# MPEG-4
## Multimedia for our time

ROB KOENEN
*KPN Research*

Consider the kinetophonograph, one of the greatest multimedia inventions of all time. Invented in 1889 by William Dickson, it reproduced brief flickering images with barely synchronized sound. Dickson's employer, Thomas Edison, reportedly thought its sound and image quality poor, and let the device languish. Robust sound cinema would not appear for another 40 years.

Despite its tremendous potential, today's revolutionary machine for multimedia, the Internet, bears an unfortunate resemblance to that aborted project. And other types of communication channels, such as mobile phones, are hardly in the picture. But this time, the great opportunity will not go unexploited, courtesy of MPEG-4, the name given to a revolutionary communications standard released this month (the acronym is pronounced "M-Peg").

The standard, developed over five years by the Moving Picture Experts Group (MPEG) of the Geneva-based International Organization for Standardization (ISO), explores every possibility of the digital environment. Recorded images and sounds co-exist with their computer-generated counterparts; a new language for sound promises compact-disk quality at extremely low data rates; and the multimedia content could even adjust itself to suit the transmission rate and quality.

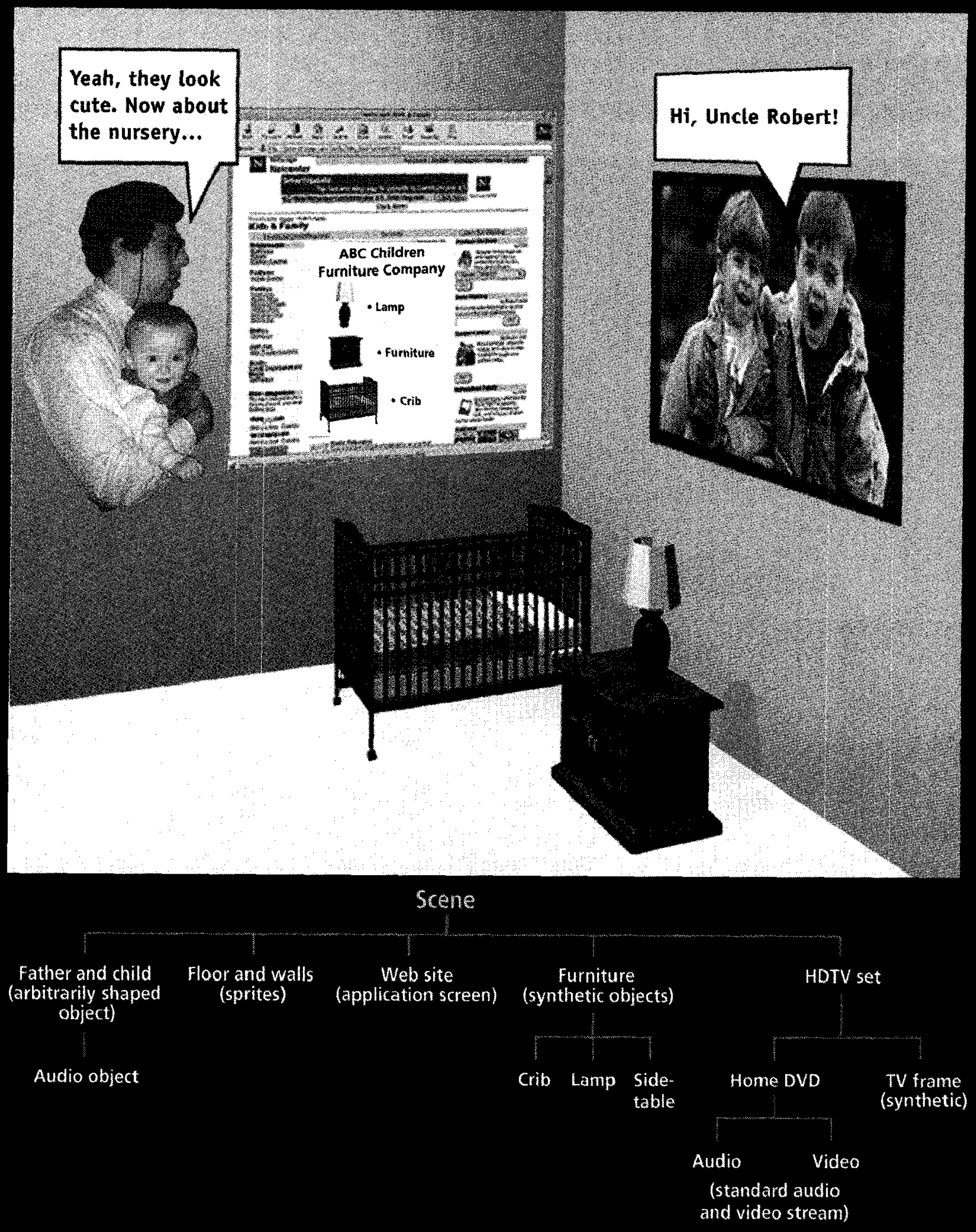
Possibly the greatest of the advances made by MPEG-4 is that viewers and listeners need no longer be passive. The height of "interactivity" in audiovisual systems today is the user's ability merely to stop or start a video in progress. MPEG-4 is completely different: it allows the user to interact with objects *within* the scene, whether they derive from so-called real sources, such as moving video, or from synthetic sources, such as computer-aided design output or computer-generated cartoons. Authors of content can give users the power to modify scenes by deleting, adding, or repositioning objects, or to alter the behavior of the objects; for example, a click on a box could set it spinning.

