♦ Supporting Next-Generation Multimedia Services Over a Broadband Access Network

Danny De Vleeschauwer

Broadband access networks, traditionally deployed for data services, increasingly need to support streaming services as well. Some papers have predicted the bit rate an access link should be able to carry in order to support a future mix of all services. These predictions typically take two trends into account: the fact that the information is offered in ever more detail (e.g., the use of high definition instead of standard definition video) drives the bit rate up, while improvements in codec technology have a tendency to decrease the bit rate. In this paper, we estimate the bit rate a user is likely to consume based on the limits of human perception. From this estimation, we infer what an access link in principle should be able to carry and how far the current wire-bound network technologies are from this ideal situation.

Introduction

While broadband access networks were originally designed to support web browsing and email, today a richer set of services, e.g., instant messaging, voice calls, exchange of large files, and streaming multimedia services (in particular, television services), need to be supported. In order to offer the users of these services a good quality of experience, all stages of the network should be well dimensioned. In the aggregation, metro and core network statistical multiplexing can be exploited, relying on the fact that not all users are active at the same time. In this paper, however, we concentrate on the last mile link. The last mile link provides a user (or a household of users) access to these services and should be dimensioned such that it can accommodate the peak bit rate.

In anticipation of the richer set of services a user is likely to consume, the access bit rates offered by access providers to residential users are increasing. Ovum details the highest access bit rates currently offered by various access providers in the top ten most competitive access network markets in [26]. Where fiber to the premises (FTTP) is widely deployed, 100 Mbps is a common downstream access bit rate, while where digital subscriber line (DSL) or cable is still predominant, the downstream bit rate offered is in the range of 20 Mbps to 50 Mbps.

Some studies [6, 10, 24] have investigated how the bit rate of the last mile link should evolve in order to support a future mix of broadband services. These predictions typically make a distinction between the portion needed for streaming services and the portion needed for data services. For the former service, often two competing trends are taken into account. On the one hand, a user consumes the streaming services at higher resolution requiring a larger bit rate, while on the other hand, multimedia compression

Bell Labs Technical Journal 14(1), 73–88 (2009) © 2009 Alcatel-Lucent. Published by Wiley Periodicals, Inc. Published online in Wiley InterScience (www.interscience.wiley.com) • DOI: 10.1002/bltj.20355

Panel 1. Abbreviations, Acronyms, and Terms

3D—Three dimensional
ADSL—Asymmetric DSL
AVC— Advanced video coding
CIE—Commission International de l'Eclairage
CM—Cable modem
CMTS—CM termination system
DSL—Digital subscriber line
DSLAM—DSL access multiplexer
DOCSIS—Data Over Cable Service Interface Specification
FEC—Forward error correction
FTTP—Fiber to the premises
EPON—Ethernet PON
GPON—Gigabit PON
HD—High definition
HG—Home gateway
HRTF—Head-related transfer function
IEEE—Institute of Electrical and Electronics Engineers
ITU—International Telecommunication Union
ITU-T—ITU Telecommunication Standardization Sector
IP—Internet Protocol
IPv4—IP version 4
IPv6—IP version 6
MAC—Medium access control
OLT—Optical line termination
ONU—Optical network unit
PON—Passive optical network
SD—Standard definition
SHV—Super hi-vision
TV—Television
VDSL—Very high speed DSL
WER—Word error rate

techniques continuously improve, thus driving the bit rate requirement down. For the latter service it is taken into consideration that typical data file sizes keep increasing, while the user's patience for the file to download is wearing thinner as he or she becomes used to high-speed Internet access.

The argument presented in this paper differs from these traditional predictions in three important aspects. First, we estimate the bit rate required per user while the traditional studies assess the bit rate per household. To tie our result in with the traditional studies, the required bit rate per user obtained in this paper needs to be multiplied by the number of (simultaneously active) users in the household. Second, rather than observing the current tendencies, we assess if the two competing trends, i.e., the increase in resolution of the multimedia objects and the increase in compression gain, will find some balance. In order to identify this balance, we start from the limits of human perception. Third, we make no distinction between streaming and data services. The traditional studies assume that for data services a data file needs to be completely downloaded before the user can access (e.g., manipulate, view, or listen to) it. We argue in this paper that if the information in the data file is organized properly in what we will refer to as a progressive file format, the requirements for supporting data services become very similar to the ones for streaming services. That is, sustaining an adequate bit rate is sufficient as long as the first relevant, intelligible part of the information stored in the data file is presented to the user fast enough. For that reason, we will consider delay requirements as well, on top of bit rate requirements.

The per-user bit rate requirements we identify are lower bounds to support high quality services consumed by humans. Although currently most streaming services are designed with these bounds in mind, data services typically are not. In this paper, we argue that as the size of the data objects increases, the boundary between both types of services blurs, and hence, that it is beneficial to design data services with these bounds in mind too. Note too that similarly to the traditional studies, we only concentrate on services to be consumed by humans and as such we do not take traffic associated with machine-to-machine communication (e.g., sensor networks) into account in this paper.

Limits of Human Perception

Networked applications to be consumed by humans all currently rely on the exchange of audiovisual information. Although there have been some experiments with haptic interfaces, the sense of touch is not yet widely used in networked applications. The other traditional human senses, i.e., the senses of taste and smell, are even further from being supported in networked applications.

Note that in this paper we take the broadest definition of audiovisual information: all pieces of information that need to be viewed and/or listened to are considered to be audiovisual information. As such, text, pictures, audio clips, and video footage are regarded as audiovisual information. Typically, we concentrate on large pieces of information that are likely to put a high demand on the network. Traditionally a distinction is made between streaming services, where the information is offered to the user under the form of a stream, and data services, where the user interacts with a file containing the information. Data exchanged between computers, not destined for human consumption, is not considered in this paper.

Human perception is not perfect, as witnessed most prominently by optical illusions. In the next paragraphs we describe the limits of human perception relevant in the context of networked applications.

