only causes problems with A/D and D/A conversion. Figure 519 shows that a sample being taken too early or too late: the *jitter timing error* - results in the wrong value: the jitter *level error*. As time information is ignored by the DSP processes and distribution of a digital audio system, only the level errors of an A/D converter are passed to the system's processes. At D/A converters however, samples being output too early or too late distort the audio signal in level and in timing. The listener to the system hears both A/D and D/A converter jitter level errors, but only the D/A converter's timing error.

*figure 519: the result of jitter: DA timing error and AD & DA level errors*



All major mixing consoles on the market today specify a jitter peak value of the internal PLL's output of several nanoseconds. This value is about a factor 1000 under the human hearing threshold of 6 microseconds - we propose to assume that such small timing errors can not be detected by the human auditory system.

Jitter level errors generated by this timing error however fall in the audible range of the human auditory system. For small jitter timing errors, with a sine wave audio signal A(t) and a noise shaped jitter signal J(t) with bandwidth B, the jitter level error E(t) is generated as presented in figure 520*[5S].

*figure 520: the level error generated by a sine wave as a result of noise shaped jitter*

*figure 520a: the audio signal: a sine wave with frequency $f_a$*

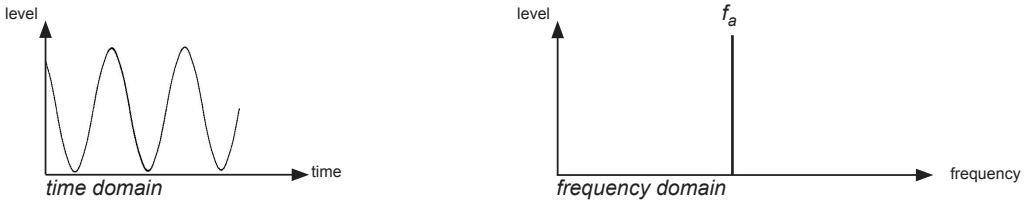



*figure 520b: the jitter signal*

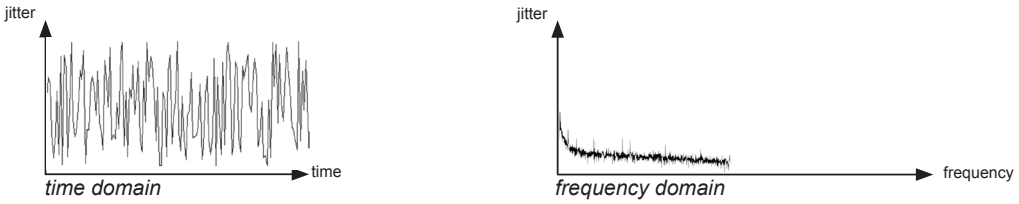
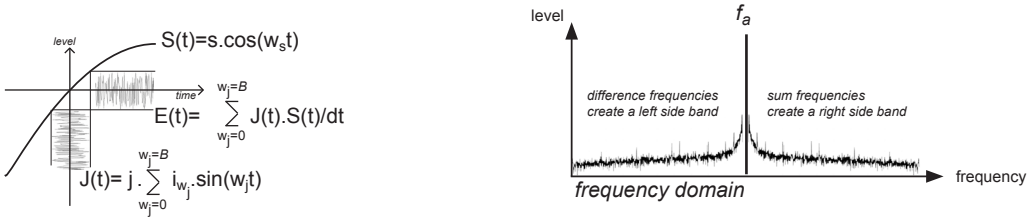



*figure 520c: the resulting level error is linear with the frequency of the audio signal, creating left and right side bands around $f_a$*




The result for a single frequency component in the audio signal is presented in figure 520c. For sinusoidal jitter, the term J(t).S(t)/dt can be represented by the expression $E(t) = s.j.w_S.\{^1/_2 \sin((w_S+w_j)t)+ ^1/_2 \sin((w_S-w_j)t)\}$. Adding up all frequencies in the jitter spectrum, it can be shown that the jitter spectrum folds to the left and the right of the audio frequency - this 'jitter noise picture' can be produced by any FFT analyser connected to a digital audio device that processes a high frequency sine wave. Repeating this calculation for every frequency component in a real life audio signal gives the resulting total jitter level error. The overall peak value of the jitter level error (E) is linear with the derivative of the audio signal: the higher the frequency, the faster the signal changes over time, the higher the jitter level error. The worst case is a 20kHz sine wave, generating a jitter level error at a 64dB lower level. As most energy in real life audio signals is in the low frequencies, the majority of the generated jitter level errors will be far below -64dB.

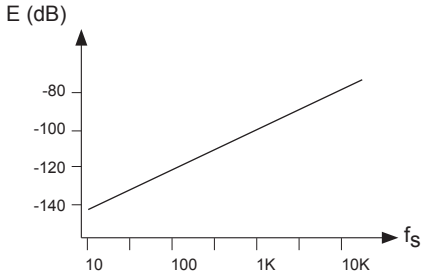
*figure 521: jitter level error peak value E as a function of the audio signal's frequency*

$E = 20.\log(j.2.\pi.f_s)$

E = jitter level error peak value in dB
j = jitter signal peak amplitude (s)
$f_s$= audio signal frequency (Hz)

for j=5 ns:

| $f_s$ (Hz) | E (dB) |
|---|---|
| 10 Hz | -124 |
| 100 Hz | -110 |
| 1,000 Hz | -90 |
| 2,000 Hz | -84 |
| 10,000 Hz | -70 |
| 20,000 Hz | -64 |



**Low frequency jitter correlation**

As PLL circuits follow the frequencies in the master word clock's jitter spectrum below the PLL's corner frequency, the low frequency jitter in all devices in a digital audio system will be the same - or *correlated*. Theoretically, for a signal that is sampled by an A/D converter and then directly reproduced by a D/A converter, all jitter level errors generated by the A/D converter will be cancelled by the jitter level errors generated by the D/A converter, so there is no jitter level error noise. In real life there is always latency between inputs and outputs, causing the jitter signals to become less correlated for high frequencies. If the latency between input and output increases, then the frequency at which the jitter is correlated will decrease. At a system latency of 2 milliseconds, the correlation ends at about 40 Hz - this means that in live systems, low frequency jitter (that has the highest energy) is automatically suppressed, but high frequency jitter in A/D converters and D/A converters just add up. In music production systems - where the audio signals are stored on a hard disk, posing a latency of at least a few seconds - and of course in playback systems where the latency can grow up to several years between the production and the purchase of a CD or DVD, the low frequency jitter signals are no longer correlated and all jitter level errors will add up.

For packet switching network protocols using the Precision Time Protocol[*5T] (*PTP*), such as Dante and AVB, the synchronisation is partly covered by the receiver's FPGA logic, adjusting a local oscillator to run in sync with up to 10 synchronisation packets per second. This means that the equivalent corner frequency of a PTP receiver is under 10Hz - correlating only for very low frequencies. In such systems, the influence of an external wordclock distributed through low latency networks as in figure 513c is not significant.

**audibility of jitter**

Assuming a $0dB_{fs}$ sine wave audio signal with a frequency of 10kHz as a worst case scenario, a jitter signal with a peak level of 5ns will generate a combined A/D and D/A jitter noise peak level of:

$$E_{A/D+D/A} = 20.\log(2.5.10^{-9}2.\pi.10.10^{3}) = -64dB_{fs}$$

When exposed to listeners without the audio signal present, this would be clearly audible. However, in real life jitter noise only occurs with the audio signal in place, and in that case masking occurs: the jitter noise close to the audio signal frequency components will be inaudible, so the average audio signal's spectrum will mask a significant portion of the jitter noise.

Note that the predicted level is the jitter noise peak level generated by $0dB_{fs}$ audio signals. In real life, the average RMS level of jitter noise will be lowered by many dB's because of the audio program's crest factor and the system's safety level margins used by the sound engineer. Music with a crest factor of 10dB played through a digital audio system with a safety level margin of 10dB will then generate jitter noise below $-84dB_{fs}$.

The audibility of jitter is a popular topic on internet forums. Often a stand-alone digital mixing console is used in a listening session, toggling between its internal clock and external clock. In these comparisons it is important to know that such comparison sessions only work with a stand-alone device. If any other digital device is connected to the mixer, then clock phase might play a more significant role in the comparison results than jitter.

In uncontrolled tests, many subjective and non-auditory sensations have a significant influence on the result. More details on quality assessment methods are presented in chapter 9.

In multiple clinical tests, the perception threshold of jitter has been reported to lie between 10 nanoseconds[*5U] for sinusoidal jitter and 250 nanoseconds[*5V] for noise shaped jitter - with actual noise shaped jitter levels in popular digital mixing consoles being below 10 nanoseconds.

# 6 Distribution & DSP issues

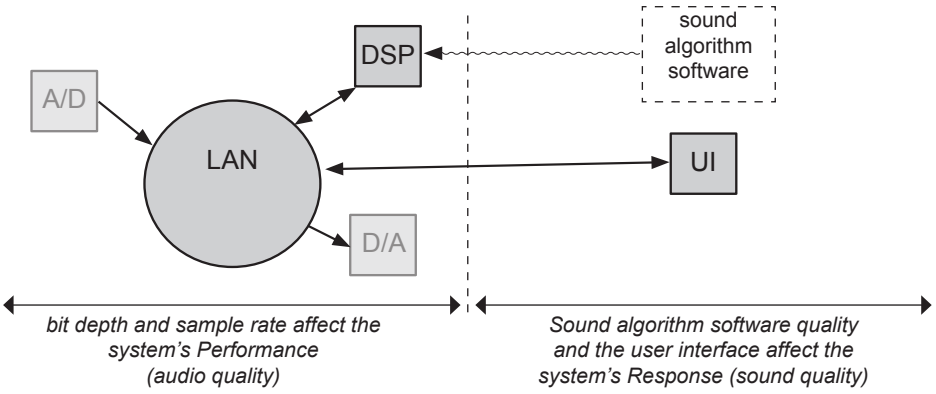
The core functionality of a networked audio system is determined by the *audio network protocol* and its hardware to take care of the distribution of audio signals throughout the system, and *Digital Signal Processors* (DSP) to process (*change*) the audio signals. This chapter focuses on the core distribution and DSP functionality.

The *sound quality* of a networked audio system is not affected by the distribution system. A properly set up networked (or any digital) connection will transfer digital audio signals without changing the samples - the only function of a distribution system is to distribute, and not to process. The DSP hardware itself also does not affect the sound quality of a networked audio system - the internal processes in a DSP core can be regarded as distribution processes - moving samples to and from the DSP's memory.

The DSP algorithms however do affect *sound quality* by definition - as the default status of a DSP algorithm is to just pass audio information from input to output without any change. Every change a DSP algorithm poses to an audio signal is therefore intended - it's the most important part of a digital audio system's *Response*.

The *audio quality* of a networked audio system on the other hand is not affected by the DSP algorithm - as all DSP algorithms are fully intended. Instead, only the bit depth and the sample rate of the distribution system and DSP hardware architecture affect the *Performance* of a system's core functionality.