Perhaps the most immediate need for MPEG-4 is defensive. It supplies tools with which to create uniform (and top-quality) audio and video encoders and decoders on the Internet, preempting what may become an unmanageable tangle of proprietary formats. For example, users must choose among video formats such as QuickTime (from Apple Corp., Cupertino, Calif.), AVI (from Microsoft Corp., Redmond, Wash.), and RealVideo (from RealNetworks Inc., Seattle, Wash.)—as well as a bewildering number of formats for audio.

In addition to the Internet, the standard is also designed for low bit-rate communications devices, which are usually wireless. For example, mobile receivers and "Dick Tracy" wristwatches with video will have far greater success now that the standard is in place. But whether

**THE LATEST MULTIMEDIA STANDARD EXCELS AUDIOVISUALLY, HUSBANDS EVERY BIT, AND INVITES THE VIEWER TO JOIN THE ON-SCREEN ACTION**

0018-9235/99/$10.00©1999 IEEE

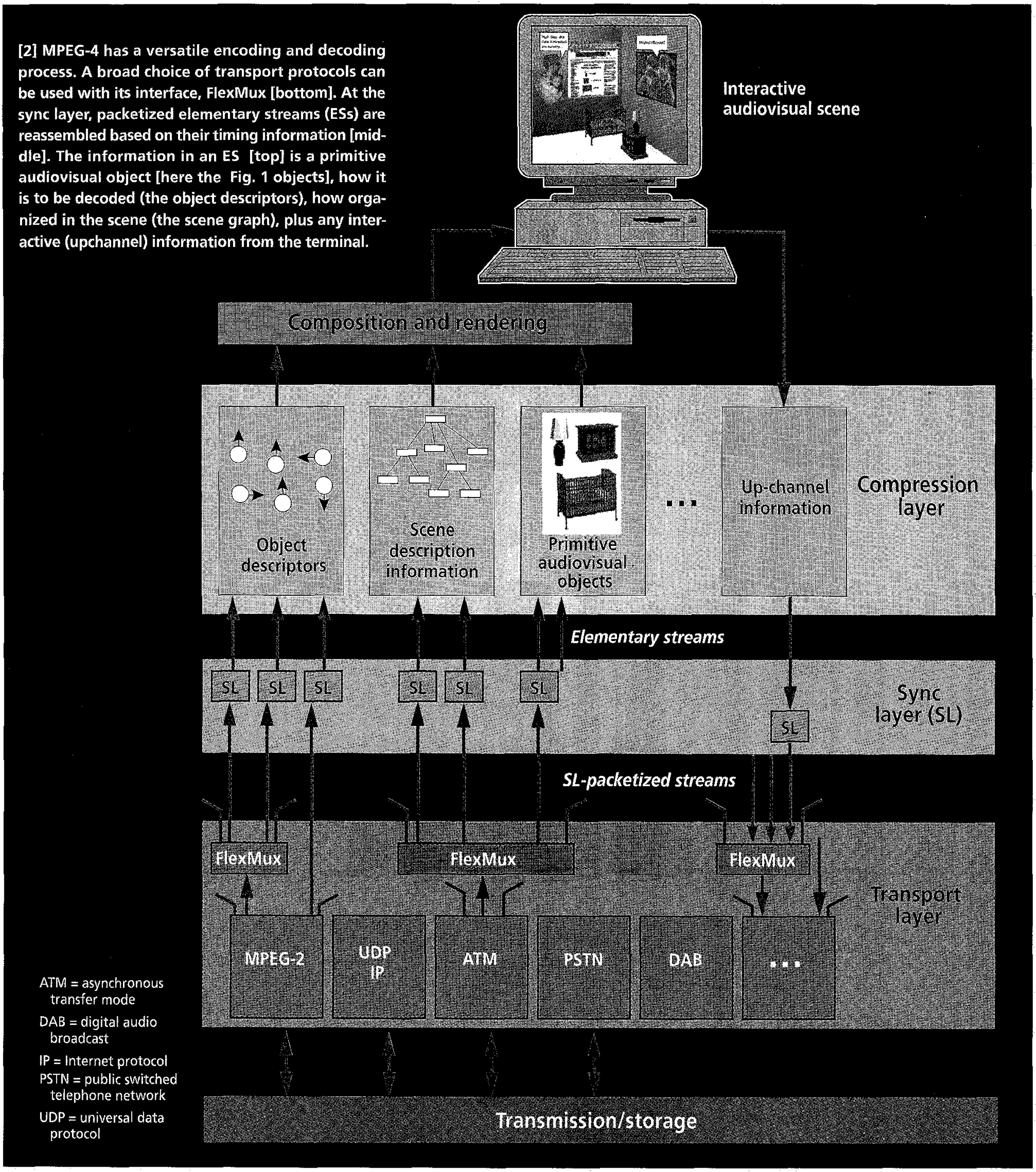

Yeah, they look cute. Now about the nursery...
ABC Children Furniture Company
• Lamp
• Furniture
• Crib
Hi, Uncle Robert!
Scene
Father and child (arbitrarily shaped object)
Audio object
Floor and walls (sprites)
Web site (application screen)
Furniture (synthetic objects)
Crib
Lamp
Side-table
HDTV set
Home DVD
Audio
Video
(standard audio and video stream)
TV frame (synthetic)

