Putting the Science Back into Loudspeakers

John Watkinson

The rapid developments in microelectronics and associated signal processing techniques have made possible DVD and Digital Television Broadcasting with multi-channel sound. In comparison with these developments, the loudspeaker has all but stood still. The understanding of the human hearing system and its requirements has never been better, yet the traditional loudspeaker continues to ignore those requirements. Worse still, the topic of loudspeakers has become a squalid pit of pseudoscience and subjectivism whereas the product has become a commodity to be made as cheaply as possible.

We have reached the unsupportable position where loudspeakers are frequently not good enough to assess the quality of other audio components. This leads to flawed experiments whose outcomes are unreliable. The most serious effect has been the widespread adoption of lossy audio coding algorithms. This must be at least partly due to the fact that their flaws cannot be heard on the legacy loudspeaker.

In this paper, I propose to show that all that is needed to demonstrate dramatic improvements in loudspeaker realism is a scientific approach based on a thorough study of the human auditory system. This alone defines the performance criteria and meeting those criteria is just a matter of engineering rigor.

Information theory has been of great service to the engineering community in predicting what will happen in real signal paths or channels, leading to the design of higher performance systems. In my work on loudspeakers I have used information theory as a tool to understand what is happening and, not surprisingly, found it extremely relevant. What is surprising is that the approach appears not to have been used earlier. This may be because a scientific approach cuts across the traditional empiricism.

Traditional loudspeaker measurements are disreputable. It is widely known that several speakers having the same measurements can sound quite different. To a scientist this can only mean that the measurements are either insufficiently accurate or incomplete. Little wonder that extensive subjective assessment is traditionally thought necessary in the development of a loudspeaker.

On the introduction of stereophonic sound there was no parallel introduction of a unit in which the accuracy of the stereophonic image could be expressed. With the subsequent development of surround sound there is still no such unit. In my view stereophonic and surround sound reproduction should pay as much attention to the spatial accuracy as it does to the traditional aspects such as frequency response. We know from psychoacoustics how accurately the human hearing system can resolve direction, and if the loudspeaker industry were serious about quality, there would be an accepted method of testing directional realism, a unit, and an agreed threshold for high quality.

However, this has not been done. In the absence of a measuring technique and a unit, the spatial accuracy of loudspeakers is substantially neglected despite being a major contributor to lack of realism. Those who consider it important are thwarted by the lack of a unit by which different designs could be compared. Imaging accuracy has to be described as best one can in plain language and this is simply unscientific. Imagine the state of the astronomy community today if a unit of resolution had not been available to measure the imaging performance of telescopes.

Another area in which loudspeakers are disreputable is in the neglect of the time domain. The traditional view is that all that matters is to be able to reproduce continuous sine waves over the range of human hearing.

A very small amount of research and thought will reveal that this is a misguided view. Frequency response is important, but not so important that the attainment of an ideal response should be to the detriment of realism. One tires of hearing that "phase doesn't matter" in audio or "the ear is phase deaf". These are outmoded views which were reached long ago in flawed experiments and which are at variance with the results of recent psychoacoustic research.

The ear works in two distinct ways, which it moves between in order to obtain the best outcome from the fundamental limits due to the Heisenberg inequality. The Heisenberg inequality states that as frequency resolution goes up, time resolution goes down and vice versa. Real sounds are not continuous, but contain starting transients. During such transients, the ear works in the time domain. Before the listener is conscious of a sound, the time domain analysis has compared the time of arrival of the transient at the two ears and established the direction. Following the production of a transient pressure step by a real sound source, the sound pressure must equalise back to ambient.

The rate at which this happens is a function of the physical size of the source. The ear, again acting in the time domain, can measure the relaxation time and assess the size of the source. Thus before any sound is perceived, the mental model has been told of the location and size of a sound source.

In fact this was the first use of hearing, as a means of perceiving a threat in order to survive. Frequency analysis in hearing, consistent with the evolution of speech and music came much later. After the analysis of the initial transient, the ear switches over to working in the frequency domain in order to analyses timbre. In this mode, the mode that will be used on steady state signals, phase is not very important. However, the recognition of the initial transient and the relaxation time are critical for realism. Anything in a sound reproduction system which corrupts the initial transient is detrimental.

