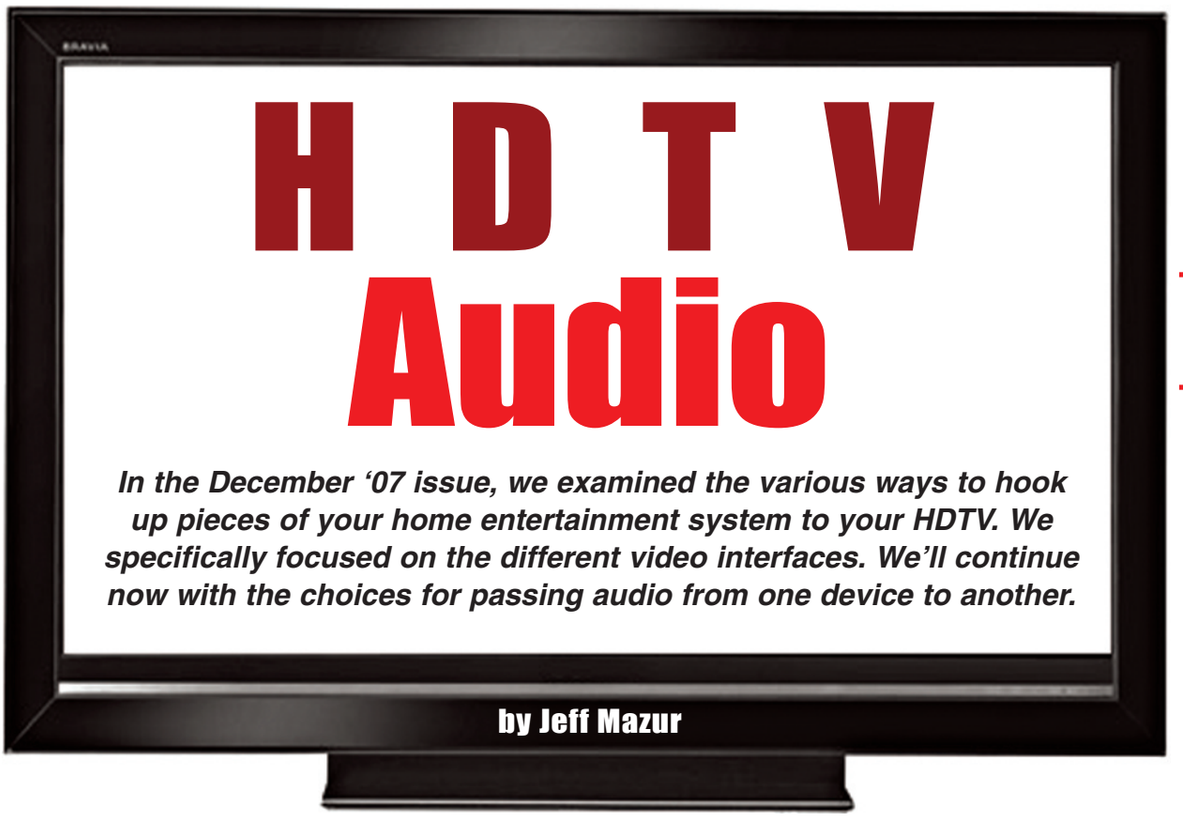


HDTV Audio

In the December '07 issue, we examined the various ways to hook up pieces of your home entertainment system to your HDTV. We specifically focused on the different video interfaces. We'll continue now with the choices for passing audio from one device to another.

by Jeff Mazur

Once again, the most common connection by far is the standard analog stereo pair using RCA jacks and cables. With good quality cable and connectors, this method can provide excellent results. The most common issue with analog audio connections is its susceptibility to picking up hum and/or other extraneous signals, especially from components within your system (or perhaps from the ham operator who lives next door!). To solve this issue — as well as complete the total conversion to binary 1s and 0s — there are three basic ways to pass audio signals digitally between devices: coax, optical, and HDMI.

Figures 2-4 are courtesy of Wikipedia, the free encyclopedia (licensed to the public under the GNU Free Documentation License).