[1] Different types of multimedia that can be transmitted with MPEG-4 appear in the scene above, a man and his infant son on-line with his offstage wife. The tree chart below, called a scene graph, represents the media as independent or compound objects. One compound object comprises the father and child (an arbitrarily shaped video) and the audio track of his voice. Other objects are the floor and walls—"sprites," here used for easily changed backgrounds; the Web site of the furniture store—an application mapped as a screen texture; and the computer-generated (synthetic) furniture the father has chosen from the Web site for his wife to look at and interactively move around. Simultaneously playing on a synthetic HDTV set is a movie from the family's home digital versatile disk (DVD).

wired or not, devices can have differing access speeds depending on the type of connection and traffic. In response, MPEG-4 supports scalable content, that is, it allows content to be encoded once and automatically played out at different rates with acceptable quality for the communication environment at hand.

On the other end of the quality/bit-rate scale, future television sets will no doubt accept content from both broadcast and interactive digital sources. Accordingly, MPEG-4 provides tools for seamlessly integrating broadcast content with equally high-quality interactive MPEG-4 objects. The expectation is for content of broadcast-grade quality to be displayed within World Wide Web screen layouts that are as varied as their designers can make them. However, the standard's potential for encoding individual objects with the extremely high quality needed in studios has the recording industries very much on the alert.

Recently, digital copying of audio from the Internet has become a popular and—to the music industry—increasingly worrying practice. For video, the same situation will arise when MPEG-4 encoding and higher bandwidths become widespread and as digital storage prices continue to drop. Accordingly, MPEG designed in features for protection of intellectual property and digital content [see "Protecting the property," next page].



**[2] MPEG-4 has a versatile encoding and decoding process. A broad choice of transport protocols can be used with its interface, FlexMux [bottom]. At the sync layer, packetized elementary streams (ESs) are reassembled based on their timing information [middle]. The information in an ES [top] is a primitive audiovisual object [here the Fig. 1 objects], how it is to be decoded (the object descriptors), how organized in the scene (the scene graph), plus any interactive (upchannel) information from the terminal.**

ATM = asynchronous transfer mode
DAB = digital audio broadcast
IP = Internet protocol
PSTN = public switched telephone network
UDP = universal data protocol

### No time to rest

Does the world need yet another standard for audio or visual information? The history of the MPEG standards in itself provides an answer. MPEG-1, which debuted in 1992 and is still widely used in the Far East, is essentially a digital video player: it plays out audio and video in linear streams, allowing the same type of access as a home VCR. (So-called streaming video or audio implies that the content need not be downloaded in full before it begins playing, but is played out even as it is being received and decoded.)

In 1995 MPEG-2 was introduced for compression and transmission of digital television signals (the digital versatile disk [DVD] also uses MPEG-2 coding). But, although MPEG-2 can be used to control streams from a server or to retrieve data from a broadcast carrousel, as in MPEG-1 its audio and video are still inherently linear: interactivity is limited to operating common VCR modes such as fast forward or slow motion.

The next MPEG standard is MPEG-4, the extraordinary strength of which stems from its radical object-oriented paradigm [see "MPEG-7 and the missing numbers," p. 31, for more information on where MPEG standards have been and where they are going].

### The utility of objects

At the atomic level, to use a chemical analogy, the audio and video components of MPEG-4 are known as objects. These can exist independently, or multiple ones can be grouped together to form higher-level audiovisual bonds, to coin a phrase. The grouping is called composition, and the result is an MPEG-4 scene [Fig. 1]. The strength of this so-called object-oriented approach is that the audio and video can be easily manipulated.

Visual objects in a scene are described mathematically and given a position in a two- or three-dimensional space. Similarly, audio objects are placed in a sound space. When placed in 3-D space, the video or audio object need only be defined once; the viewer can change his vantage point, and the calculations to update the screen and sound are done locally, at the user's terminal. This is a critical feature if the response is to be fast and the available bit-rate is limited, or when no return channel is available, as in broadcast situations.

MPEG-4's language for describing and dynamically changing the scene is named the Binary Format for Scenes (BIFS). BIFS commands are available not only to add objects to or delete them from the scene, but also to change visual or acoustic properties of an object without changing the object in itself; thus the color alone of a 3-D sphere might be varied.

BIFS can be used to animate objects just by sending a BIFS command and to define their behavior in response to user input at the decoder. Again, this is a nice way to build interactive applications. In principle, BIFS could even be used to put an application screen (such as a Web browser's) as a "texture" in the scene.

BIFS borrows many concepts from the Virtual Reality Modeling Language (VRML), which is the method used most widely on the Internet to describe 3-D objects and users' interaction with them. BIFS and VRML can be seen as different representations of the same data. In VRML, the objects and their actions are described in text, as in any other high-level language. But BIFS code is binary, and thus is shorter for the same content—typically 10 to 15 times.