*figure 601: distribution and DSP functionality of a networked audio system*



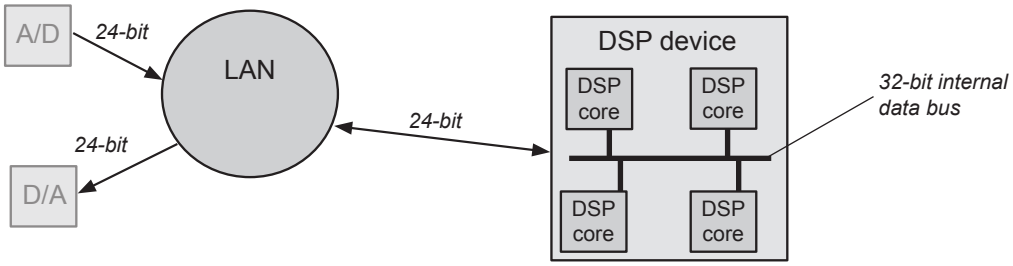

## 6.1 I/O distribution

If the bit depth of a networked audio system's A/D and D/A converters is 24-bits, then of course the connection to and from the network also has to be at least 24-bits to preserve the 24-bit resolution. Most conventional digital audio formats such as AES3 (AES/EBU), AES10 (MADI), and Ethernet compliant audio network protocols such as EtherSound and CobraNet are 24-bit digital audio formats. All these formats are linear - they transport the digital audio samples without changing them. All digital audio formats therefore 'sound' the same. However, formats do differ in latency - with the Ethernet compliant audio protocols such as EtherSound, CobraNet and Dante posing a higher latency than the conventional point-to-point connection formats such as AES3 and AES10. The implications of latency differences in correlated signal paths are described in detail in chapter 5.6.

## 6.2 Interconnected DSP distribution

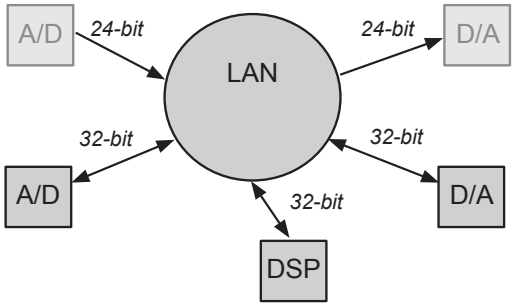
DSP MAC operations are performed at a bit depth higher than the 24 bit sampling bit depth to keep the accumulated quantization noise of rounding errors - generated with every MAC operation - below 24-bits. As most systems utilize multiple DSP chips, most DSP manufacturers support internal audio connections between DSP chips at a bit depth of 32 bits or more - reserving 8 or more bits for headroom and quantization noise to preserve a minimum resolution of 24 bits. If there is only one DSP device in the system, 24-bits are enough to connect to A/D converters and A/D converters in the system without loss of resolution.

*figure 602: 32-bit internal data bus for DSP, 24-bit external format to A/D and D/A converters*




If multiple DSP devices have to be connected in a system using external digital audio formats, 24-bit formats can be used, but this will increase the system's quantization noise above the 24-bit LSB level of -144dB. This is not a big problem if the signal path crosses such a connection only once - for example to connect to a digital speaker processor or a digital effect. But if a signal path uses the connection multiple times - for example in cascaded mixing consoles or in distributed *free configuration DSP systems*, the quantization noise might increase towards audible levels (above -120dB$_{FS}$, inside the audio universe dynamic range as presented in chapter 4.2). To support high resolution interconnection of devices, some devices offer 32-bit interconnection, such as the CL and Rio series supporting Dante 32-bit mode*[6A].

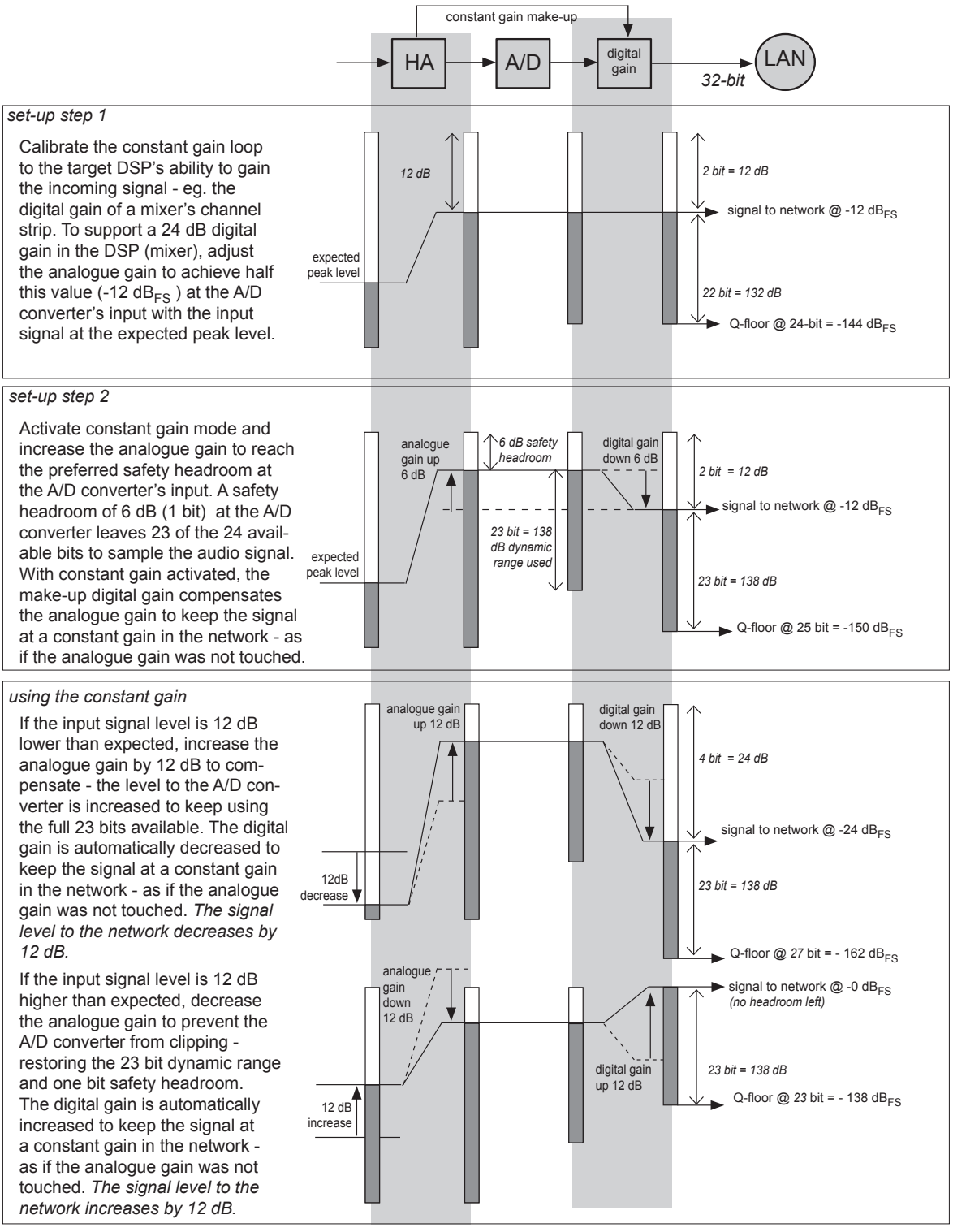
*figure 603: high resolution interconnected devices, eg. using Dante in 32-bit mode*



## 6.3 Constant gain A/D converters

In digital audio systems with ‘constant gain’ to keep the digital levels constant if analogue gain is changed (see chapter 7.3), the A/D converter device has DSP built-in to adjust the digital gain before the audio signal is sent to the network. In such a system, analogue gain is dedicated to the function of driving the A/D converter at the most efficient level (with the lowest quantization noise floor), leaving digital gains in the DSP processes to manage mixing input levels fully independently from the analogue gain. When constant gain is activated, a certain headroom is reserved to allow for digital make-up of the analogue gain. Figure 604 presents constant gain level diagrams with a 6 dB safety headroom and +/- 12 dB make-up range, supporting +24 dB digital gain in the DSP receivers (eg. mixers). To provide the three extra bits required for the worst case situation make-up, the network must support at least 27 bits to allow the use of constant gain without a decrease in resolution. A 24-bit digital audio protocol could be used, but then only a worst case resolution of 21 bits can be supported.

*figure 604: setting constant gain levels in a 32-bit distribution system.*



## 6.4 DSP architecture

DSP hardware is available in many forms - usually in the form of devices that combine multiple DSP chips to make one DSP system. Chips can be generic microprocessors from manufacturers such as Intel or AMD, dedicated DSP chips from manufacturers such as Yamaha, Analog Devices, Motorola or Texas Instruments, or generic Field Programmable Gate Arrays (FPGA) offered by many manufacturers. In theory, any digital audio algorithm can be compiled for all of these platforms - the choice of DSP platform has no effect on the *sound quality* of an algorithm. However, the *audio quality* between DSP platforms and their implementations differ - with the main issues being DSP power, latency and bit depth.

### DSP power

DSP power is the amount of processing available for MAC operations. Similar to the developments in the computer industry - with Moore's law predicting a doubling of computing power every 1,5 years*[6B] , DSP chips have developed to offer impressive amounts of DSP power, accommodating digital audio systems as a single DSP chip, as a combination of DSP chips connected through an internal data bus in a single device, or as a combination of DSP devices connected through an external data bus.

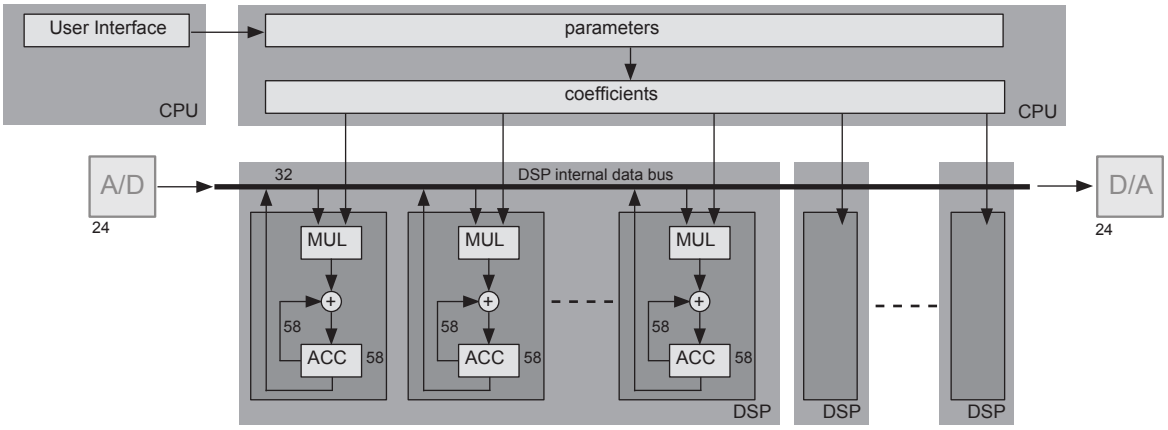
### Latency

Digital audio processing such as mixing, compression, equalisation requires a certain amount of DSP power. Dedicated DSP chips perform operations in parallel, resulting in a very low latency per operation. The total latency of the system is then dependent on the number of operations required to perform the DSP processes in a signal chain - for example a channel strip in a mixing console. *Native**[6C] systems - using generic microprocessors - perform MAC operations for all signal chains in the system one by one, causing much more latency. But because generic microprocessors have developed  to very high speeds - mostly well above 1GHz - the latency of a powerful native system can be low enough to allow the use in live systems.