Vision

First, we consider the limits of the human visual system. **Figure 1** shows a sketch of how the human eye processes visual information. Visual information consists of an evolving light intensity pattern that depends on two place coordinates and one time coordinate. It is customary to express a location on the retina by specifying the horizontal and vertical angle the line from the optic center to that particular location on the retina makes with the line between the optic center and the fovea (i.e., the most sensitive part of the retina). As such locations on the surface of the retina, i.e., the place coordinates, are expressed in degrees. The dashed lines in Figure 1 illustrate that there is a horizontal visual angle of about 15 degrees between the fovea and the center of the blind spot where the optic nerve leaves the retina and which is insensitive to light. In the nasal and temporal direction, the retina is sensitive to light up to 90 degrees

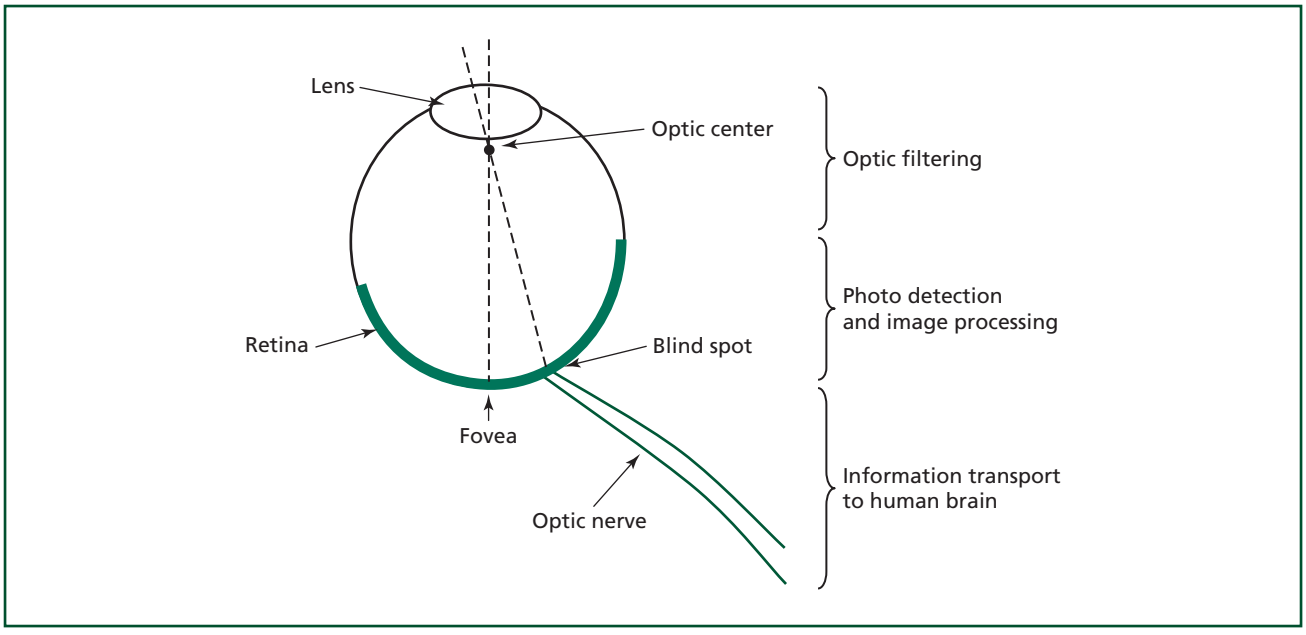


Figure 1.
Symbolic representation of the working of the human eye.

and 70 degrees, respectively. The visual angle of one eye is about 160 degrees horizontally and 170 degrees vertically, while binocular vision has a horizontal and vertical visual angle of 120 degrees and 135 degrees, respectively [33].

A light pattern entering the eye is projected on the retina. Since the opening of the eye is not infinitesimally small and the lens is not perfect, this is equivalent to a linear filtering. The retina consists of three layers of cells [21]. A first layer is made up of two types of photoreceptors, i.e., rod and cone cells. The rods are mainly responsible for vision under low light conditions (i.e., night vision), while the three types of cones enable color vision under normal lighting conditions. Since the density of the cones is highest in the fovea and falls off rapidly away from the fovea (i.e., outside the disk with diameter 10 degrees there are practically no cones), cones are mainly responsible for central vision, while since the density of the rods is highest in the periphery, rods are mainly responsible for peripheral vision. The second layer of cells in the retina is made up of bipolar, horizontal, and amacrine cells, while the third layer consists of ganglion cells. These two layers sequentially process the information picked up by the photoreceptors. The output of the ganglion cells is communicated over the optic nerve to the brain.

Although the retina around the fovea has a resolving power of about 0.4 arcmin (with 1 arcmin being 1/60 of a degree), the optical system of the eye has some limits too. A person with "20/20 vision" is just able to discern a detail of 1 arcmin in his or her central vision [18]. This corresponds to the ability to separately distinguish the (five) segments of an "E"-shaped letter of 9 millimeter (mm) height at 20 feet. Outside this small angle of central vision, the acuity is a lot smaller. However, using saccadic motion (i.e., small, rapid, unconscious eye movements) the eye scans the environment and thus builds an internal picture of the outside world, such that this high resolution is required in a larger visual angle.

As the retina cannot follow rapidly changing light patterns, it acts as a temporal filter. Typically the response of the retina starts to decrease at 30 Hz and any frequencies beyond 50 Hz cannot be discerned [18].

The Commission International de l'Eclairage (CIE) has defined a "standard colorimetric observer" [11]. This standard states that in order to completely determine a color, three values have to be specified. Various three-dimensional color spaces were defined consisting of the CIE-XYZ, CIE-LUV, and CIE-Lab color space. Moreover, the Weber-Frechner law states that the ratio of a just noticeable increment ΔI of a stimulus and the intensity I of that stimulus is constant over a whole range of intensities: $\frac{\Delta I}{I} = c$. It follows that human perception is logarithmic. Taking into account the dynamic range of the human eye, quantizing each of the three color components with a logarithmic quantizer with 1,024 levels (requiring 10 bits) is sufficient to keep the quantization error below the visible threshold, although 256 levels (requiring 8 bits) are often used too.

Taking all these considerations into account, a conservative estimate for the bit rate associated with the evolving light pattern falling on the retina can be made: for example, assuming a visual acuity of 1 arcmin; a horizontal and vertical viewing angle of 120 and 90 degrees, respectively; a refresh rate of 50 Hz ; and 8 bits for each of the three color components yields an information flow of about 47 Gbps entering each eye.

The three layers of the retina process this information flow and transport a reduced information flow over the optic nerve to the brain. The bit rate associated with the output of the ganglion cells, i.e., the bit rate traveling over the optic nerve, has been estimated in [20]. By measuring the bit rate produced by various types of ganglion cells and counting the number of ganglion cells of each type in the human retina (which contains about 1,000,000 ganglion cells in total), the authors claim that natural images yield an information flow of about 10 Mbps over each optic nerve. This means that the retina compresses the visual information it receives by a factor more than 1,000.

