**Current content technology: appropriate or not up to the task?**

adolfomaria.rosasgomez@telefonica.com

Madrid, March 2014.

## 1. INTRODUCTION

'Content' is a piece of information. We may add 'intended for humans'. That maybe de most defining property of content in front of other pieces of information. In a world increasingly mediated by machines it would not be appropriate to say that machines 'do not consume content', they do, on our behalf. Key idea is that any machine that processes content does it just immediately on behalf of a human consumer, so it processes content human-like. STBs, browser players, standalone video players, connected TVs,… may be deemed 'consumers' of content with a personality of their own (more on this through the article) but they are acting over information that was intended for humans. They are mimicking human information consumption models, as opposed to M2M (Machine to machine) information processing which happens in a way totally unnatural to humans.

Which is the state of the art in 'content technology'? By 'content technology' I mean: ways to represent information natural for humans, devices designed to take information from humans and provide information to humans. Are these technologies adequate today? How far are we from reaching perceptual limits? Are these technologies expensive/cheap? Do we have significantly better 'content technology' than we used to have or not?

## 2. AUDIOVISUAL

The most natural channel to feed info to humans is the combination of audio + video. Our viewing and hearing capacities work together extremely well. We humans are owners of some other sensory equipment: a fair amount of surface sensible to pressure and temperature (aka skin), an entry-level chemical lab for solids and liquids (tongue) and a definitely sub-par gas analyser (nose). As suggested by these descriptions we cannot be really proud about our perceptions of chemicals but we do fairly well perceiving and interpreting light, sound and pressure. This is not due to the quality of our sensors, which are surpassed easily by almost every creature out there, but due to the processing power connected to those sensors. We see, hear and feel through our brain. We humans have devoted more brain real state to image and sound processing than other animals.

It is no coincidence thus that technology intended to handle human information has been focused on audio visual. Relatively recently a branch of technology took off: 'Haptics', that may give our pressure sensors some fun (multi-touch surfaces, force-feedback 3D devices, gesture capture…) but 'Haptics' is still underdeveloped if we compare it to audio-visual technologies.

So we have created our information technology around audio-visual. Let's see where we are.

3. PERCEIVING LIGHT

We are able to perceive light. We develop brain interpretations linked to properties of light: intensity, wavelength (colour is our interpretation of light wavelength + composition of wavelengths). There are clear limits to our perception. We just perceive some of the frequencies. We cannot 'see' below red frequencies (infra-red) or above violet (ultra-violet). We have an array of light sensors of different types: 'cones' and 'rods' and a focus system (an organic 'lens'). Equipped with this hardware we are able to sample the light field of the outer world. I have chosen to say 'sample' because there is still no consensus about how do we process data from time varying light fields: does our brain work in entire 'frames'? Does it keep a 'static image' memory? Does it use a hierarchical scene representation? Does it sample always at the same pace (at a fixed rate)? We know very few things about how we process images. Nevertheless some of the physical limits to our viewing system come from the 'sensory hardware' not from the processor, not from the brain.

Lens + light sensor: aperture, angular resolution, viewing distance

If you browse any good biomechanics literature (ie: Richard Dawkins: Climbing mount improbable, or http://micro.magnet.fsu.edu/primer/lightandcolor/humanvisionintro.html), you can find that human eyes lens has in average an aperture of 10 deg. We form 'stereoscopic' images by overlapping information from two eyes. We have roughly 125-130 million light receptors per eye, most of them are rods unable to detect colour, and less than 5% (about 6-7 million) are cones sensible to colour. They are not laid out in a comfortable square grid so our field of view maybe represented delimited by something like an oval with its major axis horizontal. Central part of this oval is another oval with mayor axis vertical where the two vision cones overlap, this small spot is our stereoscopic-high-resolution view spot, though we still receive visual info from the resting part of the 'whole field of view' which may be 90º vert. and 120º hor.