### Bit depth

A signal that is processed in a DSP chip can include the results of many MAC operations - sometimes thousands when Infinite Impulse Response (IIR) or Finite Impulse Response (FIR) algorithms are used. As every MAC operation produces a result of double the resolution of the processed samples coming in from the processor's data bus, the result has to be rounded before it can be written back - each time producing a rounding error (or quantization error). If there are thousands of MAC operations for a certain DSP process, these errors all add up to significant levels. To keep the accumulated error level outside of the audio universe (of 120dB$_{FS}$), DSPs use a higher internal bit depth - normally 32 bits or more. Figure 605 presents a Yamaha DSP architecture with a 32-bit data bus and a 58-bit accumulator bit depth.

*figure 605: DSP architecture with 32-bit data bus and 58-bit accumulator*



## 6.5 Fixed point vs floating point

For distribution of samples in a digital audio system, samples are normally represented by integer *fixed point* values with a certain bit depth. In DSP data buses however, often a *floating point* data format is used to allow a higher range of numbers to be used - implying a higher dynamic range. In a floating point representation, the available bits are divided in a *mantissa* - representing the value, and an *exponent* - representing the weight. In a 32-bit floating point representation such as used in many FPGA chips and native systems, 24 bits are used for the mantissa and 8 bits for the exponent, allowing for a 'virtual 280-bit depth' or '1,680 dB dynamic range'. For algorithm programmers this allows less strict programming compared to fixed point representation, as the limited dynamic range of fixed point 32 bits is no longer a constraint with complex processes using many MAC operations. However, there is a catch: the resolution of a floating point value is reduced with the bits needed for the exponent - in case of a 32-bit floating point representation, the resulting resolution of the mantissa is 24-bits at the best case, but even lower when levelling mistakes are made when programming or applying the algorithms. Figures 606A and 606B show a MAC operation in 32-bit fixed point and floating point representations that add a 32-bit -48 $dB_{FS}$ signal to a 32-bit -3 $dB_{FS}$ signal. The result in a floating point system has reduced the -48 $dB_{FS}$ input signal to 16 bits, while the fixed point representation still provides a 24-bit resolution.

*figure 606A: 32-bit fixed point addition of two 32-bit signals*

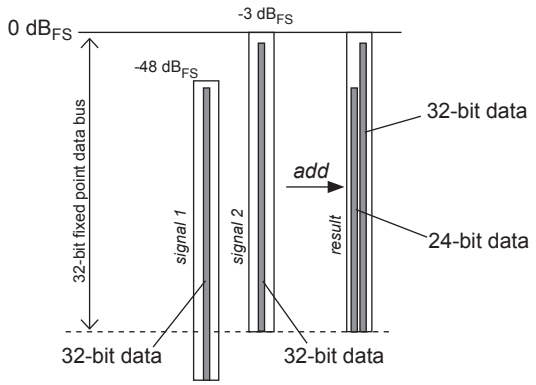
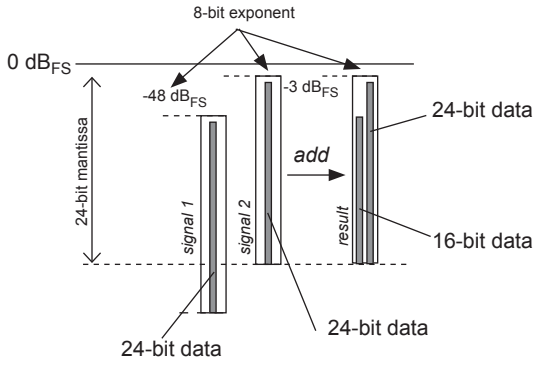



*figure 606B: 32-bit floating point addition of two 32-bit signals*




Motorola*[6D] and Yamaha DSP's have an internal DSP MAC operations bit depth of more than 56 or 58 bits - so for these DSP platforms the choice between fixed point and floating point is not very significant as the mantissa for a 56 bit or 58 bit value still has enough dynamic resolution to support algorithms with many MAC operations. Analog Devices SHARC (Super Harvard ARChitecture)*[6E] chips, Texas instruments*[6F] chips, FPGA chips and native DAW software and plug-ins can support both 32 bit and 64 bit floating point and fixed point modes, with some programmers using the 32-bit mode to offer a higher DSP power against lower costs. The issue is not significant if the programmer of the application used 64-bit mode, but both 32-bit fixed point and floating point modes pose dynamic range challenges to the algorithm programmer and sound engineer using the system. Fixed point systems support the highest resolution, but pose a risk of clipping: the algorithm programmer and the sound engineer have to constantly monitor the signal levels in the signal chain and take care that the level is below 0$dB_{FS}$ always. Floating point systems support a lower resolution with the risk of losing dynamic range of low level signals when mixed with high level signals, but there is never a risk of clipping in the floating point digital domain. In both cases, the sound engineer has to take care of the signal levels: in fixed point systems not to clip, in floating point systems not to lose dynamic range at low levels.

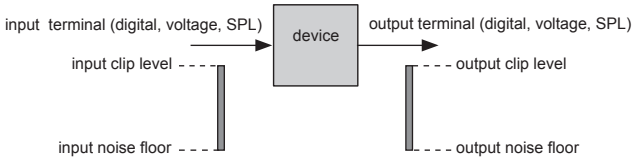
## 6.6 DSP user interfaces

When DSP algorithms are offered to a system's sound engineer as variable processes (*'colouring tools'* in a *natural sound* system), the way the sound engineer applies the DSP algorithm is significantly affected by the provided user interface. This means that the Response of a system is affected not only by the system's audio quality and sound quality, but also by the system's user interface quality. As with sound quality, this brings up similar quality discussions: individual sound engineers often prefer different user interfaces. Applying the same DSP algorithm with different user interfaces will lead to different sound qualities depending on the user interface preferences of the sound engineer. In many cases where the same DSP algorithms are available in a range of devices with different sizes of user interfaces, sound engineers can improve the response of systems with compact user interfaces by realizing that the DSP algorithm offers the same response as the ones with more elaborate user interfaces.

# 7. Level issues

All analogue and acoustic inputs and outputs - or *terminals* - of an audio device can accept a certain maximum level input signal - above this level the device's electronic circuit or mechanical construction connected to the terminal will *clip*. Also, the connected electronic circuits possess a *noise floor* that is independent from the signal level flowing through them.

*figure 701: terminal clip levels and noise floors*




As clipping of terminals cause unintended distortion, and noise floors limit the audio signal in the level dimension, system designers and sound engineers spend a considerable amount of their working time to manage the levels in audio systems to produce the highest dynamic range, and to make sure the signal never causes a terminal to clip. Also, signal chain levels must be constantly monitored and controlled to prevent clipping in case the SPL of the sound source or the output level of one of the system's processes is higher than anticipated.

## 7.1 '0dB$_{FS}$'

An audio system comprising of two or more audio devices possesses four or more terminals, each with their own clip level and noise floor. A signal in such a system passes all terminals on its route from the system input to the system output - risking to clip each terminal, and also picking up all noise levels. The ratio of the system's lowest clip level and the accumulated noise level constitutes the system's *dynamic range* or *clip to noise ratio.* There are three basic methods to route a signal through the successive devices in an audio system: *random level*, *matched noise floors* and *matched clip levels* - or *0dB$_{FS}$* - where FS stands for *Full Scale*: the highest level a terminal can handle without clipping. Figure 702 presents the three methods for a system of two devices, displaying the dynamic range of each terminal as a gray bar with the top representing the terminal's clip level and the bottom representing the terminal's noise floor. The resulting dynamic range is the ratio between the lowest clip level in the signal path and the accumulated noise floor at the last terminal.

*figure 702: signal level alignment methods through 4 terminals*

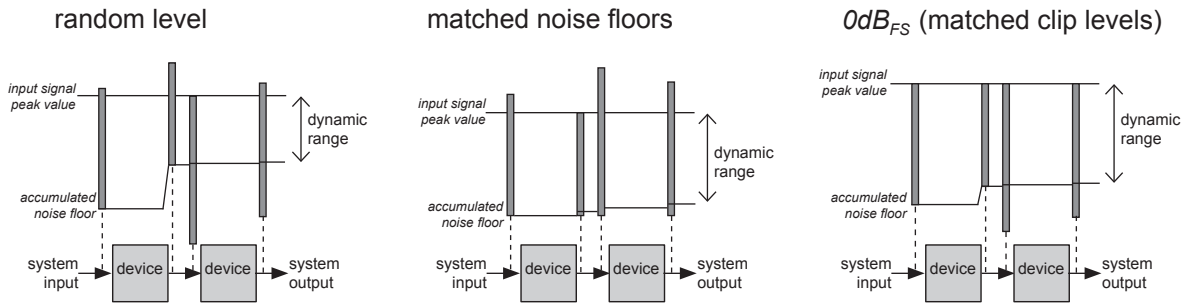
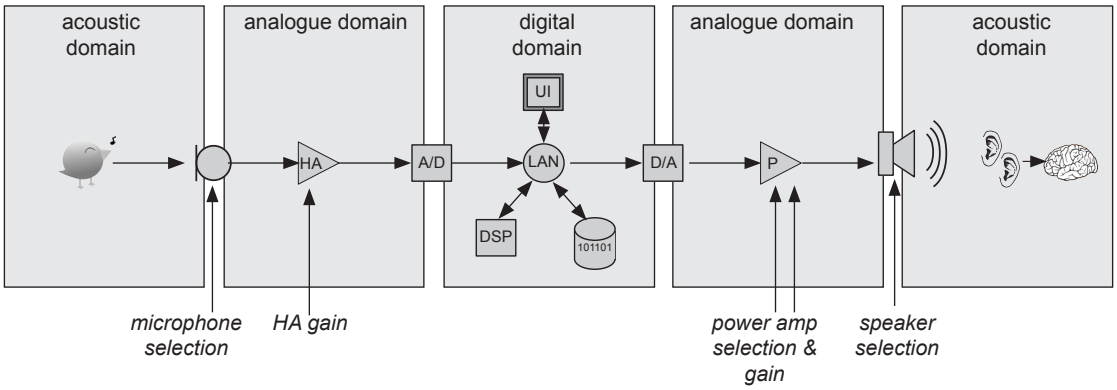



Figure 702 shows that when a random level alignment is used to route a signal through the audio system (which is any configuration other than matched noise floors or 0dB$_{FS}$), the system's dynamic range is the smallest because the dynamic range of individual terminals always overlaps with others. Matching noise floors minimises overlapping - resulting in a higher dynamic range, but it has the risk that the dynamic range of the last two devices in the signal path - the power amplifier and the speaker - are not always optimally used, causing higher costs. The 0dB$_{FS}$ method has the highest dynamic range because the noise floors are kept at the lowest level at each terminal - resulting in the lowest accumulated noise floor. Also, the 0dB$_{FS}$ method offers the lowest risk of clipping because all terminal clipping levels are aligned to the same level. Only the first terminal, which is normally the microphone's acoustic input, can clip as a result of excessive input SPL - all other terminals can never clip (at unity gain). The conclusion is that the 0dB$_{FS}$ method is the optimal strategy to make the most use of a system's dynamic range with the lowest risk of clipping and the lowest cost.