The two information flows traveling over the optic nerve, one associated with each eye, cross at the chiasm, where some nerve fibers cross over and others do not, such that the left part of the observer's visual field is sent toward the right part of the brain, and vice

versa. There is no visual processing, neither in the nerves nor in the chiasm. Besides the preprocessing in the retina, most visual processing is performed in the visual cortex at the back of the head.

Under normal circumstances, there is considerable correlation in the signals of both eyes: the images contain the same objects, but these objects are slightly shifted with respect to each other in the left and right eye. This correlation is used in the visual processing in the visual cortex to extract depth information [2]. How the information flows through various parts of the visual cortex and the bit rate associated with these flows is currently unknown.

Audition

Second, we consider the limits of the human aural system. **Figure 2** shows a sketch of how the human ear processes aural information. Aural information consists of pressure vibrations in the air: an aural signal presented to one ear depends only on the time coordinate. The aural signal is picked up by the pinna (i.e., the auricle) and travels in the aural canal to hit the tympanic membrane (i.e., the eardrum). Through a series of small bones, this signal is transferred to the cochlea. This process is equivalent to a low-pass filtering.

The human ear is sensitive in the frequency range from 20 Hz to 20 kHz, such that a sampling rate of at least 40 kHz is required. The Weber-Frechner law applies to audition as well and the dynamic range of the human ear spans 140 dB. At least 256, and preferably 1,024, logarithmically spaced levels (i.e., 8 to 10 bits per sample, respectively) are needed for the quantization error to be indiscernible to a human listener. Taking these considerations into account, a rough estimate of the information flow presented to one of the human ears contains about 0.5 Mbps (with sampling rate of 50 kHz and 10 bits per sample).

The cochlea acts as a spectral decomposition filter [3]: each part of the cochlea is tuned to a specific frequency. Ganglion cells inside the cochlea pick up the band-pass signals and feed them into the aural nerve. To our knowledge, there has been no detailed study estimating the bit rate associated with the signal traveling over the aural nerve. However, if these cochlear ganglion cells produce about the same bit rate as the retinal ganglion cells—and taking into account that there are about 30,000 ganglion cells in

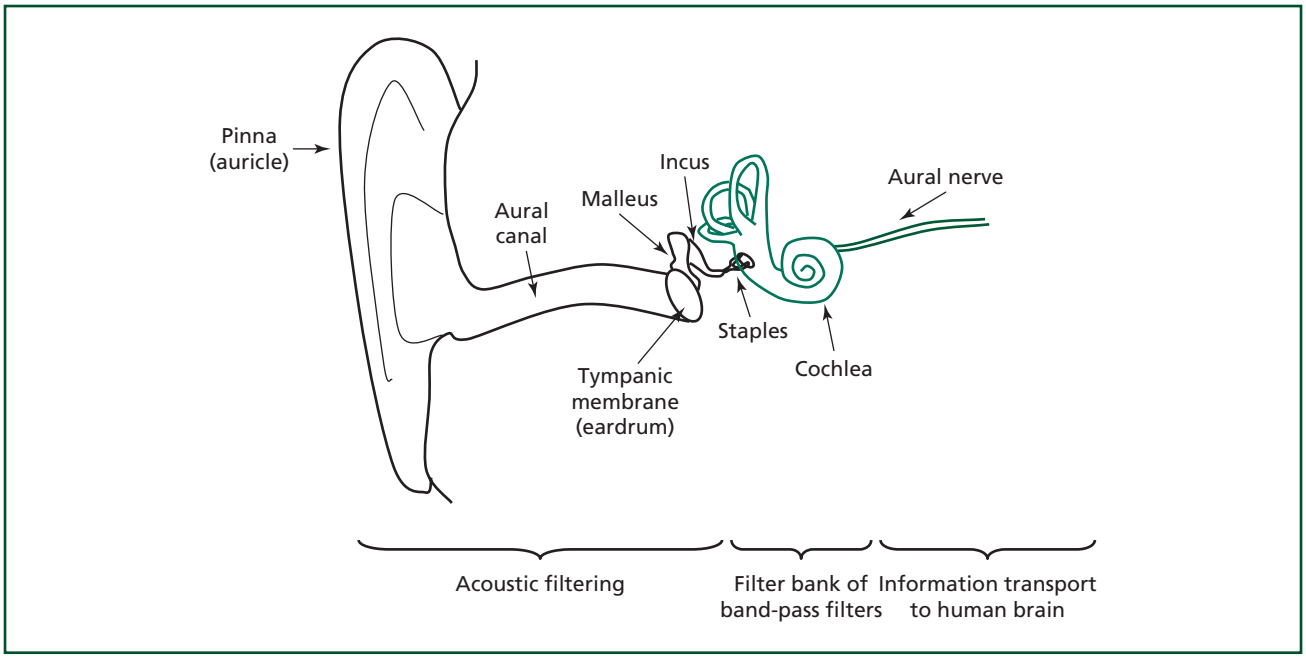


Figure 2.
Symbolic representation of the working of the human ear.

the cochlea—the cochlea does not compress the aural information very much. Note that the cochlea has a second function, i.e., providing a sense of balance, which we do not discuss in this paper.

Further processing of the aural signals of both ears occurs in various parts of the brain, e.g., in the primary aural cortex at the side of the head. There is currently no body of knowledge on the bit rates involved in the internal transport of this processed information.

Reaction

Finally, we discuss the limits of the human reaction. The proverbial "blink of an eye" has been measured to last about 300 milliseconds (ms) [34]. Although reflex reactions take less time, i.e., tens of milliseconds, reactions where the human brain needs to process some information before taking action typically take a few hundred milliseconds as illustrated by the following examples:

- Trained sprint athletes are assumed to be cheating if they exert pressure on their starting blocks any time before 100 ms after the starting shot sounds.
- According to [14], most applications will experience "transparent interactivity" if the one-way delay is smaller than 150 ms. This means that if a user takes an action (e.g., asks a question) he or she does not expect a reaction (e.g., a reply) before 300 ms after he or she took the action.
- In [22], some experiments are reported related to the quality perceived by a user changing channels on a television system. In order to have a "good" quality of experience, the first image of the new channel needs to appear about 300 ms after the user took action to change channel.

In addition to this bound on reaction time, the human brain is unable to process stimuli in rapid succession. In particular, when a user is concentrating on a specific stimulus, he or she is capable of discerning other stimuli up to 180 ms after the first stimulus and from 450 ms onward of this first stimulus, but the human brain is unable to process stimuli in the time frame between 180 ms and 450 ms [27]. This phenomenon is referred to as the "attentional blink." A similar mechanism exists in audition, but the period of "deafness" starts immediately after the aural stimulus the user was paying attention to. This phenomenon is referred to as temporal masking.