Optical resolution can be measured in Line pairs per degree (LPD) or cycles per degree (CPD). Humans are able to resolve 0.6 LP per minute arc (1/60 Deg). We can tell two points (each one lying on one different line of a line pair of contrasting colour) are different when they are apart by more than 0.3 minute arc (see for instance www.clarkvision.com or www.normankoren.com/Tutorials/MTF.html ). Visual acuity defined as 1/a where 'a' is the response in LPD is 1.7, and 'a' is the abovementioned 0.6 LP per min arc. This is called 20/20 vision. You standing at 20 feet from target see the same detail as any normal viewer standing at 20 feet. If you have better acuity than average you can see the same detail standing farther, for instance 22/20. If you are subpar you need to get closer for instance 18/20. As a reference a falcon's eye can be as good as 20/2. Cells in the fovea (central region of view field spanning 1.4 deg from total 10 deg) are better connected to the brain: 1 cone -1 nerve and cones are more tightly packed there so spatial resolution in the fovea is higher than in the peripheral view. The ratio of connections in the peripheral view can drop as much as 1:20, which means that 20 light-receptors sum up their signals into a single signal they feed to a single nerve.

I know that this way of measuring the resolution power of our eyes is cumbersome, but by the way is the only right method! Let's do some practical math. Let's say that we read a book or we read on a tablet. Normal reading distance may be 18" = 45.7 cm. Our eyesight cone at that distance is just 8cm in base-radius (spot height). We see 0.6LPminarc*600 min arc=360LP in 8cm vertical so we can tell 720 'points' to be different in a vertical high-contrast alternating colour strip of points. This is 720 points/8cm = 90 points/cm or 228 points per inch (ppi). You may have noticed that you cannot tell two adjacent dots printed by modern laser printers (300-600 dpi) at normal reading distance. It would not be fair to say that 250 dpi suffices to print so we can read comfortably at normal distance, as far as different printing technologies may need more than 1 inkdrop (a dot) to represent a pixel. This is the reason state of the art printing moves between 300-600 dpi and it does not make much sense to go beyond. *(P.S: You may notice there are printers that offer well above that: 1200-1400 ppi, but most of them confuse 'inkdrop dots' with pixels. They cannot represent a single pixel with a single drop or dot. There are also scanners boasting as much as 4000dpi…but this is an entirely different world as it may make sense scanning a surface at much closer than viewing distance so we can correct scanner optical defects to produce a right image for normal viewing distance.)*

Assuming that display/print technology is not too bad translating pixels to dots we can say that a good surface for reading at 18"must be capable of showing no less than 250 dpi or ppi so you can take full advantage of your eyes. By these standards 'regular' computer displays are not up to the task, as they have 75-100 ppi and typically are viewed at 20" so they would require over 200 ppi. iPad Retina seems a more appropriate display as it has been designed with these magic numbers in mind. Retina devices have pixel densities from 326 ppi (phone) through 264ppi (tablet) down to 220 ppi (monitor). As the viewing distance for a phone is less than 10", 326ppi fits in the same acuity range that 264 ppi for 18" and so does 220 ppi for 20". Other display manufacturers have followed on the trail of Apple: Amazon fire HD devices have followed and then surpassed Retina displays : Kindle Fire HD 8.9" (254 ppi), Kindle Fire HDX 7" (323 ppi), Kindle Fire HDX 8.9" (339 ppi). Specially note that the HDX devices are tablets while they use pixel densities that Apple Retina reserves for phones…so these tablets are designed to fit eyes much better than standard. Newer phones like HTC One (468 ppi), Huawei Ascend D2 (443 ppi), LG Nexus 5 (445 ppi), Samsung Galaxy S4 (443 ppi)… go that same way. ( http://en.wikipedia.org/wiki/List_of_displays_by_pixel_density)

What about a Full HD TV or a 4K TV? We can calculate the optimal viewing distance for perfect eyesight. Let's do the math for a 56" Full HD TV and a 56" 4K TV assuming aspect ratio 16:9 and square pixels. The right triangle is: 16:9:18.36 which is homothetic to ~ 48.8: 27.45: 56, so in vertical 27.45" we have 1080 points (Full HD) which is 39.34 ppi or double for 4K. Let's round to 40 ppi for a 56" FullHD and 80ppi for a 56" 4K TV. So optimal viewing distance corresponds to 360 points per 5 Deg matching to 40ppi: 360/40= 9 inches viewed at 5 Deg or Distance in inches= 9/tan (5 Deg) = 103 inches= 2.6 m for Full HD and 1.3 m for 4K. So if you have a nice 56" Full HD TV you will enjoy it best by sitting closer than 2.6 m. If you are lucky enough to have a 56" 4K TV you can sit as close as 1.3 m and enjoy to the max of your eyes resolving power

Colour perception.