Whilst audio electronics can accurately handle transients, the traditional loudspeaker destroys both the transient and the relaxation time measurement. Lack of attention to the time domain in crossover networks leads to loudspeakers which reproduce a single input step as a series of steps, one for each drive unit at different times. The use of resonance in reflex cabinets masks the relaxation time in the audio signal. Instead the relaxation time of the loudspeaker is superimposed on the transient to that all sounds appear to come from sources whose size is the size of the loudspeaker, not the size of the original source.

Active crossover techniques which avoid these difficulties are known, but have largely been ignored in today's loudspeakers. The reflex loudspeaker technique was a reasonable one when audio amplification and signal processing was expensive, and when the damage done to realism was not understood. Today its problems are known, and superior alternatives are known, but the reflex speaker continues out of pure tradition. The number of loudspeakers which combine linear phase low frequency reproduction with time-accurate crossovers is very small, but only with this approach can the initial transient and relaxation time be correctly reproduced. Most people have never heard such precision, but upon hearing it there is universal agreement that the degree of realism is enhanced.

Considering information theory, a steady state sine wave carries no information because it has no bandwidth and any one cycle is predictable from the one before. Only transients have bandwidth and contain information because they are unpredictable. It follows that a speaker which is optimized to reproduce steady state sine waves does not necessarily have adequate information capacity.

A loudspeaker can be modeled as an information channel of finite capacity, which can actually be measured as an equivalent bit rate. This equivalent bit rate relates to the realism which the speaker can achieve.

When the speaker information capacity is limited, the presence of another restriction in the signal being monitored may go unheard and it may erroneously be assumed that the signal is ideal when in fact it is not.

The use of poor loudspeakers simply enables other poor audio devices to enter use. When loudspeakers have such poor image forming or directional capabilities, how can they be used to assess these capabilities in microphones? This has led to the widespread use of spaced microphones for stereo recording. There is no scientific explanation for how spaced microphones reproduce a virtual image between a pair of loudspeakers, and so it should come as no surprise that there is no agreement on what the spacing between the microphones should be. What is certain is that spaced microphones give contradicting information to the listener. The initial transient is reproduced in one place, whereas the steady state sound is reproduced in a range of locations depending on frequency.

What can be said with confidence is that the greater the information capacity, realism or spatial accuracy of the loudspeakers, the less favourably will a spaced microphone recording compare with one made with coincident or soundfield techniques.

The same is true for compressors or bit-rate reducers. It follows that codecs can only meaningfully be assessed on speakers of adequate information capacity. It also follows that the definition of a high quality speaker is one which readily reveals compression artifacts. The only audio quality criteria we have for sound reproduction is that performance actually meets psychoacoustic requirements.

The appropriate criteria can only be found by subjective tests. Consequently I have been doing some research into stereo perception. This has led to some interesting conclusions, particularly regarding loudspeakers and compressors, which turn out to be related.

I have found that non-ideal loudspeakers act like compressors in that the distortions, delayed resonance and delayed re-radiation they create conceal or mask information in the original audio signal. If a real compressor is tested with non-ideal loudspeakers certain deficiencies of the compressor will not be heard. Others, notably the late Michael Gerzon, have suggested that compression artifacts which are inaudible in mono may be audible in stereo. I have found this to be true. I have also found that the spatial compression of non-ideal stereo loudspeakers conceals real spatial compression artifacts.

The ear is a lossy device because it exhibits masking. Not all of the presented sound is sensed. If a lossy loudspeaker is designed to a high standard, the losses may be contained to areas which are masked by the ear and then that loudspeaker would be judged transparent. Douglas Self has introduced the term "blameless" for a device whose imperfections are undetectable; an approach which commands respect. However, the majority of legacy loudspeakers are not in this category. Audible defects are introduced into the reproduced sound in frequency, time and spatial domains, giving the loudspeaker a kind of character which is best described as a signature or footprint.

Lossy compression does not preserve the original waveform and seeks to be blameless by placing the noises where they will be masked. Naturally one would want to carry out listening tests to see if this goal had been achieved. If blameless loudspeakers are used, the test is valid. However, the legacy loudspeaker is not blameless and does not preserve the waveform either. When a loudspeaker has a signature, how are we to know that compression artifacts are not being masked by the speaker signature? When listening in series, as we must, on hearing a deficiency, how are we to determine whether this was the codec or the speaker?.