Technological Aspects of End Devices

In this section we discuss how the limits of human perception can be exploited in transducers (e.g., displays and speakers) and compression technology.

Transducers

First we discuss audio transducers (e.g., speakers and headphones). Their ultimate aim is to present both ears of a listener with signals that are indistinguishable from the signals that the listener would hear when he or she was really present at the scene where the sound was produced. The considerations of the previous section provide guidelines on the required spectrum, the quantization, and the dynamic range of the signals.

The two signals presented to both ears (e.g., via earphones) are very similar. The small differences in these two sounds allow the human brain to determine the spatial location from where the sound is originating. Not only the fact that one signal is slightly delayed and attenuated with respect to the other, but also the fact that sounds are filtered by the shape of the human head and auricle provide cues with respect to the spatial location. The head-related transfer function (HRTF) [35] describes how a sound stemming from some location is altered (i.e., filtered) by the human head before the filtered version is presented to one of the ears. As the HRTF contains a filter for both ears and for each possible location, it is in fact a filter bank. A recording with two microphones spaced apart as far as the human ears are apart is not enough: in such a recording one signal is only slightly delayed and attenuated with respect to the other and this allows some spatial localization, but it does not allow a listener to discriminate between a sound being produced in front of or behind the head. To allow this distinction, the sound needs to be recorded with two microphones embedded on the place of the ears in an artificial human head thus mimicking the HRTF. Alternatively, the sound only needs to be recorded with one microphone if the HRTF is known. From this

Table I. Examples of visual acuity of a person with normal vision.

Visual acuity	Resolution		Unit
	0,4	1	Arcmin
Sheet of paper (of 11″ × 8.5″) viewed at arm's length (30cm)	728	291	dots/inch
SD screen viewed at 6 screen heights with 4:3 aspect ratio	1432	573	lines/height
HD screen viewed at 3 screen heights with 19:9 aspect ratio	2865	1146	lines/height

cm—Centimeter
HD—High definition
SD—Standard definition

single recording and the spatial location of the source, the two signals to be presented via earphones to the human ears can be calculated via a filtering operation with the HRTF.

Offering the two correct signals to both ears of the listener with speakers instead of earphones requires a careful setup of an array of speakers. The "5.1 surround sound" arrangement with a speaker left front, center front, right front, left back, right back, and an additional speaker for low frequencies is commonly used in movie theaters at the moment. In [25], a "22.2 surround sound" system was proposed to create an acoustic image with 22 speakers in a three-dimensional arrangement with two additional speakers for low frequencies.

Next, we discuss video transducers (e.g., displays, goggles). Their ultimate aim is to present both eyes of a viewer with images that are indistinguishable from the images that the viewer would see if he or she were present at the scene in person. The requirements are that the images are projected onto the retina with enough spatial and temporal resolution and that the color space is quantized adequately.

Current displays do not cover the complete viewing angle. Although there is no universally accepted definition of standard definition (SD) and high definition (HD), we define SD as a screen with a 4:3 aspect ratio that is to be viewed at six screen heights, while HD is a screen with a 16:9 aspect ratio that is to be viewed at three screen heights [13]. **Table I** provides some examples of how the visual acuity of a regular person determines the required resolutions for displays. This is corroborated by [8] and [32], which conclude that an HD screen needs just more than 1,000 lines per screen height. In [25], a super hi-vision (SHV) display was introduced with more than 4,000 lines per screen height to be viewed at a distance of 0.75 of the screen height. **Table II** shows the visual angle that these displays cover, under assumption of square pixels.

All these displays present the same image to the left and right eye, and as a result, the viewer has no intense depth perception. In order to convey a sense of depth, the images to be presented to the left and right eye need to be slightly different. More or less, the same objects should be depicted in the images, but they need to be projected on slightly different locations on the retina. This difference in location, referred to as parallax, allows the human visual system to infer depth [4]. Three-dimensional (3D) dis-

Table II. Visual angle of an SD, HD, and SHV display.

	Vertical angle (degree)	Horizontal angle (degree)
SD screen viewed at 6 screen height with aspect 4:3 ratio	10	13
HD screen viewed at 3 screen height with aspect 16:9 ratio	19	33
SHV screen viewed at 0.75 screen height with 16:9 aspect ratio	67	100

HD—High definition
SD—Standard definition
SHV—Super hi-vision

plays that present both images on the same screen and rely on glasses (to be worn by the viewer) to filter the correct image from the mix of two have been around for a long time. Recently, 3D displays based on a grating of lenticular lenses became available [19]. The latter displays typically provide more than two views, so that it is not as critical where the observer assumes a position in front of the screen.

Compression Technology

Up to this point, we have discussed how to present both ears and both eyes of a human observer with the adequate stimuli such that he or she does not observe unnatural artifacts such as frequency aliasing or quantization artifacts. If these artifacts are to be avoided, the transducers need to output the bit rates derived in the previous section. We also discussed that the human brain processes these stimuli, and that in processing these stimuli, the information content is reduced. Although the most striking example is the retina compressing the information contained in the light pattern projected on it by a factor of more than 1,000, it is very likely that deeper in the brain, similar processes compress (and separate) the video and audio information even further.

The aim of video and audio compression is to represent the signals destined for the human eyes or ears in signals of a lower bit rate, such that a decoder can reconstruct close copies of the original signals from these compressed signals. Since the compressed signals contain less information, the original signals cannot be decoded with infinite fidelity. In other words, compression inherently introduces distortion. The aim of compression is that the decoded signal is free from unnatural artifacts and contains the essential information that the producer of the original signal wanted to convey. Most state-of-the-art codecs transform the input signals into semantically more meaningful signals (mimicking what the human brain does) and remove the redundancy in those resulting signals without introducing unnatural artifacts. Remark that most state-of-the-art codecs use prediction to attain a high compression gain. This introduces some delay in the encoding and decoding process.

It is not always necessary to reproduce the original information with the highest fidelity in the sense that the reconstructed signal needs to be perceptually indistinguishable from the original signal. Often the reconstructed signal is perceptually very different from the original one, while the observer still experiences it as quite natural. In this section, we discuss the possible gradation in fidelity of audio and video information and the state of the art in compression technology.