Humans have three types of cones (the colour receptors) each one sensible to a different range of light wavelengths: red, green, blue. Colour is NOT an objective property of light. Colour is an interpretation of two physical phenomena: 1) wavelength of radiation, 2) composition of 'pure tones' or single-wavelength radiations. A healthy eye can distinguish more than 16 M different shades of colour (some lab experiments say even as much as 50 M). As we have commented cones are scarce compared to rods so we do not have the same resolution power for colour as we have for any light presence. Pure 'tones' range from violet to red. They are called 'spectral colours'. Non spectral colours must be produced by compositing any number of pure tones. For example white, grey, pink… need to be obtained as compositions.

Colour is subjective. Within a range, different people will see slightly different shades of colour when presented with exactly the same light. (The same exact composition of wavelengths). This is due to the way cones react to light. Cones are pigmented and thus when receiving photons of a certain wavelength range their pigment reacts triggering a current to the nerve. But cone pigment 'quality' varies from human to human so they may trigger a different signal for the same stimulus and on the contrary they may trigger the same signal for a slightly different stimulus. The colour we see is an interpretation of light. It happens that different lighting conditions may render exactly the same electrical response. This means that the composition of wavelengths to produce some colour output is not unique; there are a number of input combinations that render the same output (metamers).

Light intensity range.

To make things even more difficult, the colour response (to light) function of our eyes depends on intensity of radiation. Cones may respond to the same wavelength differently when the intensity of light is much higher or much lower (bear in mind that intensity relates to the energy carried by the photons…it is crudely the number of photons reaching the cone per time unit, it has nothing to do with individual energy of each photon that solely relates to its wavelength.) Our eyes have a dynamic sensitivity range that is truly amazing, it covers 10 decades. We can discern shapes in low light receiving as less as 100/150 photons and we can still see all the way up to 10 orders of magnitude more light!!!. Of course we do not perceive colour information equally well all across the range. When we are in the lower 4 decades of the range we need to sum up all possible receptors to trigger a decent signal, so we 'see' mostly through rods (scotopic vision) and through many peripheral rods that are less individually connected to nerves so many of them share a nerve thus losing in spatial resolution but trading in sensibility as very low photon counts per receptor may excite the nerve when summed from several receptors. When we are in the 6 upper decades of intensity range we can perceive colour (photopic vision) although in extreme intensity we just perceive washed or white colours. Some authors and labs have checked the human intensity range for a single scene (see www.clakvision.com). This range is different of the whole range, as the eye is able of 10 decades but not in the same scene, only through a few minutes of adaptation to

low/high light. For a single night scene with very low light (gazing at stars for instance) the range is estimated to be 6 decades 1:10^6; for daylight the range is estimated to be 4 decades 1:10^4.

What about current display devices? Are they good to represent colour in front of our eyes?

State of the art displays consist of a grid of picture elements (pixels) each one formed by three light emitting devices selected to be pure tones: R, G, B. The amount of light that is emitted by each device can be controlled independently by polarizing a Liquid Crystal with a variable voltage, allowing more or less light to go through. The polarization range is discretized to N steps by feeding a digital signal through a DAC to the LC. Superposition of light from three very closely placed light emitters produces a mix of wavelengths, a colour shade, concentrated on one pixel.

