Digital Formats for Images, Audio and Video

Peter Wittenburg, Reiner Dirksmeyer, Hennie Brugman, Gerd Klaas (MPI for Psycholinguistics)

Digital audio and video formats became very important when the transition was made from traditional magnetic tapes as the main recording and preservation technology to computer-based digital methods. Increasingly often linguists and other collaborating researchers are confronted with acronyms such as MPEGx, MD, AVI etc but don't exactly understand where these are about and what implications decisions may have. This paper wants to give a brief overview about the most relevant encoding formats (codecs), file formats and application programming interfaces (APIs). This may help interested users to find their way in the jungle of acronyms.

We have to distinguish three levels: (1) The stream of data - be it the acoustic wave - has to be encoded digitally such that it can be represented and processed as sequences of "0" and "1" on our digital machinery. We call this the encoding of media streams. At the playing side consequently their has to be a decoding process that creates waves or moving images again. Therefore the term "codec" (coder/decoder) is often used to refer to this level. (2) The stream of data has to be packaged into file formats that can be handled by the operating system and application programs. Application programs need to know how to interpret the chunks of information contained in a file and how to retrieve typical metadata such as sample frequency. This information is normally contained in headers (first x bytes in a file) in a specific sequence and format. (3) Some acronyms used in the media world refer to more complex objects (media and linked annotations). To write and read such complex objects the designers as in the case of Quicktime MOV files offer APIs (Application Programmer Interfaces), i.e. to access such files one has to use the provided programs or develop a new one making use of that API.

Audio Encoding Standards

Due to technological developments a number of different standards emerged for the encoding of acustic waves. Linear PCM[1]2 (Pulse Code Modulation) is the most simple and direct way

[1] For certain telephone and other applications codecs were developed that use non-linear methods in both dimensions: voltage and time. These (ADPCM, A-law, μ-law, ...) will not be commented here.

to digitize a waveform. The given voltage range is divided in equidistant steps. For 16 bit processing this means that waveforms can be represented in more than 64.000 equidistant values. In general this is said to be sufficient for speech signals. For signals with a large dynamic range – where one has very silent and very loud parts as for example in Beethoven symphonies but still wants to preserve the details – some archivies recommend to use 24 bits. We assume that for the work within DOBES 16 bit is sufficient. Each value is represented by 2 bytes. In general the samples of the waveform are taken at equidistant time steps defined by the sampling frequency. According to Nyquist only those frequency components are represented for which at least two points are taken for one period. Speech – in particular from children and female voices – are said to have frequency components to up to 7 kHz. Therefore, for speech recordings 16 kHz is the minimal sample frequency to cover the details. For Hifi recordings signal components of up to 20 kHz are seen as relevant[3], therefore the sample frequencies for most state-of-the-art equipment is set to 44.1 or 48 kHz. Some argue that over-sampling should be done and therefore require 96 kHz for archive recordings. We assume that 44.1 or 48 kHz is sufficient for all work in DOBES.

Both the ATRAC (used in MiniDisc Recorders) and the MP3 (used in stand-alone recorders and in the MPEG video codecs) compression algorithms apply psychoacoustic filters that filter out signal components that our ears cannot perceive as the documents claim. Frequency and time masking is performed, i.e. information is deleted and cannot be recovered. MD use a fixed setup and their algorithm is proprietary (not open) while for MP3 the bit-rate can be chosen and its algorithm is documented openly.

In MP3 the bit-rate determines how good the original waveform is approximated. For speech 192 kbps is in general sufficient to achieve acceptable listening quality. The MD recorders re-synthesize the signal to make it externally available, i.e. computer captioning is done via an analogue outlet which also means a 3 dB signal quality decrease. For MP3 algorithms are available that turn the compressed representations into linear PCM representations. Analysis results (Campbell, van Son, Wittenburg) have shown that most of the usual linguistic analysis operations such as pitch extraction and spectral analysis can be done without getting large deviations compared to analyzing the original waveform. Nevertheless, due to the loss of information it is strongly recommended for archiving purposes to use the best recording quality possible, since we don't know what future generations may want to do with the material. It has to be mentioned that in many cases the transformation of compressed formats to other compressed formats will not be without introducing severe artifacts.