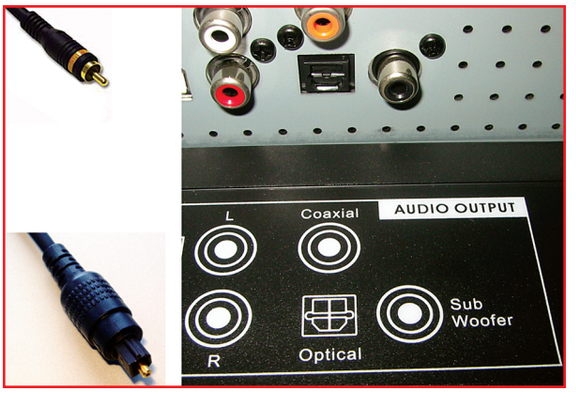


FIGURE 1. Digital audio connections (top, coax and bottom, optical).

S/PDIF (Sony/Philips Digital Interconnect Format)

Named after the two companies that developed this interface, S/PDIF is a means to carry audio between devices in a digital format. The signals can be carried over standard 75 ohm coaxial cable using RCA jacks (or BNC connectors in professional equipment) or via optical fiber (glass or plastic, usually terminated with F05 connectors). See Figure 1.

The optical connection — created by Toshiba and also known as TOSLINK — uses 1 mm fiber terminated in a 5 mm connector. While earlier cables were restricted to less than 15 feet, you can now buy high quality TOSLINK cables up to 100 feet in length. TOSLINK can carry data signals of up to 125 Mbits/s, which allows for three audio channels. However, it is usually used to carry a single pair of stereo audio signals.

As an electrical signal, S/PDIF is represented by a roughly 1V digital pulse train using Biphase Mark Code (BMC) to carry the audio data. While no specific sampling rate or bit depth is specified in the standard, audio is usually carried as either 48 kHz (DAT) or 44.1 kHz (CD) data with either 20 or 24 bit samples. We'll describe the actual data format in a moment.

HDMI

We've already discussed the HDMI interface that can carry digital video between devices. HDMI also includes support for up to eight channels of uncompressed digital audio at a 192 kHz sample rate with a

24 bits/sample, as well as compressed streams such as Dolby Digital, or DTS. HDMI also supports one-bit audio, such as that used on Super Audio CDs at rates up to 11.3 MHz. With version 1.3, HDMI now also supports lossless compressed streams such as Dolby TrueHD and DTS-HD Master Audio.

Digital Audio Basic

Digital audio connections can be used to connect various components of your home entertainment system such as from a cable or satellite STB (Set Top Box) to the TV. Since audio is transmitted digitally in the ATSC DTV signal, this will often be the best choice. Other components (e.g., a CD player) also handle audio natively in a digital form. However, devices that handle audio as an analog signal — including the equipment used to record or create TV audio at its source — must first convert the analog signal to digital. This process is known as digitizing and is a good place to start when discussing digital audio.

To digitize an analog signal, we basically perform two separate functions. First, the signal is sampled at regular intervals to determine its value at each discrete point in time. This is usually the function of a sample-and-hold circuit. Next, each sample is quantized, or converted from an analog voltage to a particular digital representation of that value.

The sampling rate determines what frequencies can be carried digitally; information theory tells us that only frequencies below one-half of the sampling frequency (also referred to as the Nyquist frequency) can be represented accurately. Signals above this limit will cause extraneous frequencies (i.e., distortion) to appear due to an effect known as aliasing. In other words, we need at least two samples per cycle of the highest frequency we wish to digitize.

The quantization of each sample determines how many bits will be used to represent each sample. The more bits, the higher the precision will be of each sample. This translates into the dynamic range of a signal, or the difference between its lowest and highest values. Under ideal conditions, it also represents the maximum signal to noise ratio (SNR), which is related to the number of bits by the following formula:

$$SNR = 20 \log 2^N = \text{approx } (6 \times N) \text{ dB}$$

where N = number of bits.

For example, a 20-bit converter theoretically could obtain an SNR of 120 dB (if there are no other sources of noise). In practice, the maximum signal level is usually reduced by 20 dB of headroom to prevent clipping. This still leaves an SNR of approximately 100 dB. In comparison, normal audio tape typically only achieves an SNR of about 60 dB.

As you can see, digitizing an analog signal is all about compromise. You need to sample at a high enough rate so as not to miss changes in the signal that occur between the samples. And we need enough bits to represent each sample so that the difference between the actual analog value and its closest digital representation (a.k.a., quantization error) is not very much. Of course, increasing either of these values means that there will be more digital data that needs to be carried and processed.

On the positive side, once a signal has been digitized it can be transmitted much more efficiently and without many of the side effects of noise and distortion present in the communication channel used. More importantly, it can be compressed digitally so that redundant and/or unessential data can be discarded. This is one of the main reasons that our TV signals are undergoing the transition to digital.