Today most LCD panels are 8 bpc (bit per channel) and only the most expensive are 10 bpc. That means that each pure tone (R, G, B) in each pixel can be modulated through 256 steps (8 bit per channel) so roughly 2^24 tones are possible (16.7 M). The best panels support 2^30 tones (1073.7 M). VGA and DVI interfaces only provide 8bpc input: RGB-24. To excite a 10 bpc panel a DisplayPort interface or an HDMI 1.3 with DeepColour: RGB-30 enabled is needed. Video sources for these panels may be PCs with high end video cards that support true 30 bit colour or high end Blu-ray players with DeepColour (and a Blu-ray title encoded in 30 bit colour of course!). As we are capable of distinguishing over 16 M shades you can think that 8 bpc could be barely enough but here comes the tricky part…who told you that the 16 M shades of an 8bpc panel are 'precisely these' 16 M different shades that your eye can see? They are not for most cheap panels. Even when a colorimeter may tell us that the panel is producing 16M different shades, our eyes have a colour transfer function that must be matched by the source of light that pretends to have 16 M recognizable shades. If not, many of the shades produced by the source will probably lie in 'indistinguishable places' of our 'colour transfer function' rendering effectively much less than 16 M distinguishable shades. Professional monitors and very high end TVs have 'corrected gamma' output. This means that they do not attempt to produce a linear amount of polarization to each channel (R, G and B) in the range 0-255. Instead of that they have pre-computed tables with the right amount of R, G and B that our eyes can see all through the visible gamma. They use internal processing to 'bias' a standard RGB-24 signal to render it into 'the gamma of shades your eyes can see' with a preference of some shades and some disrespect for others so they can render RGB-24 input into truly 16M distinguishable shades. To achieve this goal these devices store the mapping functions (gamma correction functions) in a LUT (Lookup Table) that is a double entry table that produces the right voltage to excite the LCD for each RGB input. That voltage may have finer steps in some parts of colour space where human eyes have more 'colour density'. For this reason an 8 bit DAC won't be enough. More often an 8 bit signal is fed through the LUT to a 10 or 12 bit DAC. You see, more than 8 bits of internal calculus space are needed to handle 8 bit input so many 8 bpc displays are said to be 8 bpc panels with 10 bit LUT or 12 bit LUT. Today the best displays are 10 bpc panels with 14 bit LUT or 16 bit LUT. It is also worth mentioning a technique called FRC (Frame Rate Control). Some manufacturers claim they can show over 16.7 M colours while they use only 8 true bit panels .They double the frame rate and excite the same pixel with two alternating different voltages. So they fake a colour by mixing two colours through 'modulation in

time'. This technique seems to work perceptually but it is always good to know if your panel is true 10 bit or instead 8 bit + FRC. Once available this technique it has been used to make regular monitors cheaper by going all the way down to 6 bit + FRC.

We can conclude that today's high end monitors that properly implement gamma correction are proficient to show us the maximum range of colours we are capable of seeing (normal people see something above 16 M shades of colour, maybe even 50 M) when using LCD panels that are true 8 bpc or better (true 10 bpc). Unfortunately mainstream computer monitors and TVs do NOT have proper gamma correction and those that have the feature rarely are correctly calibrated (especially TVs), so digital colour is not yet where it should be in our lives. Most cheap monitors take 8 bit per channel colour and feed it right to an 8 bit DAC to produce whatever ranges of colour it ends up being, resulting in much less than 16 M viewable shades of colour. Many cheap computer screens and TVs are even 6 bit + FRC. By applying a colorimeter you can discover the gamma that your device produces and match it to some 'locus' in a standard 'colour space'. A standard colour space is a bi-dimensional representation of all colour shades the eye can see. This representation can be built only for some fixed intensity level. This means that with more or less intensity the corresponding bi-dimensional representation will be different. You can imagine a 'cone' with vertex in 'intensity 0 plane'. Slices of this cone (one for each fixed intensity value) lay in parallel planes occupying each one a 'locus' (a connected bi-dimensional plot) that gets bigger and richer as intensity increases until we get to optimal intensity and then starts to get washed out when intensity is above optimal. The locus of viewable shades takes in most representations the shape of a deformed triangle with pure Red, Green and Blue in the three vertices. Different colour spaces differ in shape but more or less all of them look like a triangle deformed into a 'horseshoe'. You may know Adobe RGB and sRGB. All these representations are subsets of the viewable locus and tried to standardize respectively what a printing device and a monitor should do (when they were created). Today's professional monitors can match 99.X% Adobe RGB which is more ample than sRGB. Most TV sets and monitors can only produce a locus much smaller than sRGB.

Refresh rates, interlacing and time response

How do we see moving objects? Does our brain create a sequence of frames? Does our brain even have the notion of a still frame? How do technology-produced signals compare to natural world in front of our eyes? Are we close to fooling ourselves with fake windows replacing the real world?

It turns out that we can only see moving objects. Yes that is true. You may be disturbed by this statement but no matter how solidly static an object is and how static you think you are… when you stare at it you are moving constantly your eyes. If your eyes do not move and the world does not move your brain just can see nothing. The image processor in our brain likes movement and searches for it continually. If there is no apparent movement in the world our eyes need to move so the flow of information can continue.