Audio File Formats

There were a couple of file formats used to capture audio information such as NIST, AIFF etc. The de facto standard today is the WAVE format (.wav). It basically specifies how chunks of data can be read, in particular the format chunk informing about parameters such as the sample frequency etc and the data chunk containing the data in a specific byte order. WAVE is a particular sub-format of what was called RIFF (Resource Interchange File Format) which was created for different sort of applications. In fact all major programs support the WAVE format and there are converters to other formats such as AIFF etc. In practically all cases WAVE formatted files contain linear PCM data (differing resolution and sample frequency). Therefore, it is used as a synonym for linear PCM which is not fully correct.

Image Formats

For still images the difference between encoding and file formats is not obvious for the most cases, i.e. an encoding standard often also covers a file format. Therefore, we will not make this distinction here.

For still images also a number of encoding schemes are well-known. TIFF (Tagged Image File Format) is not standardized, it is more of a framework where different sub-communities have created their TIFF standard. Each manufacturer of high-resolution scanners produces his own TIFF version. Although it allows storing of compressed images as well, it is in general used for encodings that are comparable to PCM for waveforms. A picture is optically mapped to a lateral sensor that has a certain spatial resolution (number of image points in x and y dimensions yield a matrix of pixels) and for every pixel color and brightness are represented by a number of bits[4]. Therefore TIFF in general stands for uncompressed representations of image information. The major programs can handle a variety of TIFF formats although for specific versions (for example LANDSAT images) special viewers and converters are necessary.

Most popular is the JPEG format *(Joint Photographic Expert Group)*. It stands for a certain way to compress image information and for a file format. It makes use of discrete cosine transformations and the compression is achieved by cutting off high frequency components. Therefore sharp edges (lines) are smeared out. The compression factor can be chosen and of course the compression is lossy. Since all still cameras and most programs support JPEG this format is the de facto standard. Therefore, archives have to accept it. Since the format is openly described conversions to other formats are easily possible.
There are a few other but less important formats such as GIF (Graphics Interchange Format) and PNG (Portable Network

[2] "DAT Recording" is often used as synonym for good quality linear PCM recording. This is completely misleading, since the term "DAT" refers to a tape format (Digital Audio Tape). Correct is that for example the Sony DAT recorders used electronic circuits that generated high quality linear PCM (16 bit, 44.1/48 kHz). New types of recorders such as Flash-Card Recorders also support this type of high quality recording settings.

[3] Here the perception capacity of the human ear is used as indication instead of using human production characteristics.

[4] There are different color and brightness encoding schemes such as Grayscale, Pseudocolor, RGB, YcbCr and CMYK. This paper cannot go into the details.

Graphics). GIF was very popular at the start of the Web, since it is a highly reduced format. It just has 256 color encoding levels, does lossless compression and can be transferred very quickly via Internet. Since the owning company wanted to get money for every web-graphic, since the representation of color is so limited and since the network speeds increased it lost its importance. PNG supports lossless compression, supports a large color depth as JPEG and has a number of other excellent features making it very attractive. However, it is not very well supported by hardware and software builders.

JPEG2000 is a new standard that is intended to overcome some of the drawbacks of JPEG. Its compression is based on modern wavelet technology and therefore more optimal than for JPEG. It is fully specified in the mean time and aside to the core definition it covers various extensions such as for motions, file formats, APIs etc. Yet there is not much software that supports JPEG2000.

We should mention the SVG graphics format which is used to represent scalable vector graphics and is supported by W3C as a web-standard. ELAN makes use of this format to identify graphical shapes.

Video Encoding Standards

Uncompressed video which would be comparable to linear PCM in the audio world would amount to more then 250 Mbps, i.e. 30 minutes of video would require about 100 GB of storage. These two numbers indicate that uncompressed video still means a too heavy load on our current computer machinery.

Most relevant are the codecs worked out by the MPEG group (Moving Pictures Expert Group). They all do compression in the spatial and time domain and allow to define the amount of signal reduction. They also combine image and audio waveform encoding and define the packaging of the information stream such that decoders know how to unpack the stream and re-synthesize perceivable information. MPEG1 was the first compression algorithm of this type and allows bit rates between 1 and 3 Mbps (mega-bits-per-second). In the spatial domain the discrete cosine transformation is applied to compress the signal. In the time domain group of pictures are defined that include a keyframe (full pixel representation) and a number of frames (prediction- and bi-directional-frames) that are highly compressed over time. They encode difference values between frames in a tricky way. PAL video[5] is delivered with two interlaced fields both covering half of the image and fields are sampled at a time resolution of 20 msec. MPEG1 only digitizes one of the frames, therefore it offers only a pixel resolution of 352*288 (also called SIF) and a time resolution of 40 msec. The resulting bit stream can be handled by CDROM drives which is the reason why MPEG1 was used for CDROM technology.