The hearing mechanism has an ability to concentrate on one of many simultaneous sound sources based on direction. The brain appears to be able to insert a controllable time delay in the nerve signals from one ear with respect to the other so that when sound arrives from a given direction the nerve signals from both ears are coherent causing the binaural threshold of hearing to be 3 - 6 dB better than monaural at around 4 kHz. Sounds arriving from other directions are incoherent and are heard less well. This is known as attentional selectivity.

Human hearing can also locate a number of different sound sources simultaneously presented by constantly comparing excitation patterns from the two ears with different delays. Strong correlation will be found where the delay corresponds to the interaural delay for a given source. This delay-varying mechanism will take time and the ear is slow to react to changes in source direction. Oscillating sources can only be tracked up to 2 - 3 Hz and the ability to locate bursts of noise improves with burst duration up to about 700 milliseconds. Location accuracy is finite.

Stereophonic systems should allow attentional selectivity to function such that the listener can concentrate on specific sound sources in a reproduced stereophonic image with the same facility as in the original sound.

We live in a reverberant world which is filled with sound reflections. If we could separately distinguish every different reflection in a reverberant room we would hear a confusing cacophony. In practice we hear very well in reverberant surroundings, far better than microphones can, because of the transform nature of the ear and the way in which the brain processes nerve signals. Because the ear has finite frequency discrimination ability in the form of critical bands, it must also have finite temporal discrimination.

This is good news for the loudspeaker designer because the ear has finite accuracy in frequency, time and spatial domains. This means that a blameless loudspeaker is not just a concept, it could be made real by the application of sufficient rigour.

When two or more versions of a sound arrive at the ear, provided they fall within a time span of about 30 milliseconds, they will not be treated as separate sounds, but will be fused into one sound. Only when the time separation reaches 50 - 60 milliseconds do the delayed sounds appear as echoes from different directions. As we have evolved to function in reverberant surroundings, most reflections do not impair our ability to locate the source of a sound. Clearly the first version of a transient sound to reach the ears must be the one which has traveled by the shortest path and this must be the direct sound rather than a reflection. Consequently the ear has evolved to attribute source direction from the time of arrival difference at the two ears of the first version of a transient.

Versions which may arrive from elsewhere simply add to the perceived loudness but do not change the perceived location of the source unless they arrive within the inter-aural delay of about 700 microseconds when the precedence effect breaks down and the perceived direction can be pulled away from that of the first arriving source by an increase in level. This area is known as the time-intensity trading region. Once the maximum inter-aural delay is exceeded, the hearing mechanism knows that the time difference must be due to reverberation and the trading ceases to change with level.

Unfortunately reflections with delays of the order of 700 microseconds are exactly what are provided by the legacy rectangular loudspeaker with sharp corners. These reflections are due to acoustic impedance changes and if we could see sound we would double up with mirth at how ineptly the sound is being radiated. Effectively the spatial information in the audio signals is being convolved with the spatial footprint of the speaker. This has the effect of defocusing the image. Now the effect can be measured.

Intensity stereo, the type obtained with coincident mikes or panpots, works purely by amplitude differences at the two loudspeakers. The two signals should be exactly in phase. As both ears hear both speakers the result is that the space between the speakers and the ears turns the intensity differences into time of arrival differences. These give the illusion of virtual sound sources.

A virtual sound source from a panpot has zero width and on blameless speakers would appear as a virtual point source. As a result stereo reverb is added and this is audible between the point sources. A similar result is also obtained with real sources using a coincident pair of mikes. In this case the sources are the real sources and the sound between is reverb/ambience.

Upon reproducing such a stereo signal with the legacy square box speaker, the point sources have been spread by the speaker footprint so that there are almost no gaps between them, effectively masking the ambience. This represents a lack of spatial fidelity, so we can say that rectangular loudspeakers cannot faithfully reproduce a stereo image, nor can they be used for assessing the amount of reverberation added to a "dry" recording. It will be shown shortly that such speakers cannot meaningfully be used to assess compression codecs.

A compressor works by raising the level of "noise" in parts of the signal where it is believed to be masked. If this belief is correct, the compression will be inaudible. However, if the codec is tested using a signal path in which there is another masking effect taking place, the results of the test are meaningless. It is our experience both from theoretical analysis and practical measurement that legacy loudspeakers have exactly such a masking process, both temporally and spatially.

If a stereophonic system comprising a variable bit rate codec in series with a pair of speakers is considered to be a communication channel, then it will have a finite information rate in frequency, temporal and spatial domains. If this information rate exceeds the capacity of the human hearing mechanism, it will be deemed transparent.