We just do not know if there is something like a frame memory in our brain, but it seems that our eyes continually scan space stopping by the places that show more movement. To understand a 'still frame' (reasonably still…, let's say that for humans something still is something that does not change over 10 to 20 ms) our eyes need to scan it many times looking for movement/features. If there is no movement eyes will focus on edges and high contrast spots. This process will give our brain a collection of 'high quality patches' where the high resolution channel that is the fovea has been aimed to , selected by its image characteristics (movement, edges, contrast) that may not cover completely our field of vision, so effectively we may not see things that our brain reputes as 'unimportant'. Our pretended 'frame' will look like a big black poster with some high resolution photos stuck all over following strange patterns (edges for instance), surrounded by many low resolution photos all around the HQ ones, and a lot of empty space (black you may imagine).

It seems we do not scan whole frames. It seems we can live well with partial frames and still understand the world. This cognitive process is helped by involuntary eye movement and by voluntary gaze aiming. Our brain has 'image persistence'. We rely on partial frame persistence to cover a greater amount of field view by making old samples of the world last in our perceptual system being added to fresh samples in a kind of 'time-based collage'. Cinema and TV benefit from image persistence by encoding the moving world as a series of frames that are rapidly presented to the eye. As our brain scans the world from time to time it does not seem very unnatural to look at a 'fake world' that is not just there all the time but only some part of the time. Of course the trick to fool our brain is: don´t be slower than the brain.

Cinema uses 24 frames per second (24 fps) and this is clearly slower than our scanning system so we need an additional trick to fake motion: we achieve it through over-exposure of film frame. To capture a movement through let's say one second spanning 24 film frames, we allow film to get blurred with the fast movement by overexposing each frame so the image of the moving object impresses a trail on film instead of a static image. If cinema was shown to us without this overexposure we will perceive a jerky movement as 24 fps is not up to our brain's capabilities to scan for movement. Most people will be much more comfortable with properly exposed shots played at 100 fps. The use of overexposure defines 'cinema style' movement as opposed to 'video style' movement. People get used to cinema and when a movie is shot not on film but on video with proper exposure the 'motion feeling' is different, they say 'too lifelike, not cinema'.

Today we see a mixture of techniques to represent motion. In TV we have been using PAL and NTSC standards that were capturing 576 and 480 horizontal lines respectively to form frames in a tricky way. They would capture half a frame (what they call a 'field') by sampling just the even lines, then just the odd lines, taking a field every 1/50 s in PAL and every 1/60 in NTSC. This schema produces 25 fps or 30 fps in average, but see that in fact they produce 50 fields per second or 60 fields per second. Due to the abovementioned image persistence two fields seem to combine in a single image, but notice that two consecutive fields were never sampled at the same time, but shifted 1/50 s or 1/60 s so if displayed simultaneously they won't match. Edges will show 'combing' (you will see dents like in a comb). This is precisely what happens when an interlaced TV signal arrives at a progressive TV set and you must turn fields into frames. Of course there are de-

comb filters built in modern TVs. I want just to point out that with 95% of TV sets out there being progressive monitors capable of showing 50/60 fps progressive, interlaced TV signals just do not make sense anymore…, but we still 'enjoy' interlaced TV in most places of the world. Of course de-interlacing TV for progressive TV sets comes at a cost: image filters result in a poor image quality. De-comb filters produce un-sharpening. The perceptual result is a loss of resolution. Maybe you do not realise how deep interlaced-imagery is in our lives: most DVD titles have been captured and stored as interlaced. This makes even less sense than broadcasting interlaced signals. DVD is a Digital format. You are very likely to play it on a Digital TV set (a progressive LCD panel) so why bother interlacing the signal so your DVD or your TV or both will need to de-interlace it?. Plainly it does not make sense, and by the way it worsens image quality. Even some Blu-ray titles have been made from interlaced masters. Here nonsense gets extreme but it happens anyway. We may forgive DVD producers as when DVD standard came up most TV sets were still interlaced (cathode ray tubes), but having 'modern' content shot in interlaced format today is plain heresy.

## 4 PERCEIVING SOUND

The human hearing system is made of two channels, each one acquiring information independently, both mapping information to a brain area. In the same way two eyes combine information for stereoscopic vision.

The hearing sensory equipment is complex. It is made of external devices designed for directional wave capture (ear, inner duct and tympanic membrane). We cannot aim ears (do not have the muscles some animals have), just turn our head which is a 'massive movement' subject to great inertia and thus slow, so our hearing attention must work all the time for surrounding sounds. There are internal mechanisms (chain of tiny bones: ossicles) and pressure transmission that are intended to transform the spectral response of the human hearing amplifying some frequencies more than others. At the very internal part of the human hearing system there is the cochlea, a tubular, spiralling duct that is covered with sensitive 'hairs'. Is in this last stage that individual frequencies are identified. All the rest of the equipment: internal and middle part and the ear is just a sophisticated amplifier. As in the visual system there are limitations that come from the sensory equipment, not the brain.