More important, unlike VRML, MPEG-4 uses BIFS for real-time streaming, that is, a scene does not need to be downloaded in full before it can be played, but can be built up on the fly. Lastly, BIFS allows defining 2-D objects such as lines and rectangles—something currently not possible in VRML. (MPEG and the Web 3-D Consortium are working together on making MPEG-4 and VRML evolve consistently.)

To give even more support for interactive applications, the next release of the MPEG-4 standard, which is to be published in December 1999, will define MPEG-J. This is an MPEG-4-

## Protecting the property

Commerce in multimedia content on the Internet will never flourish unless means are found to combat illegal copying. MPEG-4's role here is twofold. First, a data field for the identification of intellectual property (IP) is contained in the descriptor attached to the elementary stream that is used to transport each audio or visual object. Second, the standard specifies interfaces to the proprietary systems that can manage and protect IP, like the conditional access systems for pay TV services.

Probably the best candidates for identifiers are those issued in international systems such as the International Standard Audio-Visual Number, which is to audiovisual material what the International Standard Book Number (ISBN) is to books. Ownership information may change when rights are sold, but with such a number the current holder of the rights to it can always be identified.

Not all content carries such a number, though. Instead this field can be used to identify IP by a number of key-value pairs, such as "Composer/John Lennon." MPEG's parent standards body, the International Organization for Standardization (ISO), cannot force the key-value pairs to be correct or even present. A criminal might be able to strip the data and still produce a syntactically valid MPEG-4 bit-stream. Enforcing correctness here can be achieved only through legislation, not by standardization.

The outlook for strictly technological defense is not as bad as it might seem, however. MPEG-4 provides hooks to proprietary management and protection schemes, which would probably deploy encryption of the content and embedded IP information. It would be unwise to standardize these proprietary schemes themselves; the design of secure overarching systems should be under the control of industry groups with particular application needs; furthermore, these schemes are often attacked and can become insecure, crippling the entire standard.

In addition to IP associated with content that must be managed and protected, there is the IP embodied in the encoding and decoding algorithms. In short, companies want recompense for any work of theirs that appears in their implementation of this very MPEG standard. With MPEG-2, it is relatively simple: encoders and decoders come in the form of hardware, and for every box sold, IP holders receive a small patent fee.

Although there will be hardware MPEG-4 decoders, many players will take the form of software packages, and it will be much more difficult to collect these fees for them. Hence, it may be necessary at least to audit the copying of software implementations of MPEG-4 players, and sometimes to disable it altogether. Information then could be embedded in MPEG-4 streams to manage or prohibit proliferation of the associated players.

However, in many business models, software decoders are to be given away free of charge, and people encouraged to spread copies rather than prevented them from doing so; the earnings would come from selling content. Some MPEG members are therefore investigating ways in which the patent fees can be dealt with jointly with content royalties. (Under ISO's bylaws, MPEG as a whole is forbidden to deal with patents.)

In these plans, patent holders would get, for instance, a small percentage of the content revenues, and decoders could indeed be freely distributed. Some manufacturers with IP invested in MPEG-4 apparently find it hard to say goodbye to the old, hardware-centric, "a-few-cents-per-device" IP model. In the Internet world, however, this model is certainly difficult, if not doomed.

—R.K.

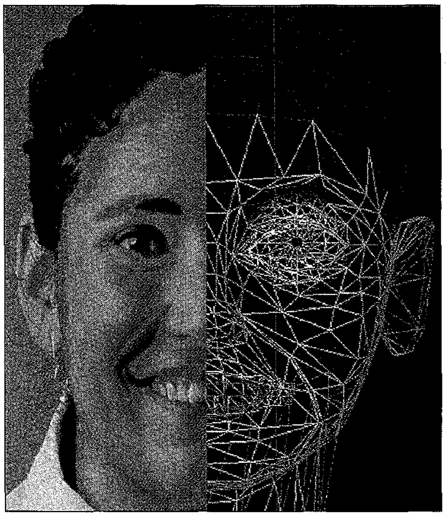
specific subset of the object-oriented Java language. MPEG-J defines interfaces to elements in the scene (objects or compound objects), network resources (such as available bandwidth and bit-error levels), terminal resources (such as processing power and stack memory), and input devices.

The Java interfaces automatically let the content scale down in an intelligent way relative to the player, and thus will allow authors not only to create highly interactive multimedia content, but also to make optimal use of terminal and network resources. For instance, the background might be decoded only once instead of continuously, or a few not-so-important objects might be omitted altogether.

## Wrapping the data

Just as MPEG-4's representation of multimedia content is new and versatile, so is its scheme for preparing that content for transportation or storage (and, logically, for decoding) [Fig 2]. Here, objects are placed in so-called elementary streams (ESs). Some objects, such a sound track or a video, will have a single such stream. Others objects may have two or more. For instance, a scalable object would have an ES for basic-quality information plus one or more enhancement layers, each of which would have its own ES for improved quality, such as video with finer detail or faster motion.

Higher-level data describing the scene—the BIFS data defining, updating and positioning the media objects—is conveyed in its own ES. Here again the virtue of the hierarchical, object-based conceptions in MPEG-4 can be seen: it is easier to reuse objects in the production of new multimedia content, or (to say it another way) the new production is easier to modify without changing an encoded object itself. If parts of a scene are to be delivered only under certain conditions, say, when it is determined that enough bandwidth is available, multiple scene description ESs for the different circumstances may be used to describe the same scene.