Assuming that the digital part of a networked audio system is always referenced to $0dB_{FS}$, there are five main parameters to align the system's acoustical and analogue terminals: the selection of the microphones, power amplifiers and speakers, the setting of the analogue gain of the head amp, and the voltage gain of the power amplifier.

*figure 703: $0dB_{FS}$ acoustic and analogue alignment parameters of a networked audio system*




Microphone selection

The microphones must be suited to handle the SPL of the sound source - with appropriate directional characteristics to best fit the sound source and environment characteristics, and of course the preferred Response ('sound').

Speaker selection

At the other end of the system, the loudspeakers must be suited to generate the required amount of SPL at the listener's position - and of course to have the preferred Response. To support electro-acoustic design, loudspeaker manufacturers specify the *sensitivity* of the speaker - being the SPL delivered to a listening position of 1 meter with an input of 1W at nominal impedance at 1kHz, allowing calculation of SPL at any distance using software such as EASE. Also, a maximum SPL is specified - often using the AES2-1984 (R2003) recommended practice for the specification of loudspeaker components*[7A]. Over-specification of loudspeakers leads to unused power and thus high costs, while under-specification of loudspeakers leads to unintended harmonic distortion that can not be classified as Response anymore - and of course the potential failure of the speaker.


Power amplifier selection

The selection of amplifiers must support the power requirements of the selected speakers. Similar to the selection criteria for speakers, over-specification of amplifiers lead to unused power and high noise floors, while under-specification of amplifiers lead to unintended distortion that can not be classified as Response - and of course the potential failure of the amplifier or speaker.


HA gain setting

The head amp gain allows the microphone's peak output voltage - generated by the sound source peak SPL - to match the A/D converter's peak input voltage. This sets the microphone input SPL and the microphone output voltage to match the system's $0dB_{FS}$ reference - creating the maximum dynamic range and eliminating the possibility of the A/D converter's analogue input to clip.

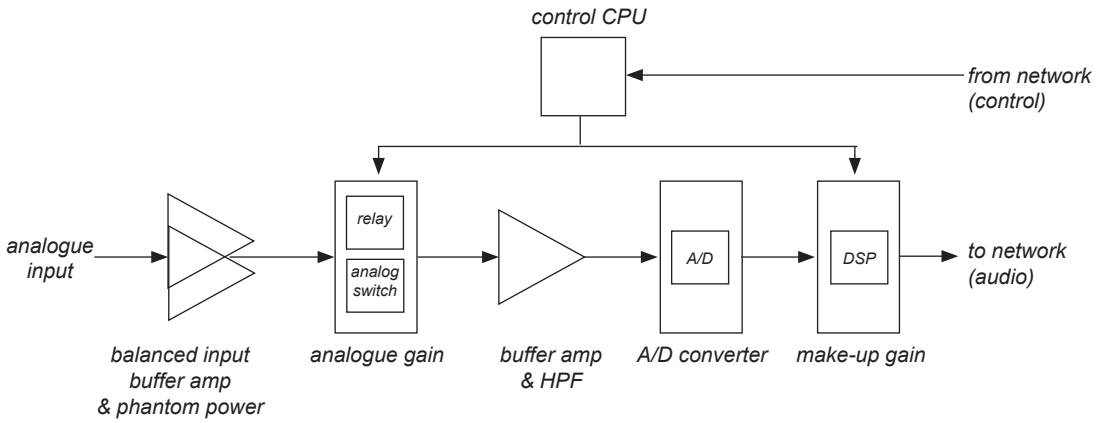

Power amplifier gain setting

The power amplifier gain allows the speaker's maximum voltage input - matched to the speaker's maximum SPL output by the sensitivity and peak SPL specifications - to be aligned with the D/A converter's peak output at the system's $0dB_{FS}$ reference - creating the maximum dynamic range and eliminating the possibility of the power amplifier and speaker to clip.

## 7.2 Head amps

As networked audio systems allow the placement of analogue inputs at remote locations - with the connection between the remote location and the control location (a mixing console) using an audio network protocol - the head amps that align the analogue microphone input levels with the remote input's A/D converter is also located at a remote location. To control the head amp's gain, high pass filter and phantom power, a remote head amp protocol must be used to control these functions from the control location. The head amp's gain is most commonly a combination of relays and/or analogue switches that provide a variable gain in the analogue domain. A digital make-up can be added to increase the gain step accuracy and to provide constant gain to the network (see chapter 6.3).

*figure 704: typical remote controlled head amp & A/D converter schematic diagram*




The noise floor and distortion of most high quality head amp circuits are normally very low. Designs from manufacturers with a *natural sound* philosophy provide a 'flat' transfer function from analogue input to A/D converter input - with as less as possible colouration, allowing DSP processes to apply sound colouration to the full control of the system designer and sound engineer. Some manufacturers apply a deliberate EQ curve in the analogue circuit to create a default Response that suits dedicated applications - eg. 'British sound' or 'warm sound'.

To support a flexible use of networked audio systems, it is of vital importance that the gain step accuracy is high - so that when a preset is recalled on the control user interface (mixing console), the recalled gain is as accurate as possible. The human auditory system is capable of detecting level differences of less than one dB. Also, to provide a high consistency when loading presets in a different console of the same type, or when reconnecting a sound source to another input in the system, the gain consistency between channels and between input head amps in different stage boxes must be very high. Still, because the total gain between a system's analogue input and an analogue output involves many electronic circuits, the accumulated signal chain gain consistency (or *gain error*) can be audible. This means that - even with very consistent head amps (and power amps) - every time a system is set up and connected, the head amp gains (and power amplifier gains) might have to be adjusted to comply to the design specifications.

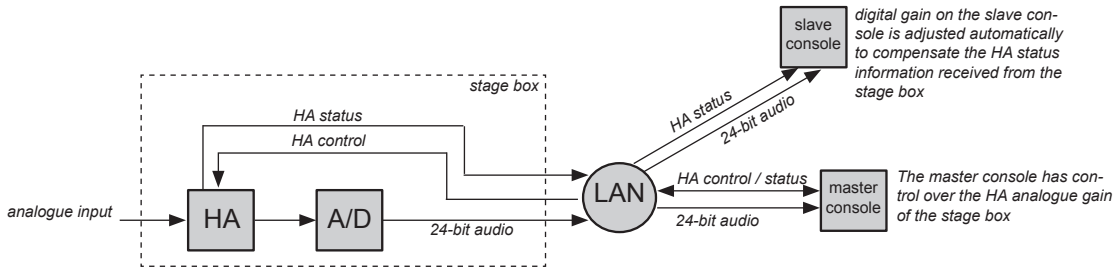
## 7.3 Gain compensation

In networked audio systems where multiple DSP processes and user interfaces (eg. mixing consoles) share the same inputs (eg. *stageboxes*), head amp control becomes an issue. When one user interface changes the analogue gain of a head amp to suit its related DSP process, the signal level to all other DSP processes changes as well. If the operators of the other processes are unaware of the analogue gain change, this can cause troubles - so in general, if multiple processes use the same inputs, the operators of the processes (in case of mixing consoles the sound engineers) have to clearly communicate with each other to adjust digital gains in the DSP process manually to compensate for the analogue gain change in the head amp. Two kinds of gain compensation schemes can be used to automate this process: *console gain compensation* and *constant gain*.

### Console gain compensation

Console gain compensation can be used in systems with two mixing consoles that are capable of handling the same HA control protocol as the stagebox. In most cases, one console is dedicated as the HA master console, the other as HA slave console. The master console controls the head amp in the stage box through a HA control protocol that flows through the same network as the audio. If the stagebox receives a gain change command (or a gain change is made locally), it executes it and sends a HA status back to the master console to update its analog gain display, and to the slave console to compensate the gain change using the slave console's digital gain. The advantage of console gain compensation is that it can be used with any digital audio format and network protocol. The disadvantage is that only stageboxes and mixing consoles that can handle the same HA control protocol can be used, and only mixers that have gain compensation implemented in the software.

*figure 705A: console gain compensation method*

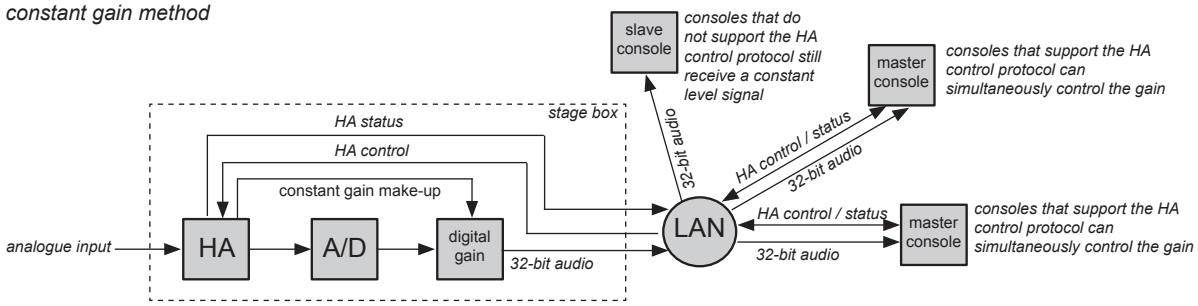



### Constant gain

Constant gain can be used if the stagebox offers dedicated DSP to compensate the analogue gain in the digital domain before sending the audio data to the network. The advantage of constant gain systems is that any DSP process (mixer) can be used in any quantity, as the compensation happens in the stagebox and not in the receiving DSP process. In a constant gain system, multiple operators can control the gain, provided they support the HA control protocol used in the stagebox. A disadvantage of constant gain is that it requires a distribution network with a higher bit rate than the audio signal to provide full resolution, eg. a 32-bit network to support constant gain for 24-bit signals.

Using a constant gain system requires a more strict understanding of the difference between digital and analogue gain functionality. In conventional digital mixing consoles, although a digital gain is always provided, the analogue gain is often used to align the analogue input level for optimal A/D conversion, as well as to set the incoming level to the suit the mixing process. In a constant gain system, the analogue gain can be used exclusively for optimization of A/D conversion (to optimise audio quality), leaving each console's digital gain to set the incoming level. More details on constant gain implications on audio quality is presented in chapter 6.3: Constant gain A/D converters.