PCM

There are many ways to represent each sample as a digital signal. The most common technique is known as Pulse-Code Modulation (PCM). This approach simply takes the output from an Analog-to-Digital Converter (ADC) and places the bits into a continuous bitstream.

Figure 2 shows a sine wave (in red) that is sampled and quantized using simple PCM. At each sample point, the digital representation of the signal's analog value is sampled and then held until the next sample point. This produces an approximation of the original signal, which is easily encoded as digital data. For example, if the sine wave in Figure 2 is quantized into 16 values (i.e., four bits), we would generate the following data samples: 1001, 1011, 1100, 1101, 1110, 1110, 1111, 1111, 1111, 1110, etc.

We could transmit these PCM samples as four-bit parallel data with a separate clock signal to indicate when each sample was taken. This is cumbersome, however, and requires the use of multi-conductor cables. Most data transmission today is done in a serial fashion. This requires that each bit of the PCM sample be clocked out onto a single serial data line. At the receiving end of this data stream, a shift register will convert the serial data back into parallel data words. To keep the receiver in sync with the transmitter, some form of clock recovery is necessary.

One of the easiest ways to do this is to make sure that the serial data changes polarities at least once during each bit-time. This is the basis for several different coding schemes, including Biphase Mark Code (BMC) — the signaling method used by both TOSLINK and the professional digital audio format established by, and referred to as, AES/EBU (Audio Engineering Society

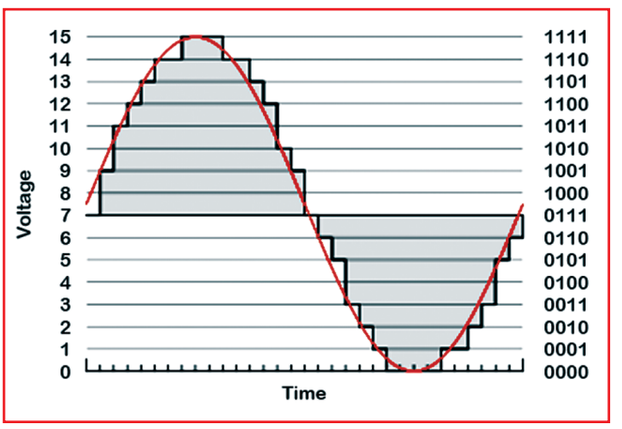


FIGURE 2. Analog-to-digital conversion of a signal using Pulse Code Modulation (PCM).

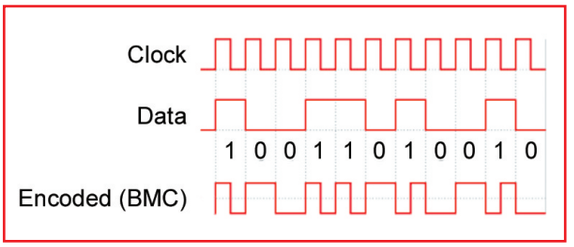


FIGURE 3. Serialization of digital data using Biphase Mark Coding (BMC).

and the European Broadcasting Union).

With BMC, the data stream changes value at the beginning of each data bit. A logic 1 is represented by having the stream change value again during the middle of its bit time; it does not change for a logic 0 (see Figure 3). BMC coding provides easy synchronization since there is at least one change in polarity for every bit. Also, the polarity of the actual signal is not important since information is conveyed by the number of transitions of the data signal.

Another advantage of BMC is that the average DC value of the data stream is zero, thus reducing the necessary transmitting power and minimizing the amount of electromagnetic noise produced by the transmission line. All these positive aspects are achieved at the expense of using a symbol rate that is double the actual data rate.

Transmission Protocol

S/PDIF and its professional cousin, AES/EBU, were designed primarily to support two channels of PCM encoded audio at 48 kHz (or possibly 44.1 kHz) with 20 bits per sample. Sixteen-bit data is handled by setting the unused bits to zero; 24-bit data can be achieved by using four auxiliary bits to expand the data samples. The low-level protocol used by both S/PDIF and AES/EBU is the same, with the exception of a single Channel Status bit.

To create a digital stream, we break the continuous audio data into smaller packets or blocks. Each block is further divided into 192 frames. Note, however, that these frames have nothing to do with frames of video. In fact, when digital audio is combined with digital video signals, there are a number of steps that must be taken to make them compatible. First off, both digitizing clocks must be synchronized to a common 27 MHz timebase. Even so, a frame of NTSC video has a duration of:

1 / 29.97 = 33.366... ms

At 48 kHz, an audio frame has a duration of:

1 / 48,000 = 20.833... µs

This makes a complete audio block 192 x 20.833 = 3,999.4 µs. The number of audio samples per video frame, however, is not an integer number:

33366 / 20.833 = 1601.6 audio samples/video frame

Because of this, it takes a total of five video frames before an even number of audio samples corresponds to an even number of video frames (8,008 audio samples per five video frames). Some video frames are given 1,602 samples while others are only given 1,601. This relationship is detailed in Figure 4.