[3] MPEG-4 allows video images to be laid over wire-frame faces to create avatars of users. Speech with naturalistic cadences can be generated from written text in step with motions of the mouth, eyes, and lips.

To inform the system which elementary streams belong to a certain object, MPEG-4 uses the novel, critical concept of an object descriptor (OD). Object descriptors in their turn contain elementary stream descriptors (ESDs) to tell the system what decoders are needed to decode a stream. With another field, optional textual information about the object can be supplied. Object descriptors are sent in their own, special elementary stream, which allows them to be added or deleted dynamically as the scene changes.

The play-out of the multiple MPEG-4 objects is coordinated at a layer devoted solely to synchronization. Here, elementary streams are split into packets, and timing information is added to the payload of these packets. These packets are then ready to be passed on to the transport layer.

## Streams here, streams there

Timing information for the decoder consists of the speed of the encoder clock and the time stamps of the incoming streams, which are relative to that clock. Two kinds of time stamps exist: one says when a piece of information must be decoded, the other says when the information must be ready for presentation.

The distinction between the types of stamp is important. In many video compression schemes, some frames are calculated as an interpolation between previous and following frames. Thus, before such a frame can be decoded and presented, the one after it must be decoded (and held in a buffer). For predictable decoder behavior, a buffer model in the standard augments the timing specification.

In terms of the ISO seven-layer communications model, no specific transport mechanism is defined in MPEG-4. Existing transport formats and their multiplex formats suffice, including the MPEG-2 transport stream, asynchronous transfer mode (ATM), and real-time transport protocol (RTP) on the Internet. Incidentally, the fact that the MPEG-2 transport stream is used by digital TV has the important consequence of allowing co-broadcast modes.

A separate transport channel could be set up for each data stream, but there can be many of these for a single MPEG-4 scene, and as a result the process could be unwieldy and waste bits. To remedy matters, a small tool in MPEG-4, FlexMux, was designed to act as an intermediate step to any suitable form of transport. In addition, another interface defined in MPEG-4 lets the application ask for connections with a certain quality of service, in terms of parameters like bandwidth, error rate, or delay.

From the application's point of view, this interface is the same for broadcast channels, interactive sessions, and local storage media. Application designers can therefore write their code without having to worry about the underlying delivery mechanisms. Further, the next release of the standard will allow differing channels to be used at either end of a transmission/receive network, say, an Internet protocol channel on one end and an ATM one on the other.

Another important addition in Version 2 is a file format known as mp4, which can be used for exchange of content and which is easily converted. MPEG-1 and MPEG-2 did not include such a specification, but the intended use of MPEG-4 in Internet and personal computer environments makes it a necessity. It will be the only reliable way for users to exchange complete files of MPEG-4 content.

## Objectifying the visual

Classical, "rectangular" video, as the type that comes from a camera may be called, is of course one of the visual objects defined in the standard. In addition, objects with arbitrary shapes can be encoded apart from their background and then placed before other video types.

In fact, MPEG-4 includes two ways of describing arbitrary shapes, each appropriate to a different environment. In the first, known as binary shape, an encoded pixel (of a certain color, brightness, and so on) either is or is not part of the object in question. A simple but fairly crude technique, it is useful in low bit-rate environments, but can be annoying—the edges of pixels are sometimes visible, and curves have little jagged steps, known as aliasing or "the jaggies."

For higher-quality content, a shape description known as gray scale, or alpha shape, is used. Here, each pixel belonging to a shape is not merely on or off, but is assigned a value for its transparency. With this additional feature, transparency can differ from pixel to pixel of an object, and objects can be smoothly blended, either into a background or with other visual objects.

One instance of smooth blending can be seen in most television weather reports. The weatherman's image seems to be standing in front of a map, which in fact is generated elsewhere. Not surprisingly, then, manufacturers of television studio equipment have expressed an interest in the capabilities for arbitrary shape objects and scene description since, conceptually, they closely match the way things are already done in a studio. In fact, MPEG video has started working on bit-rates and quality levels well beyond the high-definition television (HDTV) level that can already be achieved.

Note that MPEG does not specify how shapes are to be extracted. Doing this automatically—video segmentation, as it is known—is still a matter of intensive research. Current methods still have limitations but there are ways to get the job done: the best way of obtaining shaped objects such as the weatherman is recording them with a blue or green background, colors that can easily be filtered out.

(Actually, MPEG-4, like its predecessors, specifies only the decoding process. Encoding processes, including any improvements, are left to the marketplace. Even today, improvements in MPEG-2 compression quality sometimes show up, even though that standard was cast in stone several years ago.)

In another nod to the broadcast world, somewhat reluctantly the MPEG committee decided to support interlaced as well as progressive scanning, which is the type used in computer monitors. To the chagrin of the computer industry, interlaced video content will be around for quite some time to come.

## Thinking small

On the other end of the bit-rate spectrum, a great deal of effort has gone into making moving video possible at very low bit-rates, notably for mobile devices. MPEG-4 has been found usable for streaming wireless video transmission (making use of GSM/Global System for Mobile Communications) at 10 kb/s—the data rate in GSM currently used for voice communications!