*figure 705B: constant gain method*



## 7.4 Clip level mismatch

If analogue devices are used before or after a networked audio system's analogue inputs and outputs, such as external analogue pre-amplifiers or power amplifiers, then the corresponding terminals must be connected. If the clip levels of these terminals don't match, the level difference can reduce the system's dynamic range (in other words: waste it). As most analogue devices use buffer amp circuits to provide inputs and outputs that are independent of the connected terminal's impedance, clip levels are a fixed value - they can not be changed by the system designer or sound engineer without opening the device cabinet to perform a hardware modification.

To illustrate how clip level mismatches cause the dynamic range to be decreased, figure 706 a, b and c present the connection of the analogue input of three different power amplifiers to a Yamaha DME24N analogue output, with different clip levels of the power amplifiers input as listed in table 701.

*table 701: clip levels of the Yamaha DME24N and three different power amplifiers*

| device | terminal | clip level |
|---|---|---|
| Yamaha DME24N | output | +24 dBu |
| power amplifier P1 | input | +24 dBu |
| power amplifier P2 | input | +27 dBu |
| power amplifier P3 | input | +21 dBu |

*figure 706: $0dB_{FS}$ connection of a DME24N output to three power amplifiers with different input circuit clip levels*

*figure 706a: power amplifierP1*

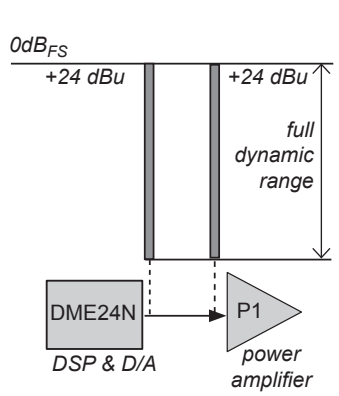



*figure 706b: power amplifier P2*

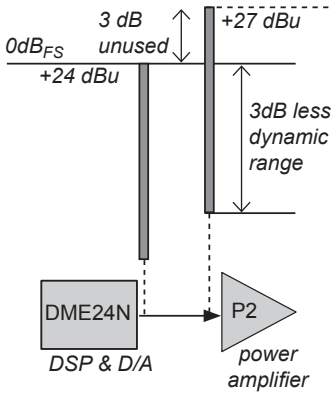
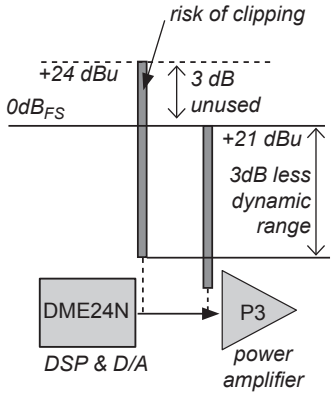



*figure 706c: power amplifier P3*




Clip level mismatches cause a system's dynamic range to be limited by as much as several dB's - raising the system's noise floor by several dB's. Specially for systems with high power loudspeakers this can be a problem - as the noise floor below $100dB_{FS}$ becomes audible for the audience situated close to the speakers. If the mismatch causes a clipping risk - as in figure 706c where the power amplifier clip level is lower than the system output's clip level - the system designer or sound engineer has to manually limit the system's output level in the digital domain.

In every networked audio system design, it is advised to check the clip level specifications at every line input (eg. wireless receivers) and output (eg. power amplifiers) of the system to know if any digital gain should be designed into the DSP processes to prevent mismatch clipping. Some devices offer internal DIP switches or jumpers with a choice of multiple clipping values. For high power high quality applications, it might be worth while to consider an external passive analogue attenuator if an internal option is not available.

## 7.5 Double A/D-D/A pass signal paths

In digital audio systems, the A/D and D/A converters and their analogue circuits are the main contributors to the system's noise floor. It makes sense for system designers to make sure that all signal paths flow through only one A/D converter and one D/A converter. If signal paths flow through such an A/D-D/A pass twice, the peak value of the noise floor is doubled, and the RMS level raises with several dB's. In the migration from analogue to digital live mixing consoles in the past 2 decades, often the connections to the speaker processing equipment remained analogue - even when the analogue processors were replaced by digital versions. In such a case, often the system's dynamic range can be improved by several dB's by simply changing the connection between the mixing console and the speaker processors from analogue to digital.

*figure 707A: single A/D-D/A pass networked audio system*

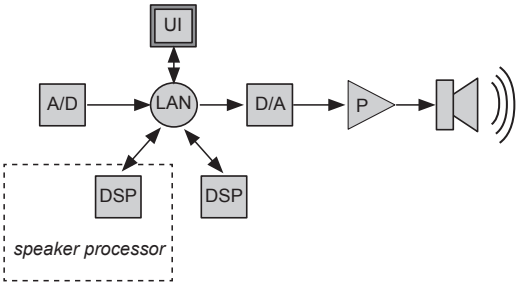
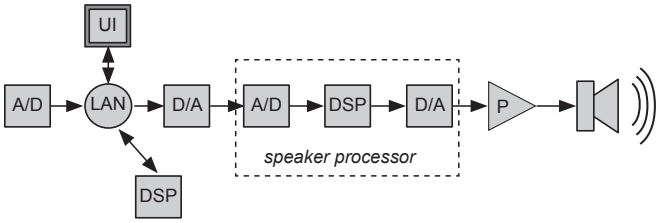



*figure 707B: double A/D-D/A pass networked audio system*




## 7.6 Unbalanced output mode

Output compression is sometimes applied in speaker processing for multi-amped speaker cabinets to protect drivers from overloading where other drivers in the same cabinet can still output more SPL. Also, full range output compression is sometimes applied to multi-cabinet speaker systems where parts of the system have spare SPL headroom. In both cases, the compression is used to make the system 'louder' - leaving out frequency ranges or locations above a certain threshold level. This could be seen as an audio quality problem - affecting a speaker system's Performance. But because it is applied as an intended process - with the goal to make the system louder at the cost of losing the designed SPL balance between loudspeakers at high volumes, it affects the Response (sound quality) - and not the Performance (audio quality).

*figure 708A: balanced mode output to a speaker system with different power ratings*

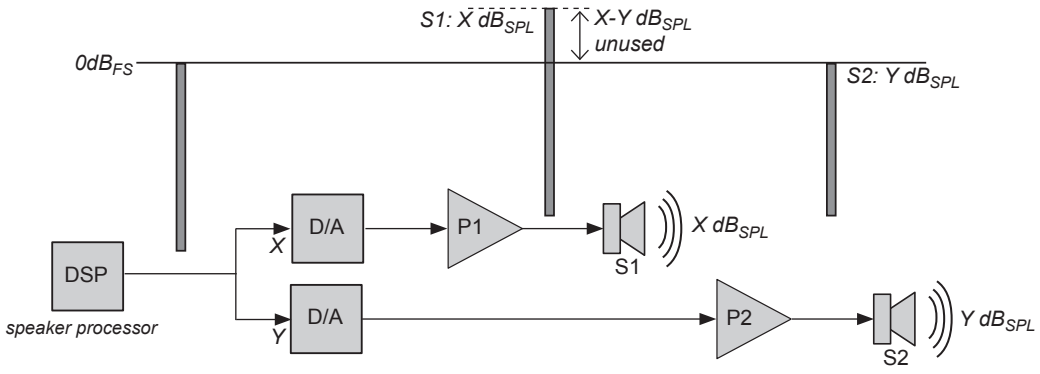
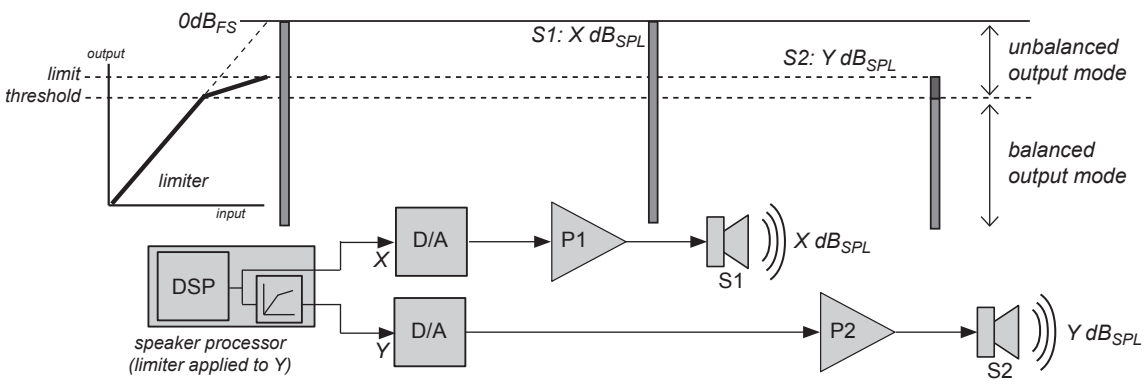



*figure 708A: unbalanced mode output to a speaker system with different power ratings*



# 8. Operational quality issues

Networking technologies introduce new audio quality issues that system designers and sound engineers have to be aware of in order to achieve a high audio quality. However, networking technologies also introduce new and exciting possibilities for system design - along with new quality issues that are not related to audio or sound, but to other fields such as costs, logistics and reliability. This chapter presents a selection of *operational quality issues*: network implications, Ethernet compliance, redundancy, and cable lengths. Similar to audio quality and sound quality, investors and rental customers can set requirements for the networked audio system to comply to operational quality requirements - with the system designers and sound engineers being responsible for compliance.

## 8.1 Network implications

In this white paper we propose the following basic definition of a network:

> *Network*
>
> *A network offers functional connections independently from a system's physical connections.*

This means that in a *networked system*, all connections in the network can be set up completely independently from the cabling. To connect a device to a network, it doesn't matter to which physical connector the device is connected. The functional connections - in the case of networked audio systems all audio connections - are established through *routing software*.

Network hardware (networked audio devices, Ethernet switches, cables) can be connected physically in three *topologies*: daisy chain, ring and star. Some network protocols only work in daisy chain or ring topologies, and some work in all topologies - but in all cases, the way the devices in a network are connected within the topology does not affect the software controlled routing. In the networked audio field, examples of networks that require a ring topology are Optocore*[8A], Rocknet*[8B] and EtherSound*[8C]. CobraNet*[8D] and Dante*[8E] support all topologies. For more detailed information on audio networks we refer to the white paper 'an introduction to networked audio systems' published on the Yamaha website*[8F].

The separation of physical cabling and functional connections in networked audio system has implications for the design and use of networked audio systems.

**Design**

In the design process of a *conventional audio system*, functionality and physical cabling are always connected - a change in a system's function requires a change in the system's physical cabling and vice versa. When designing functionality, the designer is constrained by the physical cabling possibilities. And if changes to either the functional or physical design are required to be implemented afterwards, the one is always constrained by the other.

In the design process of a *networked audio system*, the two design jobs can be separated: the functional design can be done first - without physical cabling constraints, and then the physical design - possibly even by another design engineer. The only constraint for the functional design is to stay within the network's data bandwidth - for networked audio systems usually the maximum number of channels for a single cable in the network, eg. 64 for EtherSound, 512 for Dante. Design changes afterwards can be implemented easily without changes to the cabling system - provided the system's channel count stays within the available bandwidth. This saves design labour costs and increases the speed of the design process. In many cases it also increases the reliability of the design because the separation of the two jobs make the design process less complex.