However, in the system mentioned, either the codec or the speakers could be the limiting factor and ordinarily there would be no way to separate the effects. However, if a variable bit-rate codec is available, some conclusions can be drawn. Clearly with a very high bit rate, the speakers will be the limiting factor, whereas with a low bit rate, the codec will dominate. At some intermediate bit rate, the effects will be equal. At this point the masking due to the speaker is equal to the level of artifacts from the coder. At any lower bit rate, compression artifacts will become audible over the footprint of the speaker. The worse the information capacity of the speaker, the lower the bit rate at which the artifacts are audible.

As a result by simply varying the bit rate of a coder, it becomes possible to measure the effective bit rate of a pair of loudspeakers.

In order to test these theories, we have built a number of active loudspeakers, both electrostatic and moving coil. These exhibit minimum phase, including through the crossover region, and are free of reflections in the sub-700 microsecond trading region. Not surprisingly the imaging is much more accurate and actually reveals what is going on spatially. It is possible to resolve the individual voices in double-tracked vocals where the panpots on each track have been in slightly different places.

These speakers were used to assess some audio compressors. Even at high bit rates, corresponding to the smallest amount of compression, it was obvious that there was a difference between the original and the compressed result. The dominant sound sources were reproduced fairly accurately, but what was most striking was that the ambience and reverb between was virtually absent, making the decoded sound much drier than the original.

What was even more striking was that the same effect was apparent to the same extent with both MPEG layer 2 and Dolby AC-2 coders even though their internal workings is quite different. In retrospect this is less surprising because both are probably based on the same psychoacoustic masking model. MPEG-3 fared even worse because the bit rate is lower. Transient material had a peculiar effect whereby the ambience would come and go according to the entropy of the dominant source. A percussive note would narrow the sound stage and appear dry but afterwards the reverb level would come back up. An opportunity arose to compare the same commercially available recording on CD and MiniDisc and the MD version was obviously inferior. All of these effects largely disappeared when the signals to the speakers were added to make mono which removes the ear's ability to discriminate spatially.

The effects are not subtle and do not require "golden ears". We have successfully demonstrated these effects to an audience of about 60 in a conference room on more than one occasion; hardly the ideal listening environment, but all heard it. One of us (Watkinson) was asked in one demonstration if this was only relevant to classical recordings so the demonstration was repeated with a Bruce Springsteen recording and again all heard the difference.

We are forced to conclude that because of the phenomena described here, audio codecs have reached the market, which produce audible artifacts even at high bit rates, despite exhaustive subjective testing. When one examines the results of any subjective compression test, it becomes clear that the type of loudspeakers used would have been those having the shortcomings mentioned above. As a result these subjective tests are invalid because the masking of the legacy speakers was masking the coder being tested.

We must conclude that whilst compression may be adequate to deliver post produced audio to a consumer with mediocre loudspeakers, these results underline that it has no place in a quality production environment. When assessing codecs, loudspeakers having poor diffraction design will conceal artifacts. When mixing for a compressed delivery system, it will be necessary to include the codec in the monitor feeds so that the results can be compensated. Where high quality stereo is required, either full bit rate PCM or lossless (packing) techniques must be used.

We have also shown that audio codecs can only be developed fully if blameless monitoring loudspeakers are available and vice versa.

Using a codec to measure the bit rate of a speaker gives a direct assessment of its figure of merit. The use of this technique has had some further interesting consequences. Traditional loudspeakers use ferrite magnets for economy. However, ferrite is an insulator and so there is nothing to stop the magnetic field moving within the magnet due to the Newtonian reaction to the coil drive force. In magnetic materials the magnetic field can only move by the motion of domain walls and this is a non-linear process. The result in a conductive magnet is flux modulation and Barkhausen noise. The flux modulation and noise make the transfer function of the transducer non-linear and result in intermodulation.

The author did not initially believe the results of mathematical estimates of the magnitude of the problem, which showed that ferrite magnets cannot reach the 16-bit resolution of CD. Consequently two designs of tweeter were built, identical except for the magnet. The one with the neodymium magnet has higher resolution, approaching that of an electrostatic transducer. Such precision loudspeakers and drive units require no more than an appropriate degree of rigour during the design stage, along with some high grade circuit design, but have the advantage that there usually needs to be very little change between the prototype and the production phase.