For an audio signal, it is sometimes only necessary that the message it contains can be understood. In [29], it was shown that a typical English text contains about 1 bit per letter of information. At a speaking rate of about 4 to 5 words per second, and with an average of about 4.2 letters per word, this yields a bit rate of a mere 20 bps, so that conveying a verbal message requires a very limited number of bits. Speech-to-text algorithms can be viewed as encoders in this context. Although advancements are continually being made, for systems that are not trained a priori for a specific speaker, the word error rate (WER) is still 10 percent and higher [31]. If on top of comprehending the verbal message, the listener also needs to be able to identify the speaker, the pitch and intonation of the speech signal need to be conveyed in the compressed signal too, which is what a voice coder typically does. State-of-the-art voice coders can encode the human voice at a few tens of kbps [1], but are not suitable for encoding audio signals that are not produced by the human vocal tract (e.g., music). For such signals (another codec and) a larger bit rate is needed. In [30], it is demonstrated via subjective experiments that a state-of-the-art audio codec requires about 100 kbps for stereo music. As explained earlier, the two signals to be presented to both ears (or the multitude of signals to be presented to the speakers in a surround sound system) should not be encoded separately. In fact, only one signal and its spatial location together with the HRTF need to be known at the decoder, the latter of which does not need to be transmitted, but can be permanently stored in the decoder.

Next, we consider video signals. Here too it is often not necessary to present the signals to the human eyes with the highest fidelity. Monochromatic television, which was common up to the last part of the last

century, is an example of representing the signal with a minimal fidelity. Even in current color television systems, there is no full fidelity. First, state-of-the-art displays still do not cover the complete color space. Second, as shown in Table II, video signals commonly transported today (when viewed at the distance they were designed for) only cover a part of the visual field of a viewer. Finally, in most of the cases the depth information is not conveyed. In [9] it is shown via subjective experiments that an HD display (to be viewed at three screen heights, with a 16:9 aspect ratio and with more than 1,000 lines) requires about 8 Mbps with a state-of-the-art codec for the most difficult sequences. In order to convey depth information as well, at least two images (i.e., a left and right view) need to be transported. Here again, these two signals do not differ much, so it is more economical to send a central view, depth information (in the form of a parallax field), and information pertaining to the parts of the scene that are occluded in the central view, but are visible in the other view [4]. To transport depth information in this format typically requires an additional bit rate of 10 percent.

Requirements for Audiovisual Information

From the preceding discussion we can draw the following conclusions. Aural information in the highest fidelity currently needs about 100 kbps. This is less than what travels over the aural nerve. This is due to the fact that state-of-the-art codecs take into account processes (e.g., masking) that occur beyond the stage of the cochlea in the human brain. A state-of-the-art video codec can represent visual information in a visual angle (33 degrees horizontally and 19 degrees vertically) with about 8 Mbps. Such a visual angle covers most of the cones, but only a fraction of the rods. Since, as we discussed, all rods and cones together produce about 10 Mbps, there is probably some room for improvement in video codec technology, especially if aspects of the human visual system beyond the retina are taken into account. As specified earlier, conveying depth information (i.e., two views) typically requires an additional bit rate of 10 percent. Finally, if the requested information flow can be switched to the user in a time frame expiring 300 ms after he or she took the action to assess it, the user will experience transparent interactivity with the information.

In the remainder of this paper, we will refer to these bit rate values (around 100 kbps for aural information and about 10 Mbps for visual information, augmented with 10 percent for conveying depth information) as the "audiovisual bit rate bounds" and to a response time below 300 ms as "a blink of an eye." In the next paragraphs we assess for the two types of services, i.e., streaming and data services, whether or not they are designed with these bounds in mind and, if they are not, how they still could benefit from doing so. Since we only discuss the last mile link in this paper, we assume that the desired information (i.e., stream or file) is readily available in the access network. How to dimension the aggregation, metro, and core network to achieve this is beyond the scope of this paper.

Streaming Services

High fidelity streaming services pose the highest demands on the current access networks. As such, any improvement in codec performance currently is readily embraced in order to transport these services more efficiently. It is very likely that these services will continue to be designed with the audiovisual bit rate bounds in mind for some time to come. However, in a more distant future, there still exists the possibility that the resolution (and frame rate) might be set higher than strictly needed, in effect wasting transport capacity on unobservable details. Additionally, some studies (e.g., [24]) have argued that in the future a user will consume several visual information flows simultaneously, in the same view. Even if this is the case, this does not necessarily mean that a bit rate much higher than our audiovisual bit rate bound is needed. In fact, one specific information flow out of the set of all simultaneous information flows will cover only part of the complete visual angle of the observer. The angle covered will be smaller than the one used when the observer would view that particular visual flow alone. Since the (spatial) resolution of this specific information flow is smaller, in principle, a lower bit rate is required for that flow, such that the bit rate of all flows together will be close to the specified audiovisual bit

rate bound. Finally, in 3D applications, more views might be transported than the two views a user consumes (one for each eye) in anticipation that the user might rapidly shift his or her attention from one viewpoint to another. If this shift of attention cannot be detected and reacted to in less than a blink of an eye, these additional viewpoints really are needed. (Remember, however, that in this paper we discuss the required bit rate per user, and that views consumed by different observers are counted in the budget of each person individually).

Data Services

Up to now, data services have not been designed with the audiovisual bit rate bounds in mind and it is less likely for streaming services that they ever will be. For current data services, a file, e.g., a text document, a presentation, or an e-mail message, that happens to reside in the network needs to be downloaded completely to the device on which the user wants to manipulate (i.e., view, listen, edit) it before he or she can do so. For small objects, this method of operation generates only a small bit rate, and hence, this presents no problem. However, for larger objects, e.g., objects containing a large number of images, or for video footage, this results in lengthy download times, which may become unacceptable in the future. One way to solve this is to increase the access bit rate, but there is also an alternative option. If data files were to be organized such that only the part the user manipulates would be transported in a blink of an eye to his or her local device, this prohibitively long download time would no longer pose a problem. In fact, with such file formats, the difference between data services and streaming services is blurring. In order to make the user manipulations transparently interactive, the effect of these manipulations needs to have an audiovisual effect in a blink of an eye, which can be done if the information is suitably compressed, provided the access link can support a bit rate just larger than the audiovisual bit rate bounds.

Progressive File Formats

If the observer wants to manipulate the visual information in his or her visual field of view in a very flexible way, e.g., to scale various pieces of visual information up or down, and if this information needs to be readily available (i.e., in a blink of an eye), codecs that encode the audiovisual information such that multiple resolutions are easily accessible are helpful. These codecs, referred to as scalable codecs, exist:

- The adaptive multi-rate wideband (AMR-WB) speech codec is a standard for speech signals [1],
- JPEG-2000 [12] describes a standard way of accessing multiple spatial resolutions and levels of color fidelity for still images (with an extension to sequences), and
- Annex G of ITU-T H.264 [15] (of which [28] provides an overview) was recently adopted to encode video sequences in such a way that multiple spatial and temporal resolutions, as well as various levels of color fidelity are easily accessible.