A bothersome feature of mobile operation is the probability of transmission errors, due to the low redundancy of the data. Various techniques are used in the standard both to overcome the inevitable errors as soon as possible and to enable the decoder to mask the results of errors. One such technique is using resync markers in the video bit-stream. With these, the synchronization lost after an error can be rapidly regained.

A second technique is using the standard's reversible variable-length code. This code can be uniquely decoded even when read backwards, which means that the terminal can still use all uncorrupted information from the newly found resync marker back to the place of the error burst.

Low bit-rates are accommodated as well by the use of scalable objects in MPEG-4. Nowadays, many Internet audiovisual providers ask the viewer to choose the streaming quality, depending on the bit-rate he or she can handle. So providers need to encode their content separately for each possible bit-rate. With scalable coding, though, providers need encode clips only once, with the bit-rate sensed automatically before playout, or even adjusted during it.

In scaling, a base layer conveys all the information in some basic quality, and one or more enhancement layers can be used to get a better picture if the bit-rate is available. When a scene is composed of different objects, it is even possible to send only the most important of them. This kind of scalability is completely new. Note that it also allows unequal error protection—protecting the most important objects best (remember that error protection also costs bits).

Sprites, as they are called, are another device useful for saving bits, and are typically used to code unchanging backgrounds. Currently, when, say, camera angles are changed by the viewer in an interactive application, the background must be completely updated for each change. But a sprite defining the image of the background need be sent only once. After that, new views are created by simply sending the new positions of four pre-defined points.

## Hold that smile

Of all the new and useful features in MPEG-4, perhaps the most entertaining is its ability to map images onto computer-generated shapes. When the shapes are animated, the gap between synthetic and real can be bridged quite efficiently. In principle, any mesh (currently 2-D, with 3-D in the next version of the standard) may have any image mapped onto it. A few parameters to deform the mesh can create the impression of moving video from a still picture—a waving flag, for instance. For more advanced effects, moving video images could also be mapped onto the mesh. Consider the bit-rate saving: rather than sending entirely new images for each deformation, just the directions to do so are sent, and the warping is done locally.

## MPEG-7 and the missing numbers

Confronting a ton of data without the ability to search it intelligently is almost useless, as many users of the current text-based Internet search engines know. However, with the widespread adoption of digital coding, more audiovisual data floods the net every day. Hence the urgency of MPEG-7, scheduled for approval in September 2001.

What MPEG-7 will do is standardize the description of multimedia material: pictures, sound, and moving images. Those descriptions are often referred to as metadata.

MPEG-7 will be useful to describe multimedia data regardless whether in local storage, remote databases, or broadcast. Examples are finding a shot in a movie, finding music that sounds like a favorite compact disk, or picking a digital broadcast channel from the hundreds available.

The use of MPEG-7 descriptions will expand the familiar static methods. The searcher for pictures could input a sketch as a wildcard image or else a general verbal description ("a sunset with a halo over a mountain-top" or "two dandies dueling"). Music could be found in a "query by humming" format.

A current issue in defining the standard is choosing useful descriptors at all levels in terms of their content, the previous examples being high-level, and color, texture and audio spectrum being low-level. MPEG-7 will also specify hierarchical schemas using multiple descriptors of this kind, as well as a language for defining these schemas.

A historical note. Many people have had their curiosity piqued by the strange numbering system of MPEG standards: 1,2,4, and, on the horizon, 7. The explanation is quite mundane. Several years ago, MPEG began work on an MPEG-3 standard for high-definition television (HDTV), but in 1993 it became clear that the tools needed were very similar to those in MPEG-2. Therefore, MPEG-3 was quickly abandoned, and HDTV support was included in MPEG-2.

Early on in MPEG's work, the numbers 5 and 6 were not assigned to work items, an official term for current and future projects. After starting such a new work item, the questions had logic of a sort: "shall the number for the next job be 5, which follows 4, or should it be 8, attractive in its own binary way, to follow 1,2 and 4?"

After some thought, MPEG members decided that their new work item was so different from what had gone before that they threw both ideas overboard and chose 7 as the lucky number (poor 6 never had a chance).

—*R.K.*

Pre-defined faces are particularly interesting meshes. These are computer models with a repertory of independent motions and a few common emotional states.

With MPEG-4's text-to-speech interface (discussed in more detail in a moment) the animated face is an attractive tool for building an avatar—a digital, on-line stand-in for a human or synthesized presence [see Special Report, "Sharing virtual worlds," *IEEE Spectrum*, March 1997, pp. 18–51].

The appearance of the face may be left to the decoder, or complete, custom facial models may be downloaded [Fig. 3]. The wireframe face model may have any surface, including even a snapshot of a person, though as the face is inherently 3-D, a special snapshot is required.

Any feature on the model, such as lips, eyes, and so on, may be animated by special commands that make them move in sync with speech. Parameters for defining and animating entire bodies will be published in MPEG-4 Version 2.