### Sound: differential pressure inside a fluid produced as vibration.

What we call sound is a vibration of a fluid. As in any fluid (gas or liquid), there is an average pressure at every point in space. Our hearing system is able to detect pressure variations deviating from the average. We do not detect any small variation (but almost!), and we will not detect a really strong one (at least we will not detect a sound, just pain and possibly damage of the hearing system). So there is a range of intensities in pressure: the human hearing system has an amazing

range of 13 decades in pressure (sound intensity as the energy produced per unit surface by the vibration is proportional to differential pressure). The smallest perceivable pressure difference is: 2x10^-5 Newton/m^2, the highest is 60 Newton/m^2. The intensity range is 10^-12Watt/m^2 to 10Watt/m^2. We do not detect isolated pulses of differential pressure. We need sustained vibration to excite our hearing system (there is a minimum excitation time of about 200 ms to 300 ms). And thus there is a minimum frequency and also a maximum. Roughly we can hear from 20 Hz to 20 kHz. Aging reduces severely the upper limit. We are not equally sensible to pressure (intensity) through the range. As with our eyes and colour perception, there is a transfer function, a spectral response in frequency space that is not flat.

Frequency resolution: pitch discrimination, Intensity resolution: loudness

Our hearing system retrieves the following information from sound: frequency (pitch), intensity (loudness), position (using two ears). We are able to detect from 20 Hz to 20 KHz and we can tell 1500 different pitches in the range. Separation of individually recognizable pitches is not the same across the range. There is a fairly variant transfer function. It is assumed that frequency resolution is 3.6 Hz in the octave going from 1 KHz to 2 KHz. This relates to perception of changes in pitch of a pure tone. As with colour, sound can be a perception of combined pure tones. It happens that when there is more than one pure tone, interference between tones can be perceived as a modulation of intensity (loudness) that is called 'beating' and the human ear is then more sensible to frequency. For instance two pure tones of 220 Hz and 222 Hz when heard simultaneous interfere producing a beating of 2 Hz. The human ear can perceive that effect, but if we increase let's say our pure tone from 200 Hz to 202 Hz the human ear will not perceive the change.

We perceive sound intensity (loudness) differently across the frequency range. Several different pressures can be perceived equal if they are vibrating at different frequencies. For this reason the human ear is characterized by drawing 'loudness curves'. There is a line (a contour line as level curves in geographical maps) per perceived loudness value, these lines cover the whole range of audible frequencies and they (obviously) do not cross. It is noticeable that the low threshold of loudness perception has a valley in the range 2 khz to 4 khz. There is where we can hear the less intense sounds. In that frequency range lays most of the energy of the human voice spectrum.

Sound Hardware: is it up to the task?

It seems that audio-only content is not very fashionable in these days. It does not attract the attention of the masses as it did in the past. Anyway let's take a look at what we have.

Sound storage and playback is mostly digital in our days. Since the inception of the CD a whole culture of sound devices employs the same audio capacities. CD spec: two channels, each sampled at 44.1 KHz, 16 bit samples using PCM. (Sound input is filtered by a low-pass with cut frequency at

20 KHz, and then sampled. As you may know to preserve a tone of 20 kHz you must sample at least at 40 kHz: Nyquist theorem). Is CD spec on par with human perception? Let's see.

Audio sampling takes place in time domain (light sampling happens in frequency). Each sample takes an analog value for pressure (provided converting pressure to voltage or current intensity in a microphone during sampling time), then this value is quantized in N steps (16 bits provide 2^16 =65536 steps). The resulting bit stream can be further compressed after for storage and/or transmission efficiency. Some techniques can be applied before quantization to reduce amount of data (amplitude compression), but usually data reduction is applied while quantizing leading to Differential PCM and Adaptive PCM which deviate from Linear PCM.

If we look at frequency resolution, LPCM with cut-off frequency at 20 kHz is ok. Two separate tones of less than 20 KHz will be properly sampled and can be distinguished perfectly no matter them are separated by 2-4 Hz.

If we look at pressure resolution it has nothing to do with CD spec. A loudness curve for each pressure value can be properly encoded using the CD spec. What plays here is the Microphone technology (sensitivity) and all the chain of manipulations (AD conversion, storing, transmitting, DA conversion, amplifying) analog and digital that intervene to display sound in front of your ears.