**Installation & set-up**

For the installation or set-up of *conventional audio systems*, the cable installer needs to have a detailed knowledge of the system's audio functions - as the way cables are connected affects the system's functionality. This requires experienced staff and extensive quality management procedures to include not only the physical cabling but also the audio functionality.

With *networked audio systems*, the cable installer doesn't need any knowledge on the system's audio functions - so less experienced staff can be used, and quality management procedures don't have to include the audio functionality - saving costs and set-up time.

## 8.2 Ethernet compliance

In the field of information technology (IT), Ethernet is the most prevalent world wide standard for networks. All computers, laptops, tablets and smartphones have some form of Ethernet connectivity to enable them to connect to other computers, printers, hard disks and, of course, the internet. For networked audio systems, it makes sense to include Ethernet so all of these functions can be supported additionally to the audio connections. As two other media-related fields - *video* and *lighting control* - also have embraced Ethernet as an efficient distribution protocol, a networked audio system that includes Ethernet can also include video and light control over the same cabling - increasing cost efficiency of *integrated media systems*. Finally, user interfaces increasingly utilize Ethernet, replacing *USB* and *RS232/RS485.* Examples are the many tablet *apps* available to control digital mixing consoles.

There are two ways to provide Ethernet connectivity in a networked audio system: *Ethernet embedded* systems and *Ethernet compliant* systems.

**Ethernet embedded networked audio systems**

A dedicated audio network protocol can include (or embed) an *Ethernet tunnel* that connects Ethernet ports on selected devices together as one Ethernet network. Examples are Optocore offering a 100Mb tunnel, and Rocknet offering a 10Mb tunnel. The advantage of dedicated audio protocols is that the network is fully managed by the audio designer - offering services to other functions under 'audio' supervision. The disadvantage is that the available tunnels mostly offer only unmanaged, low bandwidth Ethernet functionality.

**Ethernet compliant networked audio systems**

Audio network protocols that use Ethernet as the basic network technology offer the advantage of compatibility with appropiate Ethernet technologies and hardware on the IT market - with the choice of thousands of 'off-the-shelf' hardware components (*switches,* media converters, wireless access points). As the IT market is one of the largest markets in the world, it is also the fastest developing market - increasing bandwidth and speed constantly. Fully compliant audio protocols can make use of these developments at very low cost. Examples of low bandwidth Ethernet compliant audio network protocols are EtherSound and CobraNet - both restricted to 100Mb data rates supporting up to 64 channels per cable. An example of a high bandwidth audio network protocol is Dante - using a 1Gb data rate, supporting up to 512 channels per cable. The advantage of Ethernet compliant networked audio systems is that they offer Ethernet connectivity for any Ethernet compliant service as a standard: video, light control, stage automation systems, user interfaces and many more services can be connected without the need for extra hardware. Also, as virtually all personal computers (PC's) offer an Ethernet connection, modern Ethernet compliant audio networks such as Dante instantly allow all audio channels in the network to be recorded and played back by *Digital Audio Workstations* (or DAW's) without the need for any additional interfaces.

**Open and closed systems**

Some manufacturers of networked audio devices use a proprietary audio network protocol that is only supported by the manufacturer, and not by others. Examples are Optocore - manufactured by Optocore GmbH, and Rocknet - manufactured by Riedel GmbH. The advantage of a closed protocol is that the compatibility between protocol and hardware can be very high - as all devices come from the same manufacturer. A disadvantage of a closed protocol is that a single manufacturer can only offer a limited range of products.

To allow system designers and sound engineers to combine networked audio devices from different manufacturers, an *open* networked audio protocol can be used - most commonly licensed to many manufacturers of audio devices by a dedicated network technology company such as Cirrus Logic (CobraNet), Digigram (EtherSound) and Audinate (Dante). The advantage of an open networked audio protocol is the high design freedom for the system designer, allowing devices from multiple manufacturers to be used in a system. A disadvantage of an open protocol is the many different implementations, requiring compatibility management.

The choice of an audio network protocol is an important tool for system designers and sound engineers to achieve efficiency in the design processes, and to deliver the desired compliance to customer requirements.

## 8.3 Redundancy

Audio systems using analogue cables to connect the system's devices are insensitive to cable failures: if one single audio cable breaks, all other audio cables still work. Troubleshooting analogue cabling systems is relatively easy - as all cables can be visually inspected to run from the transmitting device to the receiving device.

With the introduction of multichannel digital audio formats, and later the introduction of networked audio protocols, high amounts of audio channels flow through only one network cable. If such a cable breaks, then all connections are lost - affecting a significant portion of the system. Troubleshooting is not as easy as with analogue cabling because it requires the use of computer software - and a member of staff who knows how to operate the software. This is why networked audio systems - and networks in general - have a redundancy protocol built in that automatically re-routes the connections to a spare cable if any cable in the system fails. All audio network protocols used in a ring or star topology offer proprietary redundancy protocols. For Ethernet compliant protocols, Ethernet itself can offer additional redundancy protocols such as link aggregation and spanning tree*[8G]

It is important for system designers and sound engineers to thoroughly know the redundancy protocols offered by the audio network protocols they are using, and to apply them when they are required. Especially for large scale live events such as corporate events, theatre productions, live concerts and broadcasting, redundancy is a key quality issue for the customer - often of equal importance to audio quality and sound quality.

## 8.4 Switches and cables

Redundancy protocols offer a solution for situations where a connection in a networked audio system fails. However, bad connectors on network switches or excessive attenuation due to long cable runs can cause intermittent connection problems that can not be solved by the redundancy protocol. System designers need to carefully study the specifications of the network interfaces, switches, connectors and cables to provide a cabling reliability that matches the customer's reliability requirements.

Excessive copper (CAT5E, CAT6) cable lengths are a common cause of intermittent connection problems. The determination of allowable lengths is a problem for system designers because many cable manufacturers do not provide maximum length specifications. Depending on the cable and connector quality, the maximum length for CAT5E and CAT6 cables is 100 meters. For mobile use, wear and tear of the cables and connectors, and the use of patch panels, might support shorter cable lengths. For longer distances, the use of fibre cabling can be considered.

For networked audio systems that are installed many times on a temporary basis - such as outside broadcast and live touring systems, road proof network switches and rugged connectors and cables can be used to provide a high reliability. Examples of rugged network connectivity systems for the audio industry are Neutrik EtherCon*[8H], Neutrik opticalCon*[8I], and Connex Fiberfox*[8J].

# 9. Quality assessment methods

In chapters 1 and 3 of this white paper, definitions are proposed to support discussions on audio quality and sound quality. Chapter 8 presented a selection of issues on operational quality. For discussions to make sense, in addition to definitions, information about the audio system is required - including judgement statements on the audio quality, sound quality and operational quality. Operational quality is generally covered by comparing the published system devices and the system design's operational specifications to the operational requirements. For the audio quality and sound quality however, things are a little more complicated. The assessment of both audio quality and sound quality is a subject of discussions in the professional audio market*[9A]. Contributing to the discussions from a 'Performance & Response' viewpoint, this chapter presents two basic methods for the quality assessment of an audio system: analysis of electrical measurements, and listening.

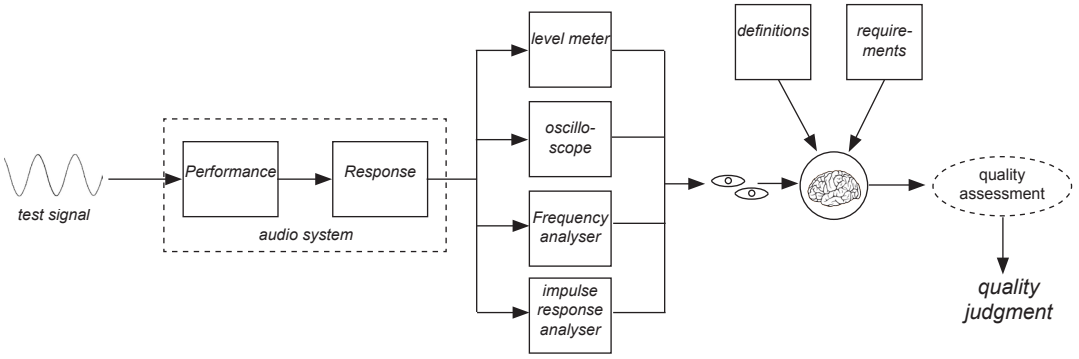
## 9.1 Quality assessment through electronic measurements

To assess the *audio quality* of a networked audio system, the *Performance* of the system can be measured by electronic measurement equipment such as level meters, oscilloscopes, FFT analysers and impulse response analysers. Because the requirements for the Performance of a system are so clearly defined - see chapter 1, the measurements can be analysed and interpreted using strict definitions and specifications. In cases where definitions and measurements of a found audio quality problem don't exist yet, a new measurement method has to be invented. But it is assumed that - after more than a century of world wide research in the field of electronic sound reproduction systems - the majority of audio quality issues in audio systems have been defined, electronically measured and analysed. Most manufacturers of electronic equipment list the most relevant measurements in their specifications - although sometimes different definitions are used, making comparison between devices from different manufacturers difficult or even impossible.

To assess the *sound quality* of an audio system, the *Response* of the system can be measured using the same electronic measurements as used for the measurement of the system's Performance. But the result of an electronic measurement doesn't say anything about sound quality by itself - it has to be translated. A complication is that the translation of electronic measurements to (individual) hearing experiences is not standardized. For example: a measured equalizer curve can be interpreted as 'good sounding' by one individual, and 'bad sounding' by the next. To obtain an 'average audience' translation matrix of physical sound characteristics to perceptional sound characteristics, clinical research is required using a large population of listeners. In contradiction to electronic Performance measurements, research on the translation of electronic Response measurements to sound quality perception has not been conducted on large scale yet*[9B]. Instead, the sound quality (Response) of many audio devices (or DSP algorithms - '*plug-ins*') are most commonly referenced to individual opinion leaders in the professional audio field, articles of respected journalists in professional audio magazines, or to the overall sound quality image of the manufacturer. Of course, also many individual listening sessions are conducted to assess the sound quality of audio devices, but then there are many complications involving the translation of the hearing sensation to the device's sound quality. This topic is presented in chapter 9.2.

Electronic measurements of analogue and digital systems is not difficult as both analogue and digital measuring equipment is widely available. The measurement of the acoustic parts of a system can be performed by the same equipment, but it requires a calibrated measurement microphone to transform acoustic signals into analogue signals. Often, individual parts of the system can be measured by either probing internal circuits in the device, or by bypassing parts of the system. A special case of bypassing is the measurement of the Performance of a system with the Response of the system bypassed - eg. switching off all processing. A big advantage of electronic measurements is that the system can be measured using a controlled test signal - making the measurement independent of the sound source, and also making it possible to reproduce the measurement at different times and locations for confirmation or to obtain a high statistical significance.