State-of-the-Art Access Network Technologies

In this section, we briefly describe commonly deployed access networks. A more in-depth analysis is outside the scope of this paper but can be found (including wireless access technologies) in [7]. We assess how much delay is introduced on and which bit rate can be reached over the last mile. On top of the delay in the last mile there is some additional delay in the network (and codec). For a well-designed network, the queuing delay should be small, such that the only delay that matters is the propagation delay of about 5 ms per 1000 km.

Before we discuss the access technologies, we point to the difference between the net and gross bit rate.

Overhead Bit Rate

The bit rates mentioned in previous paragraphs are net bit rates. In order to transport these net bit rates over a packet-based network, an overhead bit rate is required.

A first type of overhead bit rate stems from the fact that each packet needs:

- An address, typically 20 bytes in the Internet Protocol (IP) version 4 (IPv4) and 40 bytes in IP version 6 (IPv6),
- An identifier to distinguish the packets from various applications running on the same computer, and

- Often a time stamp and sequence number to identify the temporal relation of the packets and to detect packet loss respectively.

For the transport of video information, the payloads of the packets are typically just less than 1,500 bytes, such that the overhead bit rate is limited to a few percent. However, for low bit rate audio services where the packetization delay plays an important role [16], this overhead can be substantial.

A second type of overhead is needed to protect the information flow from errors incurred during transport. Often transport channels are prone to bit errors due to noise, which can in turn cause packet loss. In order to correct bit errors or to recuperate lost packets, additional bit rate is needed. Either a retransmission or forward error correction (FEC) scheme can be used [5]. The overhead required to protect the packet flow depends very much on the statistics of the bit error and packet loss process. Overhead bit rates of up to 10 percent are not uncommon.

DSL

Digital subscriber line technology enables bidirectional transport of broadband traffic over the legacy twisted pairs that were traditionally used for offering telephony services. **Figure 3** shows the DSL network architecture.

In the downstream direction, the DSL access multiplexer (DSLAM) routes the packets destined for each specific home gateway (HG) and modulates this bit stream (over a high frequency band) on the twisted pair. During transport of the bits, errors may occur due to the cross-talk between the twisted pairs in the same binder or due to impulsive noise. There are various ways to keep the bit errors and associated packet loss under a desired bound [5]. Typically, they introduce an overhead bit rate of a few percentage points and a delay of 10 to 20 ms. In the upstream direction, the transport of the packets over the associated bit stream is very similar to the downstream direction, only a different frequency band is used.

Various versions exist for DSL. While the original asymmetric DSL (ADSL) could support a bit rate of about 8 Mbps downstream and 1 Mbps upstream, its enhanced version increases the downstream bit rate to more than 20 Mbps by doubling the downstream frequency band. The very high speed DSL (VDSL) increases the downstream and upstream bit rate to even higher values by using a wider frequency band and allowing a flexible way to distribute the frequency bands both upstream and downstream. The bit rate that can be reached very much depends on the length of the twisted pair. In-house VDSL can support up to 100 Mbps symmetrically, while today's commercial offerings range from 20 Mbps to

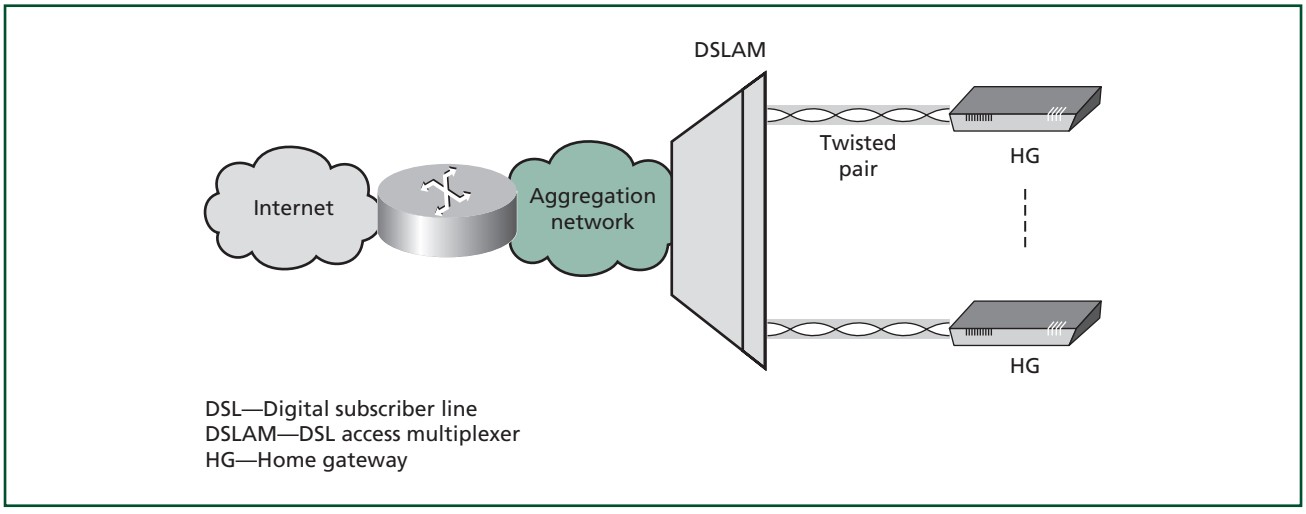
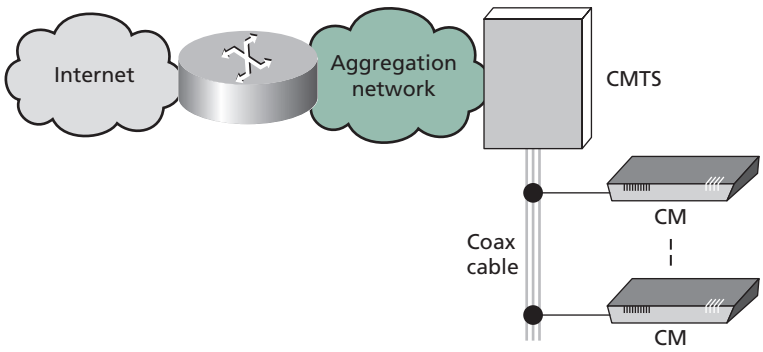


Figure 3.
DSL access network.