It is easier to look at dynamic range to compare fidelity of sound handling. As we said the human ear is capable of 13 decades (130 dB), but as it happens with the human eye not in the same scene, not in the same sound time segment. For human hearing the name of this reduced range effect is called 'masking'. Loud noises make our hearing system to adapt by reducing the range so we cannot hear faint sounds when an intense signal is playing. Some experiments (http://en.wikipedia.org/wiki/Dynamic_range) demonstrate that the CD spec (16 bit samples) can render a range of 98 dB for sine shaped signals, a range of 120 dB using special dithering (not LPCM), 20 bit LPCM can render 120 dB, 24 bit LPCM can render 144 dB. But at the same time other elements in the chain: AD/DA steps, amplifiers, transmission are very likely to reduce the range below 90 dB

So theoretically there are high end sound devices with high dynamic range, that could be paired among them so the resulting end to end system gets close to 125 dB, but that may take a fair amount of money. To technically achieve the maximum possible fidelity one way (recording) and the other way around (playing) you must ensure that all your equipment fits together not breaking the 125 dB dynamic range. For the recording segment it is no problem. Studios have the money to afford that and more. For the playing segment you will find trouble in DAC, amplifiers and loudspeakers. Cheap HW does not have the highest dynamic range. The symptom of not being up to the task is the amount of distortion that appears in the range, usually measured as maximum % of distortion in the range.
http://books.google.es/books?id=00m1SlorUcIC&pg=PA75&redir_esc=y). But anyway ¿do we need the maximum dynamic range all over the chain? Is audio quality available? Is it expensive?

Reading this (http://www.aes.org/e-lib/browse.cfm?elib=14195) and this (http://www.tomshardware.com/reviews/high-end-pc-audio,3733.html) may help us derive the conclusion that YES, we have today the quality needed to experience the best possible sound that our perceptual system can detect, and NO it is not expensive. Using computer parts and peripherals you can build a cheap and very perceptually decent sound system. Of course pressure wave propagation is a tricky science and to adapt to any possible room you may need to invest in more powerful much more expensive equipment. But for 'direct ray' sound we are fortunate, virtually anyone can afford perceptually correct equipment today.

## 5 DIGITAL ENCODING, FORMATS AND PROCESSING

### Video

With the advent of Digital technologies we have inherited a world in which video is digital. This means of course that the video signal is 'encoded' in a digital format. Today EBU, DVB and other organisations like DVD Forum and Blu-ray Disc Association have closed the variety of encoding options to a few standards: MPEG (1, 2, &4) and VC1, there are also other famous video coders: WebM, On2 VP6, VP8 and VP9. By far the more successful standard is MPEG (Motion Pictures Expert Group) which is today well consolidated after more than 25 years of existence. Most of the TV channels in the world are today delivered as a MPEG2 video over M2TS (MPEG 2 transport stream) and more recently HD TV is delivered as MPEG4-part 10 video over M2TS. Also called ISO H264 or AVC. And the latest addition from MPEG is HEVC (H265)

We are seeing a digital world that moves much faster than EBU/DVB/DVD FORUM/BLURAY DISC ASSOCIATION… these big entities may take as much as 5 years to standardize a new format. Then manufacturers NEED to change production lines and keep them for some years producing devices that adhere to the new standard (so they can derive profit from investment). So you cannot expect a breakthrough in commodity electronics for video in less than 5-7 years and that is accelerating as in the past (1960-2000), TV sets have been built essentially equal in viewing specs for more than 20 years in a row.

But as I've said today we can expect to see much more dynamism in video sources and video displays. Thanks to Internet video distribution people can encode and decode video in a wealth of video formats that may fit best their needs than regular broadcast TV. Of course the hardest limitation is the availability of displays. It doesn't make sense to encode 30 bit colour video if you do not have a capable display… but assuming we have a capable display, today high end PC video boards can be used to feed HDMI 1.4 or DisplayPort signals to these panels overcoming the limitations of broadcasting standards. For this single reason 4K TVs are being sold today. Only PCs can feed 4K content to current 4K TVs, and most of the times this content must be 'downloaded' from Internet. Today streaming 4K content would take above 25 Mbps encoded in H265 HEVC.