However, due to MPEG1's heavy reduction, its relatively poor quality and its tricky encoding scheme over time MPEG2 was defined. It is based on similar compression principles, but encodes every field. Therefore, the original spatial resolution for PAL video is 704*576[6] and for time 20 msec[7]. It is widely used by the media industry for editing machines and was accepted as a kind of backend format[8]. Therefore, it is interesting for archiving purposes. MP3 is used for audio encoding in MPEG2.

MPEG4 is the last invention of the MPEG group[9] and is in so far very different from the first two ones, that it is more of a framework for decoding and merging several different streams of media information and supporting user interaction. It comes with an improved compression algorithm for video encoding and is designed for web-based applications. Although there already codecs around there is not so much software yet supporting MPEG4. It is mainly used for web-streaming[10] purposes. It allows to set bit-rates from 500 kbps on.

Many programs support the MPEG codec lines. However, writing software for proper decoding still seems to be a difficult issue, since it still occurs that programs or new program versions don't handle MPEG streams accurately. Video encoding requires to set a large number of parameters. The MPI team provides a template with default parameters that can be used with Adobe Premiere for example. There is a trend in TV and media industry to use MPEG2/4 with only I-frames, i.e. no intermediate frames with compression over time.

There are a few other codecs that are supported by some programs. Cinepack was used in old Quicktime versions, however, cannot be recommended anymore due to its bad quality. Sorensen is another codec, but not used very frequently. There are also proprietary codecs as provided by RealVideo and Microsoft. DV (Digital Video) is another very popular digital format. It was created from Sony for their digital cameras, but it is proprietary. These codecs don't play any role for archiving purposes such as in DOBES. The following table indicates again the data rates for the most important codecs (typical values, all for PAL).

	uncomp video	DV	MPEG1	MPEG2	MPEG4
bit rate in Mbit/sec	> 250 Mbps	35	1.5	6	0.5 - ...
1 hour recording in GB	>100	16	0.7	3	0.2 - ...

Video File Formats

The encoded video stream has to be packed into a file format. Here we only can briefly mention some of them without describing

[5] NTSC Video in MPEG1: 352*240

[6] NTSC Video in MPEG2: 704*480

[7] There are various other resolutions supported which will not be reported here (HDI, QSIF, ...).

[8] MPEG2 4:2:2P is the standard that was agreed by the big media industry for their "mainstream production lines". The numbers 4:2:2 point to the relation for encoding brightness and color information (Y,U,V model). MPEG2 is the standard for DVD technology.

[9] MPEG7 and MPEG21 are not about codecs and file formats.

[10] In contrast to the download option where data is first downloaded and stored on the local machine, streaming means that data is directly presented on the screen while being sent.

the details. MPEG (.mpg) comes along with its own file format. Another very popular format is AVI (.avi). While .mpg format only includes streams encoded with the MPEG codec, AVI can include all sorts of codecs, i.e. saying that something is an AVI file does not say anything about quality and other characteristics. DV streams for example can be embedded in AVI.

APIs for Complex Streams

There is a group of formats that go beyond a simple file format such as Quicktime and SMIL. QT and SMIL (which is a recognized standard by W3C) both can contain several tracks of different information types that are linked by cross-references. This would for example allow to store video and its synchronized subtitles. An API (Application Programmer Interface) is mostly provided that tells the user how the different objects can be dealt with. Media can be encoded in different ways, i.e. saying that a video is in QT format does not say which codec is used. QT is very popular since it was pushed by Apple and since there are QT Players for Windows and MAC OS. QT is supporting a number of codecs. SMIL is very similar, an open standard and will be supported by the major web-browsers. MPI will probably offer subtitled video presentations via the web by providing SMIL documents.

MXF (Material Exchange Format) will become a very important format for media exchange, since the big TV and media companies agreed on it. It includes data and metadata and can codify different tracks of related streams. It will play an important role for future archiving, i.e. we expect that the DOBES archive at a certain moment may have to turn over to MXF.

Summary

This brief overview may have given a better insight to what currently is used and recommended. It is impossible to write a comprehensive document about these issues that will not cross the 100 pages boundary. Therefore, it is always recommended to speak with experts, since much knowledge especially about tools resides in their heads only. We can't speak about a stable situation in the area of codecs and formats, since within a 5 years time new suggestions for codecs and formats will be made. Whether such suggestions or those that are already around will become standards relevant for the documentation work and for archivists cannot be predicted. There is a clear trend to uncompressed representations, but technology has to support this.