CM—Cable modem
CMTS—Cable modem termination system
DOCSIS—Data over cable service interface specification

Figure 4.
DOCSIS access network.

50 Mbps downstream, and a few megabits per second upstream.

DOCSIS

The data over cable service interface specification (DOCSIS) allows bi-directional transport of broadband traffic over a coaxial cable that was normally used for the transport of analog television signals. **Figure 4** shows the DOCSIS network architecture.

In the downstream direction, the cable modem termination system (CMTS) broadcasts the packets destined for all cable modems (CMs) in one or more frequency bands normally used to carry analog television (TV) signals. Each CM only picks out its own packets. Currently about 32 Mbps to 40 Mbps can be carried per frequency band and this needs to be shared by all CMs on this frequency band. A coaxial cable is less prone to noise than a twisted pair, such that a correction scheme does not need to be very strong, thus yielding only a low overhead bit rate.

In the upstream direction, all CMs make use of the same upstream frequency band. As such, a medium access control (MAC) mechanism is required to regulate the requests from all CMs served by the CMTS. This has two consequences. First, the upstream bit rate has to be shared, and as such, depends on the number of CMs a CMTS serves, often referred to as the number of homes passed. Second, the delay introduced on an upstream packet is variable. In [17], it was shown through measurements that although the minimal delay is about the same as in DSL, the delay variation is quite large (up to 100 ms).

PON

The high level architecture of the passive optical network (PON) is shown in **Figure 5**. The downstream and upstream direction use different wavelengths for the transport of bits. Various versions of PON exist, of which the most prominent are Gigabit PON (GPON) standardized by the International Telecommunication Union (ITU) and Ethernet PON (EPON) standardized by the Institute of Electrical and Electronics Engineers (IEEE).

In the downstream direction, the optical line terminator (OLT) broadcasts packets destined for all optical network units (ONUs). The splitter copies all received bits over its outgoing fibers. Each ONU only picks out its own packets.

The upstream direction is shared by all ONUs. Just as in the DOCSIS-based system, a MAC protocol is required to regulate the requests of ONUs for sending upstream packets. However the PON bit rates are so large that the resulting delay and delay variation are low enough [23].

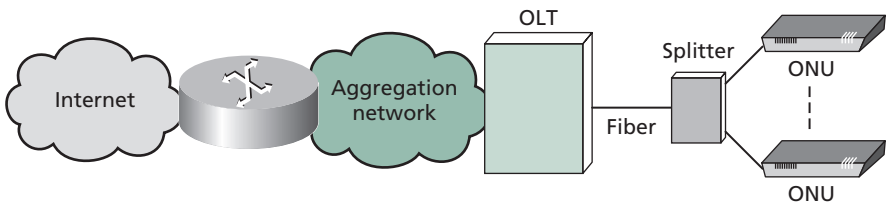


Figure 5.
PON access network.

PON uses some FEC, but much less than DSL.

The bit rates offered to the ONU by a PON system depend on the splitting factor, which is typically 32 but can be up to 128. As the bit rate on an optical fiber is typically 1 Gbps to 2.5 Gbps, PON can offer a bit rate of up to 100 Mbps per ONU, over lines of typically 10 to 20 km.

Access Technology Summary

If the access technologies are used properly, they all can attain the desired sustainable bit rate to serve at least one user and often more users. In DSL, the twisted pairs should not be too long (1 km or so), a good protection scheme should be used, and the VDSL version should be used. In DOCSIS, the number of homes passed should be kept low, and similarly in PON, the splitting factor should be small.

None of the access technologies has a problem in attaining a 300 ms round-trip delay, provided the traffic load on the aggregation network is not too high, and provided that the information the user wants to access is not too far away for the propagation delay of 5 ms per 1,000 km to become unworkably large.

Conclusions

In this paper, we have used the limitations of human perception to estimate the bit rate a human observer in principle needs to view and listen to audiovisual information at the highest quality. We argued that with perfect codecs, a net data rate on the order of 10 Mbps for visual information and the order of 100 kbps for aural information per user are sufficient to provide the user with a signal that is perceptually virtually indistinguishable from the signals that he or she would perceive if he or she was really present at the scene. When a user wants to manipulate information, we have shown that a round-trip delay of 300 ms is low enough to provide full transparent interactivity for practically all applications.

We further claimed that while currently only streaming services are designed with these bounds in mind (because they put the highest demand on the current networks), future data services, where large objects are manipulated, could also benefit from respecting these bounds, provided that the data files are properly organized. More specifically, the audiovisual information the user is manipulating should be stored and organized in such a way that each part can be easily extracted. Standards that aim to achieve this goal were discussed.

Finally, we considered commonly deployed wire-bound access network technologies and came to the conclusion that state-of-the-art access networks when properly designed can easily offer the bit rate needs and delay bounds necessary to serve at least one simultaneous user.

References

[1] 3rd Generation Partnership Project, "Performance Characterization of the Adaptive Multi-Rate Wideband (AMR-WB) Speech Codec," 3GPP TR 26.976, v7.0.0, June 2007,

<http://www.3gpp.org/ftp/Specs/html-info/26976.htm>.

[2] B. T. Backus, D. J. Fleet, A. J. Parker, and D. J. Heeger, "Human Cortical Activity Correlates With Stereoscopic Depth Perception," J. Neurophysiol., 86:4 (2001), 2054–2068.

[3] F. Baumgarte, "A Computationally Efficient Cochlear Filter Bank for Perceptual Audio Coding," Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processingbbb (ICASSP '01) (Salt Lake City, UT, 2001), vol. 5, pp. 3265–3268.

[4] D. De Vleeschauwer, "On the Smoothness Constraint in the Intensity-Based Estimation of the Parallax Field," Multidimens. Syst. Signal Process., 6:2 (1995), 113–135.

[5] N. Degrande, D. De Vleeschauwer, and K. Laevens, "Protecting IPTV Against Packet Loss: Techniques and Trade-Offs," Bell Labs Tech. J., 13:1 (2008), 35–51.

[6] C. Dixon, Bandwidth Challenges to the Digital Home, The Diffusion Group, TDG dBrief, 2007.

[7] M. Fijnvandraat and H. Bouwman, "Flexibility and Broadband Evolution," Telecommun. Policy, 30:8/9 (2006), 424–444.

[8] L. Haglund, N. Guest, S. Einerman, H. Öster, P. Björkman, and H. Graf, Overall-Quality Assessment When Targeting Wide-XGA Flat Panel Displays: Test Results and Possible Implications for Broadcasting in Europe, SVT Corporate Dev. Technol., Apr. 2002, <http://www.ebu.ch/CMSimages/en/tec_svt_widexga_final_tcm6-44922.pdf>.