We have seen that state of the art displays have just met perceptual minimum resolution (250 ppi at 18") and are getting better every day. We are seeing the introduction of decent colour handling with 10 bpc LCDs and 30 bit RGB colour. We are seeing the introduction of large format high resolution displays: 4K displays starting at 24" for computer monitors (200 ppi) and at 56" for TV sets (80 ppi). BDA has recently announced that the 4K extension to BluRay spec will be available before end 2014. In the meantime they need to choose the codec (H265 and VP9 are contenders) and cut some corners of the spec. The available displays have a proper dynamic range usually better that 1000:1 and getting close to 10000:1. At least if we take contrast ratio as if it were a 'real' intensity range, which is not. Of course our monitors cannot light up with the intensity of sunlight and at the same time or even in a different scene show a star field with distinguishable faint stars. No. Dynamic range of displays will not get there soon, but HDR (High Dynamic Range) techniques are starting to appear and they can compress the dynamic range of real world input much better than current technology, which by the way does not compress it just clips. Current cameras can take high light clipping low light or on the contrary low light clipping high light, and you are fortunate to be left to select which part of the range you want. Near future HDR cameras will capture the full range. As displays will not be on par with the full range, some image processing that is already available must be done to compress the range and adapt it to the display. *(PS: today you can process RAW files to produce your own HDR images, or even take multi-exposure shots to produce HDR files. The problem is that to see the results you must compress the range to standard RGB or otherwise you must select an 'exposure' value to view the HDR file.)* We can expect to see incredible high definition high dynamic range content in full glory using 4K 30 bit per colour displays.

Audio

Digital audio is not living up to digital video expectations. In the past decade a few high definition audio formats appeared: SACD (Super Audio CD), DVD-A (DVD Audio) and multichannel uncompressed LPCM in BD. From these formats SACD and DVDA have proved real failures. It has been demonstrated (http://www.aes.org/e-lib/browse.cfm?elib=14195) that increasing the bit count per sample above 16 bits: 20, 24, 32, 48…. and increasing sampling rate above 44.1 khz : 48, 96, 1 Mhz (DDS 1 bit) do not produce perceptually distinguishable results… so the answer is clear: we got there many years ago. We have achieved the maximum 'reasonable' fidelity with the CD spec. (OK, OK the noise floor could be improved going to sound dithering or moving from 16 to 20 bit, but anyway the perceptual effect is ridiculous and the change is not worth the investment at all.) The only breakthrough in digital audio comes from the fact that now we have more space available in content discs so we can go back to uncompressed formats and enjoy again LPCM after years of MP3 compression or other sorts of compression : DTS, Dolby, MPEG. Also state of the art audio is multichannel. So the reference audio today is uncompressed LPCM 16 bit/sample, 44.1 or 48 khz in 5.1 or 7.1 multichannel format stored on Blu-ray disc.

## 6 CONCLUSIONS

I started this article posing fairly open questions about the availability of perceptually correct technology to display image and sound in front of our eyes and ears. After careful examination of our viewing and hearing sensory equipment, and after examination of the recent achievements of the CE industry providing displays and audio equipment, and the prices of these devices and after examining the market acceptance for content and the ways to distribute content…. we can conclude that we are living in an extremely interesting time for content. We've got 'there' and virtually no one noticed. We have the technology to provide perceptually perfect content and we have the distribution paths and we (almost) have the market.

In the way to this discovery we have found that today only a very small amount of devices and content encodings have put all the pieces together, but that is changing. We will be no more delayed by broadcast standards; we will no more be fooled by empty promises in audio specs. The right technology is just at hand and the rate of price decline is accelerating. Full HD adoption took more than 15 years but maybe 4K adoption will take less than 5 years, and maybe most content will not get to us via broadcast anymore…

## 7 REFERENCES


http://www.clarkvision.com/articles/human-eye/index.html

http://www.normankoren.com/Tutorials/MTF.html

http://gene.bio.jhu.edu/resolution/resolution.html

http://micro.magnet.fsu.edu/primer/lightandcolor/humanvisionintro.html

http://www.eizo.com/global/library/basics/maximum_display_colors/

http://www.liftgammagain.com/forum/index.php?threads/monitor-eizo-cg246.2439/page-3

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

http://www.liftgammagain.com/forum/index.php?threads/monitor-eizo-cg246.2439/page-3

http://hyperphysics.phy-astr.gsu.edu/hbase/sound/earsens.html#c3

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