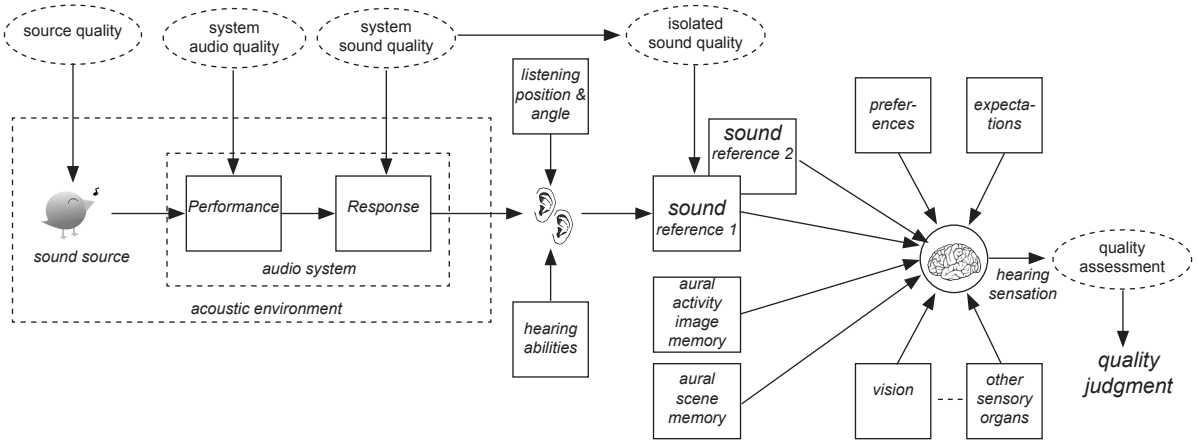
*figure 901: electronic measurement of a system's quality*



## 9.2 Quality assessment through listening tests

When it comes to the assessment of audio quality and sound quality of a networked audio system through listening (assessing the quality of the invoked hearing sensation), the human auditory system makes things extremely difficult: a hearing sensation is affected by the quality of the sound source, acoustic environment, listening position and angle, the individual's hearing abilities, preferences, expectations, short term aural activity image memory, long term aural scene memory, and all other sensory inputs of the human body such as vision, taste, smell, touch. For this reason, quality judgements based on hearing sensations have to be analysed for all factors that play a role in the hearing experience before any quality statement can be extracted about the audio system. Figure 902 presents a selection of the most important factors that play a role in the result of a listening session, with selected factors discussed in the remainder of this chapter..

*figure 902: quality assessment factors in a listening session.*




### Aural scene memory

As the definition of quality is 'conformance to requirements', always both the measurement - in the case of a listening session the hearing sensation - and the requirement are needed to come to a quality assessment. The problem is that for hearing sensations, requirements do not really exist. Every individual has different preferences or expectations for a hearing sensation, only an average hearing sensation can be assumed as requirement - at this moment only available indirectly through the sales results of theatre tickets, CD's and music downloads. This data is not very reliable, as it is heavily biased with other factors not even included in figure 902, such as social peer pressure, culture, commercial factors.
The only compatible and relevant reference for quality assessment are previous hearing sensations stored in the brain's memory as aural scenes. However, these are only relevant if they are undergone at exactly the same conditions as the conditions in the listening session. Any deviation in the environment - from all factors in figure 902 - make the memory invalid for reference. A fundamental conclusion - based on the simplified auditory processing model presented in figure 412 in chapter 4.3 - is that the assessment of the quality of a sound system can never be achieved by a single listening session because the aural scene memory is virtually always invalid. Listening tests always have to be performed in two or more sessions to create a reference and allow differential analysis to come to a quality assessment.

### Aural activity image memory

In the auditory processing model, long term aural scene memory and short term aural activity image memory are presented. Because multiple listening sessions can only be conducted one-by-one, the reference used for differential analysis always uses memory. In case of long listening sessions, only the overall aural scene can be used for comparison - as it is stored in long term memory, allowing only aggregated quality assessments. If detailed quality assessments are required, the brain's short term aural activity image memory of 20 seconds has to be used - which means that comparisons of hearing sensations have to take place within 20 seconds - using identical sound source signals. Listening sessions performed by switching between two situations while listening to an integral piece of music are not valid, as then the 20 seconds before and after the switch are never the same. The conclusion is that audio fragments used in listening sessions for detailed quality assessment have to be identical pieces of sound, shorter than 20 seconds.

**Acoustic environment, listening position and listening angle**

Listening sessions always take place in an acoustic environment. The environment can be a live space such as a concert hall, or a carefully designed acoustically optimized control room in a music studio. Only listening sessions performed in 'dry rooms' in auditory research laboratories, or a desolated open space without any wind such as in a desert on a structure high above the ground, can cancel out the acoustic environment. Apart from the acoustic environment, the listening position and angle significantly affects the hearing sensation - already a displacement between multiple listening sessions of several millimetres or degrees can change the result drastically. Two conclusions can be drawn: first, listening sessions should be undergone strictly in the sweet spot of a speaker system - facing the same direction for every session. In case of listening sessions to assess the quality of loudspeakers, a mechanical rotating system should be used to ensure that listening position and angle are always the same. Second, the result only has relevance for the listening position and angle in the acoustic environment the listening session was performed in. In any other acoustic environment the quality assessments obtained from the listening sessions are no longer valid. For live systems, the quality results are never valid because the audience in most cases consists of more than one person, who can not be located at the same listening position at the same time.

**Anticipation**

The anticipation of a result can cause test subjects to experience the result, even if there is none. Because of this *placebo effect*, all clinical trials in the medical field include two groups of test subjects, one receiving the drug under test, and the other receiving the placebo. The placebo effect also plays a role in listening tests - when a change is *anticipated*, even if there is no change, a change might somehow be detected. This is not a shortcoming of the listener, anticipation is simply one of the many factors that affect a hearing sensation - in fact, anticipation can significantly amplify the pleasantness of the hearing sensation - a concept used in many music compositions and performances. To assess the quality of a sound system however, anticipation should be avoided to prevent it from affecting the quality assessment. The simplest way to achieve this is to perform blind tests - with the test subject knowing that the audio fragment can be either different or the same.

**Expectation**

If listening tests are conducted sighted instead of blind, the test subjects can be influenced by non-auditory signals, along with previous experiences associated with those signals. For example, seeing the mechanical construction of test objects can create an expectation of the listening experience, eg. large loudspeakers are expected to reproduce low frequencies very well. Of course, knowing the product brand and remembering the brand's reputation also strongly affects the test results, rendering the outcome invalid*[9C].

**Vision & other sensory organs**

In many cases, the brain's processing of audio signals is affected and sometimes overruled by other sensory inputs - a famous example is the McGurk Ba-Ga test described in chapter 4.3. If in listening sessions the non-audio sensory inputs differ the outcome can be completely different. Even things often thought of as completely irrelevant for audio, such as the colour of cabinets and cables, or even the colour of the visual signals used to indicate sources in a listening test, can affect the quality assessment. For a series of relevant listening tests, the non-audio environment should be as constant as possible - eg. constant colours and temperature. For this reason, the consumption of food and drinks - constituting smell and taste - shortly before and during listening sessions should be avoided.

**Sound source**

All hearing sensations are affected by the quality of the sound source as described in chapter 1. To assess the quality of a system, a reference is required to compare results with the same sound source - thus ruling out the quality of the sound source. This is easy using pre-recorded materials - high resolution (eg. 24 bit 96 kHz) audio recordings can be used. It is impossible to assess a system's quality in multiple listening sessions using real-life musicians - as the musicians will never play two music pieces exactly the same. This causes the references in the aural memory to differ because of the sound source quality, and not the audio system quality. When using pre-recorded sound source materials, it is important to know if the test subject (listener) is familiar with the material because that would allow the aural scene memory (with scenes most probably generated under different conditions) to affect the new hearing sensations.

**Calibration**

To allow differential analysis of a listening session comparing a single parameter or process in a system (for example comparing one signal chain with an equaliser applied and one without), all other processes in the signal chains have to be exactly the same. When two physically different analogue devices are used, this is never the case - the gain error alone can cause up to 4 dBu level difference in case of a mixing console - significantly affecting the hearing sensation as louder signals are most commonly perceived as better sounding. This can be eliminated partially by calibrating all signal chains in the listening test to produce the same output volume. As the human auditory system is capable of detecting level differences down to 0.5 dB, listening test systems need to be calibrated within 0.5 dB or lower.

**Assessment of preferences vs. assessment of detection thresholds**

Listening tests can be performed to assess the preferences of test subjects when the differences between audio systems are high. With small differences however, it becomes increasingly difficult to assess preferences - in that case, first an assessment of detection thresholds can be performed. For this purpose, ABX testing is an accepted method - featuring blind listening to two situations A and B, then confronting the test subject with an unknown situation X, which can be either A or B. Performing an ABX test multiple times gives a statistically significant statement on whether the difference could be detected or not.

**Training**

Training strongly affects the result of listening tests. Trained listeners have learned to extract detailed information from the aural activity scene information and keep it in long term memory, not only remembering more details than untrained listeners, but also being able to report the results better - knowing the psycho-acoustic vocabulary.

Listening to audio sources as form of short-term training in AAAB test sequences introduces a preference bias as the test persons get accustomed to the A source, and might perceive the B source as less preferred. This makes AAAB tests unsuited for preference tests. For detection tests however, AAAB tests can be suited if the difference between objects are extremely small.

## 9.3 Conducting listening tests

To achieve a relevant quality assessment about the audio quality and sound quality of an audio system, we propose the following conditions to be met for valid *listening tests*:

*table 901: conditions for valid listening tests*

1) Tests must be controlled: all factors other than the audio system must be either removed or kept constant:
    * sound source (live musicians can not be used)
    * acoustic environment
    * listening position & angle
    * visible environment
    * temperature and humidity
    * smell and taste

2) at least two *listening sessions* must be performed per listening test to allow differential analysis.
    * *A single session referencing to memory is not valid.*

3) tests must be blind
    * The test subjects must not know to what reference they are listening to

4) audio materials must be shorter than 20 seconds

5) If different signal chains are used, their total gain must be calibrated within 0.5 dB

**Significance**

The abilities and characteristics of the human auditory system differ strongly from individual to individual, but also over time. Single listening tests (with multiple sessions) only provide a quality assessment of a system that is valid only for the test subject at the time of the test. To achieve statistical significance in order to generate statements that are valid for an average audience at all times, listening tests and sessions can be performed multiple times, applying general scientific statistical principles (eg. analysis of variance, $\chi^2$ tests)

**Analysis - statements on audio quality and sound quality**

The results from valid listening tests can identify audio quality issues in the Performance processes of a system, and sound quality issues in the Response processes of a system. However, there is no translation table available to translate hearing sensations to physical phenomenon in a system's circuits or software. Statements on physical phenomenon can not be made based only on listening test results. At best, electronic measurements can be proposed - based on listening test results - to find a possible physical cause of the perceived quality issue. Only if a physical cause can be confirmed, a valid statement can be made correlating the hearing experience to the physical phenomenon. All assuming that the listening tests were 'controlled' - conducted under the conditions proposed in table 901.