## Is it live or is it encoded?

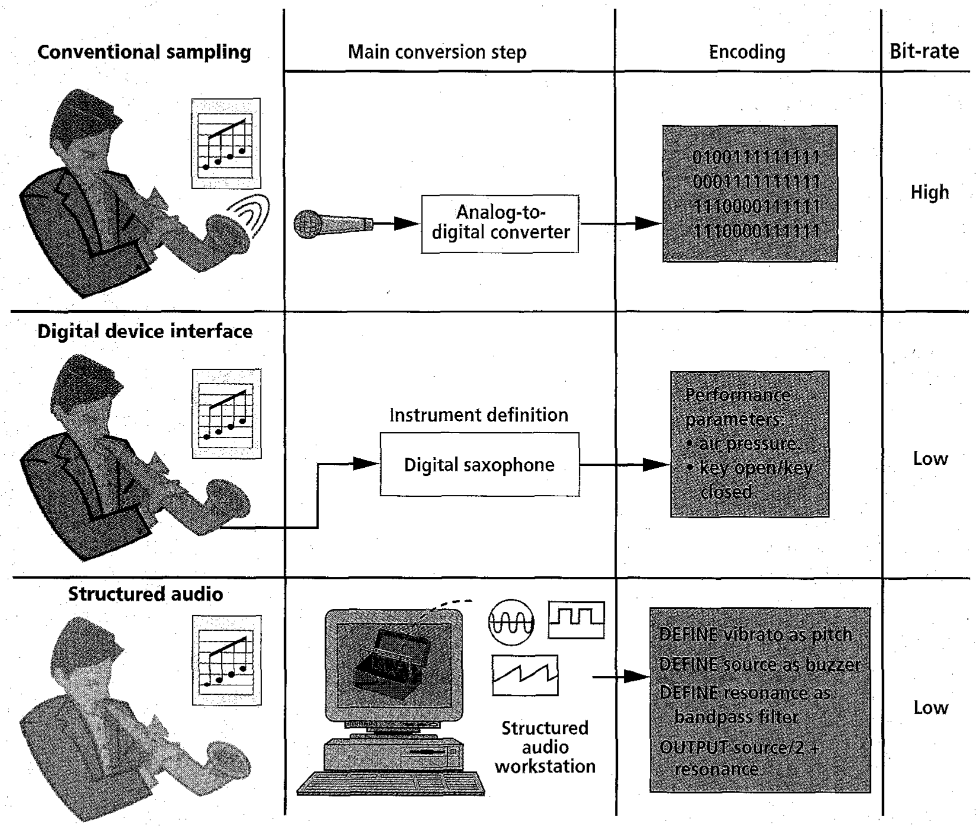
Some 10 years ago, when the Moving Picture Experts Group started its work, audio took a back seat, regarded simply as a signal associated with video. Now the audio is of the same importance—and quality—as the video.

MPEG-4 includes several audio tools for achieving a good complexity-performance ratio at bit-rates starting at 6 kb/s and reaching beyond 128 kb/s. The range embraces everything from a mono signal to transparent-quality stereo, that is, stereo with no audible degradation. The quality is not only better than for a CD, but uses less than a tenth of the latter's encoding rate of 1411 kb/s.

For audio of the highest quality, MPEG-4 includes the advanced audio coding (AAC) algorithm of the MPEG-2 standard. The algorithm gives what is termed indistinguishable-quality audio, from mono to multichannel surround sound, at considerably lower bit-rates than the mp3 audio format widely used on the Internet today. ("Indistinguishable" is a term used for a perceived quality above the broadcast one, as judged in double-blind subjective tests.)

Speech is coded by two algorithms. One parametric coder handles operation at 2–4 kb/s, or even lower in variable bit-rate mode. The other, based on a technique known as CELP (code excited linear prediction), is for operation at 4–24 kb/s. The latter uses sampling at 8 or 16 kHz for narrowband and wideband speech respectively.

Piling one trick on another, MPEG-4 can even utilize written text to generate corresponding motions of the synthetic face models mentioned above. Moreover, the text, which may be synchronized with those motions, is not spoken in a flat droning voice; rather, the speech generator (the makeup of which is left to the



```
DEFINE vibrato as pitch
DEFINE source as buzzer
DEFINE resonance as
  bandpass filter
OUTPUT source/2 +
  resonance
```

Source: Eric Scheirer, Massachusetts Institute of Technology

**[4] Structured audio (SA) is a new way of creating music that is standardized in MPEG-4. It joins conventional methods of digitally encoding music: sampling the sound stream [top] or recording the physical input to an instrument controller [middle]. (In point of fact, SA also includes the latter type of synthesis, which is similar to methods employed with the music instrument digital interface [MIDI].)**

**In a more powerful form of SA [bottom], sound synthesis primitives, such as oscillators and filters, are assembled into networks to create specific sounds. The assembly Instructions and commands to drive the network—not the network output itself—are sent to the terminal.**

developer) can be supplied with so-called prosodic parameters, that is, information about stress and changes in speed. For even greater individuality of speakers, parameters like age, gender, and accent *may also be specified.*

## Build your own instrument

MPEG-4 has entered new terrain with its provisions for "structured audio," a fertile method for creating sounds at extremely low bit-rates. The fundamental ideas come from work at the Massachusetts Institute of Technology's Media Laboratory on NetSound, which uses the popular synthesis language Csound. Rather than a single method of synthesis, structured audio is a format for *describing* methods of synthesis. MPEG-4's standard for it can thus accommodate any current or future method [Fig. 4].