[9] H. Hoffmann, T. Itagaki, D. Wood, and A. Bock, "Studies on the Bit Rate Requirements for a HDTV Format With 1920 × 1080 Pixel Resolution, Progressive Scanning at 50 Hz Frame Rate Targeting Large Flat Panel Displays," IEEE Trans. Broadcasting, 52:4 (2006), 420–434.

[10] C. Holliday, B&C Consulting, How Much Bandwidth Is Enough in the Access Network? Strategies of AT&T, Verizon and Bellsouth in the Design of the Last Mile, Information Gatekeepers, Aug. 2006.

[11] International Organization for Standardization and International Commission on Illumination, "CIE Standard Colorimetric Observers," ISO/CIE 10527, 2007.

[12] International Organization for Standardization and International Electrotechnical Commission, "Information Technology—JPEG 2000 Image Coding System: Core Coding System," ISO/IEC 15444-1, Sept. 2004.

[13] International Telecommunication Union, Radiocommunication Sector, "The Present State of High-Definition Television," ITU-R BT.801-4, 1990.

[14] International Telecommunication Union, Telecommunication Standardization Sector, "One-Way Transmission Time," ITU-T Rec. G.114, May 2003.

[15] International Telecommunication Union, Telecommunication Standardization Sector, "Advanced Video Coding for Generic Audiovisual Services," ITU-T Rec. H.264, Nov. 2007.

[16] J. Janssen, D. De Vleeschauwer, M. Büchli, and G. H. Petit, "Assessing Voice Quality in Packet-Based Telephony," IEEE Internet Comput., 6:3 (2002), 48–56.

[17] T. Jehaes, D. De Vleeschauwer, T. Coppens, B. Van Doorselaer, E. Deckers, W. Naudts, K. Spruyt, and R. Smets, "Access Network Delay in Networked Games," Proc. 2nd Workshop on Network and Syst. Support for Games (NetGames '03) (Redwood City, CA, 2003), pp. 63–71.

[18] D. H. Kelly, "Spatio-Temporal Frequency Characteristics of Color-Vision Mechanisms," J. Opt. Soc. Amer., 64:7 (1974), 983–990.

[19] C. Kim and J. B. Ra, "Noninteger View Multiplexing for 3D Lenticular Display," Proc. 3DTV Conf. (3DTV-CON '07) (Kos Island, Gr., 2007).

[20] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry II, V. Balasubramanian, and P. Sterling, "How Much the Eye Tells the Brain," Curr. Biology, 16:14 (2006), 1428–1434.

[21] H. Kolb, "How the Retina Works," Amer. Scientist, 91:1 (2003), 28–35.

[22] R. Kooij, K. Ahmed, and K. Brunnström, "Perceived Quality of Channel Zapping," Proc. 5th IASTED Internat. Conf. on Commun. Syst. and Networks (CSN '06) (Palma de Mallorca, Sp., 2006), pp. 155–158.

[23] G. Kramer, B. Mukherjee, and G. Pesavento, "IPACT: A Dynamic Protocol for an Ethernet PON (EPON)," IEEE Commun. Mag., 40:2 (2002), 74–80.

[24] Motorola, "Anticipating the Bandwidth Bottleneck: Looking at Bandwidth Usage Five Years Forward," White Paper, May 2007.

[25] E. Nakasu, Y. Nishida, M. Maeda, M. Kanazawa, S. Yano, M. Sugawara, K. Mitani, K. Hamasaki, and Y. Nojiri, "Technical Development Towards Implementation of Extremely High Resolution

Imagery System With More Than 4000 Scanning Lines," Proc. Internat. Broadcasting Conv., (IBC '06) (Amsterdam, Neth., 2006), pp. 345–352.
[26] M. Philpott and J. Coham, Benchmarking Broadband: Top-Ten Most Competitive Markets, Ovum, Feb. 8, 2008.
[27] J. E. Raymond, K. L. Shapiro, and K. M. Arnell, "Temporary Suppression of Visual Processing in an RSVP Task: An Attentional Blink?" J. Exp. Psych.: Human Percept. Perform., 18:3 (1992), 849–860.
[28] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," IEEE Trans. Circuits Syst. for Video Technol., 17:9 (2007), 1103–1120.
[29] C. E. Shannon, "Prediction and Entropy of Printed English," Bell Syst. Tech. J., 30:1 (1951), 50–64.
[30] E. G. Sheffield, J. Kean, M. Starling, J. Andrews, K. Evans, and S. Khemlani, "Results From Subjective Testing of the HD Codec at 16–96 kbps," IEEE Trans. Broadcasting, 52:2 (2006), 219–222.
[31] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, K. Sönmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu, "Recent Innovations in Speech-to-Text Transcription at SRI-ICSI-UW," IEEE Trans. Audio, Speech, and Language Processing, 14:5 (2006), 1729–1744.
[32] M. Sugawara, K. Mitani, F. Kanazawa, F. Okano, and Y. Nishida, "Future Prospects of HDTV—Technical Trends Toward 1080p," SMPTE Motion Imaging J., 115:1 (2006), 10–15.
[33] University of Utah, John Moran Eye Center, "Facts and Figures Concerning the Human Retina," <http://webvision.med.utah.edu/Facts.html>.
[34] F. VanderWerf, P. Brassinga, D. Reits, M. Aramideh, and B. Ongerboer de Visser, "Eyelid Movements: Behavioral Studies of Blinking in Humans Under Different Stimulus Conditions," J. Neurophysiol., 89:5 (2003), 2784–2796.
[35] L. Wang, F. Yin, and Z. Chen, "HRTF Compression via Principal Components Analysis and Vector Quantization," IEICE Electronics Express, 5:9 (2008), 321–325.

(Manuscript approved December 2008)

DANNY DE VLEESCHAUWER is a network strategist in the Network & Technology Strategy (NTS) group of Bell Labs with Alcatel-Lucent Bell in Antwerp, Belgium. He received an M.Sc. in electrical engineering and a Ph.D. degree in applied sciences from Ghent University in Belgium. Prior to joining Alcatel-Lucent, Dr. De Vleeschauwer was a researcher at Ghent University. His early work was on image processing, and he worked later on the application of queuing theory in packet-based networks. His current research focus is on ensuring adequate quality for triple-play services offered over packet-based networks. He is a guest professor in the Telecommunications and Information Processing department (TELIN) of Ghent University, and a member of the Alcatel-Lucent Technical Academy. ◆