In this white paper, we strongly advise *not* to draw direct conclusions about physical phenomenon in networked audio systems based on listening tests. A valid conclusion can only be drawn after confirming a found cause for the listening tests results - normally by conducting further listening tests varying the found cause parameters. We even more strongly *not* advise to draw any conclusion in general based on uncontrolled listening tests.

note: besides references to scientific publications, this reference list includes references to Wikipedia printed in *italic* font. Although the information in the Wikipedia lemma's is not guaranteed to be consistent, it provides a very accessible source of information - with further references at the bottom of each article. As a thank you, the author has made a financial contribution to Wikipedia.org.

references in chapter 1:Audio Quality


| | | |
|---|---|---|
| 1A | Quality | Quality is free, Phill B. Crosby, ISBN0070145121, McGraw-Hill, Inc. |


references in chapter 2: Networked audio systems


| | | |
|---|---|---|
| 2A | Power amp class AB, D, EEEngine | http://www.yamahaproaudio.com/global/en/training_support/selftraining/technology/eeengine.jsp |


references in chapter 4:The human auditory system


| | | |
|---|---|---|
| 4A | picture: anatomy of the human ear | *http://en.wikipedia.org/wiki/Ear (file: Anatomy of the human ear.svg)* |
| 4B | picture: cochlea cross section | *http://en.wikipedia.org/wiki/Cochlea (file: Cochlea-crosssection.svg)* |
| 4C | outer hair cells | Auditory neuroscience, Schnupp et al, P73.<br>Also: *http://en.wikipedia.org/wiki/Hair_cell (Outer hair cells - acoustic pre-amplifiers)* |
| 4D | cochlear nerve | *http://en.wikipedia.org/wiki/Cochlear_nerve (anatomy and connections)* |
| 4E | hair cells | Fundamentals of Hearing,m W.A. Yost, p88.<br>also: *http://en.wikipedia.org/wiki/Auditory_system (hair cell)* |
| 4F | neurons | *http://en.wikipedia.org/wiki/Cochlear _nerve (types of neurons)* |
| 4G | equal loudness contour ISO226 | *http://en.wikipedia.org/wiki/Equal_loudness_contour* |
| 4H | Tinitus | *http://en.wikipedia.org/wiki/Tinitus (pathophysiology)* |
| 4I | hearing damage | directive 2003/10/EC of the European Parliament and of the Council of 6 Februari 2003<br>http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2003:042:0038:0044:EN:PDF |
| | ISO1990:1999 | can be purchased from ISO.ORG. |
| 4J | threshold of pain | *http://en.wikipedia.org/wiki/Threshold_of_pain* |
| 4K | Kunchur papers | http://www.physics.sc.edu/~kunchur/papers/Audibility-of-time-misalignment-of-acoustic-signals---Kunchur.pdf<br>and http://www.physics.sc.edu/~kunchur/papers/Temporal-resolution-by-bandwidth-restriction--Kunchur.pdf |
| 4L | echoic memory | *http://en.wikipedia.org/wiki/Echoic_memory (Overview)* |
| 4M | neuron connectivity | *http://en.wikipedia.org/wiki/Neuron (connectivity)* |
| 4N | simplified auditory processing model | based on 'Binaural signal processing', Jens Blauert & Jonas Braash, Ruhr university, Bochum, Germany. pdf available at IEEEXplore.ieee.org |
| 4O | Barkhausen, phon, bark | Fastl & Zwicker, Psychoacoustics<br>http://en.wikipedia.org/wiki/Bark_scale |
| 4P | masking | Fastl & Zwicker, Psychoacoustics<br>*http://en.wikipedia.org/wiki/Auditory_masking (similar frequencies)* |
| 4Q,R | sharpness, roughness | Fastl & Zwicker, Psychoacoustics |
| 4S | localization | Auditory neuroscience, Jan Schnupp, chapter 5 |
| 4T | Haas effect | *http://en.wikipedia.org/wiki/Haas_effect* |
| 4U | AFC | *www.yamahaproaudio.com/afc* |
| 4V | visual environment | Hearing lips and seeing voices, McGurk H., MacDonald J. Nature 264 (1976): p746–p748. pdf available at www.nature.com<br>also: *http://en.wikipedia.org/wiki/Sound_localization* |
| 4R | auditory masking | *http://en.wikipedia.org/wiki/Auditory_masking (similar frequencies)* |


references in chapter 5: sampling


| | | |
|---|---|---|
| 5A | carbon microphone | *http://en.wikipedia.org/wiki/Carbon_microphone* |
| 5B | Victor Orthophonic Victrola | *http://en.wikipedia.org/wiki/Victor_Orthophonic_Victrola* |
| 5C | Magnetophon | *http://en.wikipedia.org/wiki/Magnetophon* |
| 5D | Compact Cassette | *http://en.wikipedia.org/wiki/Compact_cassette* |
| 5E | PCM | *http://en.wikipedia.org/wiki/Pulse-code_modulation* |
| 5F | CD | *http://en.wikipedia.org/wiki/Compact_disc* |
| 5G | users manual specifications: DMP7 | www.yamahaproaudio.com |
| 5H | users manual specifications: 02R | www.yamahaproaudio.com |
| 5I | analogue mixer specifications | users manual specifications & specifications sheets:<br>Midas XL4: www.midasconsoles.com,<br>Soundcraft series 5: www.soundcraft.com<br>Yamaha PM4000: www.yamahaproaudio.com |
| 5J | A/D converter | *http://en.wikipedia.org/wiki/Analog-to-digital_converter* |
| 5K | MAC operation | *http://en.wikipedia.org/wiki/Multiply–accumulate_operation* |
| 5L | 6dB per bit | Taking the Mystery out of the Infamous Formula, "SNR = 6.02N + 1.76dB,"<br>http://www.analog.com/static/imported-files/tutorials/MT-001.pdf |
| 5M | dither | *http://en.wikipedia.org/wiki/Dither* |
| 5N | dynamic range of digital systems | users manual specifications & specification sheets<br>DigiCo SD8: www.digiconsoles.com<br>Avid SC48: www.avid.com<br>Yamaha M7CL: www.yamahaproaudio.com |
| 5O | Nyquist-Shannon | *http://en.wikipedia.org/wiki/Nyquist–Shannon_sampling_theorem* |
| 5P | oversampling | *http://en.wikipedia.org/wiki/Oversampling* |
| 5Q | speed of electricity | *http://en.wikipedia.org/wiki/Speed_of_electricity* |

| | | |
|---|---|---|
| 5R | analog switch data sheet | Datasheet 74HC4053 |
| 5S | jitter | Jitter: specification and assessment in digital audio equipment, Julian Dunn, Cambridge presentation AES 93rd convention, 1992. Available through www.aes.org |
| 5T | PTP | http://en.wikipedia.org/wiki/Precision_Time_Protocol |
| 5U | sinusoidal jitter audibility | Theoretical and Audible Effects of Jitter on Digital Audio Quality, Eric Benjamin and Benjamin Gammon, Dolby Laboraties inc presentation AES 105th convention, 1998. Available through www.aes.org |
| 5V | noise shaped jitter audibility | Detection threshold for distortions due to jitter on digital audio, Kaori Ashihara, Shogo Kiryu et al, National Institute of Advanced Industrial Science and Technology - Acoustical Science and Technology 26, 1 (2005) |

references in chapter 6: distribution & DSP

| | | |
|---|---|---|
| 6A | Dante 32-bit | http://dev.audinate.com/kb/webhelp/content/yamaha/clseries/the_yamaha_cl_series_consoles_and_io_racks_use__the_new_dante_32-bit_mode_of_operation_.htm |
| 6B | Moore's law | *http://en.wikipedia.org/wiki/Moore's_law* |
| 6C | Native | *http://en.wikipedia.org/Native_processing* |
| 6D | Motorola 56K series | *http://en.wikipedia.org/wiki/Motorola_56000* |
| 6E | Analog Devices SHARC | Analog Devices - getting started with SHARC, http://www.analog.com/static/imported-files/tech_docs/GettingStartedwithSharcProcessors.pdf |
| 6F | Texas Instruments | http://ti.com/lsds/ti/dsp/c6000_dsp/c674x/products.page |

references in chapter 7: level issues

| | | |
|---|---|---|
| 7A | AES2 1984 (R2003) | available through www.aes.org |

references in chapter 8: operational quality issues

| | | |
|---|---|---|
| 8A | Optocore | http://www.optocore.com/downloads/pdf/Optocore_Basics_cabling.pdf |
| 8B | Riedel Rocknet | http://www.riedel.net/AudioSolutions/RockNetOverview/AboutRockNet/tabid/502/language/en-US/Default.aspx |
| 8C | EtherSound | http://www.EtherSound.com/ |
| 8D | CobraNet | http://www.cobranet.info/ |
| 8E | Dante | http://www.audinate.com/index.php?option=com_content&view=article&id=93&Itemid=93 |
| 8F | audio networks | 'an introduction to networked audio systems', Ron Bakker, http://download.yamaha.com/file/47399 |
| 8G | AVB | http://en.wikipedia.org/wiki/Audio_Video_Bridging |
| 8H | Cobranet system design | 'networked audio system design with CobraNet' http://download.yamaha.com/file/47405 |
| | EtherSound system design | 'networked audio system design with EtherSound' http://download.yamaha.com/file/47411 |
| 8I | Neutrik Ethercon | http://www.neutrik.com/en/ethercon/ |
| 8J | Neutrik opticalcon | http://www.neutrik.com/en/opticalcon/ |
| 8K | Connex Fiberfox | http://www.fiberfox.com/fiberfox_ebene.htm |

references in chapter 9: quality assessment methods

| | | |
|---|---|---|
| 9A | audio quality assessment | 'understanding what really matters with audio reproduction and what not', Ethan Winer, workshop AES 38th convention 2009, pdf available through www.aes.org |
| 9B | translation of listening tests | 'measurement and perception of quality in sound systems, G.R. Thurmond,11th AES international convention 1992, pdf available through www.aes.org |
| 9C | sighted listening | Hearing is believing vs. Believing is Hearing: Blind vs. Sighted Listening Tests, and Other Interesting Things, Floyd E. Toole and Sean E. Olive, Harman International Industries, Inc. international convention 1997, pdf available at www.aes.org, |

further reading:

books:

| | |
|---|---|
| The Art of Digital Audio - John Watkinson, | ISBN 0-240-51320-7 |
| William A. Yost, Fundamentals of Hearing | ISBN 0-12-370473-1 |
| Auditory Neuroscience - Jan Schnupp, Israel Nelken, Andrew King: | ISBN 978-0-262-11318-2 |
| Psychoacoustics - Hugo Fastl, Eberhard Zwicker | ISBN 987-3-540-23159-2 |

white papers and technical publications:

| | |
|---|---|
| an introduction to networked audio systems, Yamaha | http://download.yamaha.com/file/47399 |
| networked audio system design with CobraNet, Yamaha | http://download.yamaha.com/file/47405 |
| networked audio system design with EtherSound, Yamaha | http://download.yamaha.com/file/47411 |