In structured audio, descriptors for many signal-processing elements for sound synthesis, such as oscillators and digital filters, are specified, and small networks of elements are chosen to create the specific sounds. Each network, whether its ultimate sound production is the wail of a trumpet or a fire alarm, is termed an instrument. These instruments can be downloaded and then played through commands in the bit-stream.

The method is in essence a musical-score–driven synthesis. The "scoring" of the sound space uses two languages: the structured audio orchestra language (SAOL, pronounced "sail") and the structured audio score language (SASL). Instruments can be defined and downloaded with the first and controlled with the second. Skilled structured-audio programmers-cum-composers can create any sound, from real-sounding pianos to rushing water.

An important result of the description framework is that the synthetic audio fed to each terminal is identical. Thus, barring the vagaries of physical equipment that one user might have compared to another, the output is guaranteed to sound the same from terminal to terminal.

Some other audio options are already familiar to users of mass-market synthesizers. The popular musical instrument digital interface (MIDI), whose repertoire of instruments is quite limited, can be used to control the MPEG-4 audio if very fine control is not required. Wavetable bank synthesis (a technique found in many PC sound cards) can be used in simple environments.

## Spaced-out sound

Although less self-evident than with images, audio is also represented in the form of objects. An audio object can be a monaural speech channel or a multichannel, high-quality sound object. The composition process is in fact far more strictly prescribed for audio than for video. With the audio available as objects in the scene graph, different mixes from input channels (objects) to output channels (speakers) can be defined for different listening situations.

Another advantage of having audio as objects is that they then can have effects selectively applied to them. For example, if a soundtrack includes one object for speech and one for background audio, an artificial reverberation can be applied to the speech as distinct from the background music. If a user moves a video object in the scene, the audio can move along with it, and the user could also change how audio objects are mixed and combined.

Like video objects, audio objects may be given a location in a 3-D sound space, by instructing the terminal to spatially position sounds at certain spots. This is useful in an audio conference with many people, or in interactive applications where images as well as audio are manipulated.

A related feature known as environmental spatialization will be included in MPEG-4 Version 2. This feature can make how a sound object is heard depend on the room definition sent to the decoder, while the sound object itself need not be touched. In other words, the spatializations work locally, at the terminal, so again virtually no bit-transmission overhead is incurred.

Imagine spatialization when a person walks through a virtual house: when a new room of different shape and size is entered, the sound of the voice object changes accordingly without the object itself having to be changed.

## The future of MPEG-4

So where will the MPEG-4 standard be deployed? The first systems are likely to be used for Internet delivery of multimedia. One MPEG-4 server and a software decoder was shown by Philips Digital Video Systems at the September 1998 International Broadcasting Convention, in Amsterdam. And the video decoder in Microsoft's Windows Media Player now operates according to the text of the final MPEG-4 standard.

Manufacturers of mobile equipment and providers of mobile services will probably be the next adopters. Implementing the entire standard would be overkill for a mobile device. Instead, a number of profiles, which define subsets of the standard, have been included for their designers to choose from. Profiles are available for simple situations for mobile use or complex ones with advanced 3-D graphics. (Similarly, MPEG-2 profiles exist for digital TV.)

Some audio broadcasters have already expressed interest in MPEG-4, whose sound quality has been judged by the European Narrowband Digital Audio Broadcasting Group to be better than analog AM broadcasting for the same bandwidth. But so far most others in the broadcast world have ignored MPEG-4.

Some reactions have even been hostile, in the belief that the standard is intended to supersede MPEG-2. Obviously, someone in the costly process of switching over from analog television to (MPEG-2) digital operation will find yet another MPEG standard not particularly attractive.

But MPEG-4 was not meant to replace MPEG-2; instead, it enables new applications and new types of content, as well as more types of connections. In fact, the Moving Picture Experts Group is now designing ways for MPEG-4 and MPEG-2 to work together. As broadcast becomes digital worldwide and TV sets come to resemble interactive terminals a new phase in the evolution of multimedia has begun. ◆

## To probe further

Detailed information about all of the MPEG standards, including MPEG-4 and MPEG-7, can be obtained from the MPEG homepage on the World Wide Web, www.cselt.it/mpeg. Some other MPEG-related sites and their special interests are:

- garuda.imag.fr/MPEG4/ (systems).
- www.hhi.de/mpeg-video (natural video coding).
- www.es.com/mpeg4-snhc/ (synthetic visual coding).
- www.tnt.uni-hannover.de/project/mpeg/audio/ (audio coding).
- sound.media.mit.edu/~eds/mpeg4/ (structured audio).

The May and June 1998 issues of the *IEEE Proceedings* include a number of articles that supply background information on the new MPEG-4 standard.

Accessible, in-depth information on MPEG-4 and MPEG-7 is available in *Advances in Multimedia: Standards, Systems, and Networks*, edited by Atul Puri and Tsuhan Chen (Marcel Dekker, New York, 1999).

A visit to the Web3d Consortium (formerly VRML Consortium), www.web3d.org, will supply more information on the Virtual Reality Modeling Language.

## About the author

Rob Koenen is a senior consultant and project manager with KPN Research, the research facility of KPN, the Dutch telecommunications operator based in Leidschendam. He has been actively involved in shaping the MPEG-4 and MPEG-7 standards, and has chaired the requirements group within the Moving Picture Experts Group since 1996.

*Spectrum* editor: Robert Braham