Each audio frame consists of two subframes: one for each of the two discrete audio channels. Furthermore, as shown in Figure 4, each subframe contains 32 bits — 20 audio sample bits plus 12 extra bits of metadata.

There is a single Channel Status bit in each subframe, making 192 bits per channel in every audio block. This means that there are 192 / 8 = 24 bytes available in each block for higher level metadata. In S/PDIF, the first six bits are organized into a control code. The meaning of these bits is:

bit	if 0	if 1
0	Consumer	Professional
1	Normal	Compressed data
2	Copy Prohibit	Copy Permitted
3	Two Channels	Four Channels
4	—	—
5	No pre-emphasis	Pre-emphasis

In AES/EBU, the 24 bytes are used as follows:

- Byte 0: Basic control data — sample rate, compression, emphasis modes.

- Byte 1: Indicates if the audio stream is stereo, mono, or some other combination.

- Byte 2: Audio word length.

- Byte 3: Used only for multichannel applications.

- Byte 4: Suitability of the signal as a sampling rate reference.

- Byte 5: Reserved.

- Bytes 6–9 and 10–13: Two slots of four bytes each for transmitting ASCII characters.

- Bytes 14–17: Four-byte/32-bit sample address, incrementing every frame.

- Bytes 18–21: As above, but in time-of-day format (numbered from midnight).

- Byte 22: Contains information about the reliability of the audio block.

- Byte 23: CRC (Cyclic Redundancy Check) for error detection. The absence of this byte implies interruption of the data stream before the end of the audio block, which is therefore ignored.

AC-3

As previously mentioned, raw PCM data would require a large bandwidth to transmit. For surround sound, this would require approximately six channels x 48 samples/s x 20 bits = 5.7 Mb/s. With appropriate compression, however, this can be reduced to 384 Kb/s.

Dolby Digital — officially known as AC-3 (Adaptive Transform Coder 3) — is the compression scheme used to transmit audio within the ATSC DTV data stream. It can represent up to five full bandwidth (20 Hz-20 kHz) channels of surround sound (Right Front, Center, Left Front, Right Rear, and Left Rear), along with one low frequency channel (20 Hz–120 Hz) for subwoofer driven effects. This is often referred to as 5.1 surround sound.

A complete description of the

Figure 4

Bits 0 to 3

These do not actually carry any data but they facilitate clock recovery and subframe identification. They are not BMC encoded so they are unique in the data stream and they are easier to recognize, but they don't represent real bits. Their structure minimizes the DC component on the transmission line. Three preambles are possible:

X (or M): 11100010 if previous state was "0;" 00011101 if it was "1."

Y (or W): 11100100 if previous state was "0;" 00011011 if it was "1."

Z (or B): 11101000 if previous state was "0;" 00010111 if it was "1."

They are called X, Y, Z from the AES standard; M, W, B from the IEC 958 (an AES extension). The eight-bit preambles are transmitted in the same time allocated to four (BMC encoded) bits at the start of each sub-frame.

Bits 4 to 7

These bits can carry auxiliary information such as a low-quality auxiliary audio channel for producer talkback or studio-to-studio communication. Alternately, they can be used to enlarge the audio word length to 24 bits, although the devices at either end of the link must be able to use this non-standard format.

Bits 8 to 27

These bits carry the 20 bits of audio information starting with LSB and ending with MSB. If the source provides fewer than 20 bits, the unused LSBs will be set to a logical "0" (for example, for the 16-bit audio read from CDs, bits 8-11 are set to 0).

Bits 28 to 31

These bits carry associated status bits as follows:

- V (28) Validity bit: It is set to zero if the audio sample word data are correct and suitable for D/A conversion. Otherwise, the receiving equipment is instructed to mute its output during the presence of defective samples. It is used by players when they have problems reading a sample.

- U (29) User bit: Any kind of data such as running time, song, track number, etc. One bit per audio channel per frame form a serial data stream.

- C (30) Channel status bit: Its structure depends on whether AES/EBU or S/PDIF is used (see text).

- P (31) Parity bit: For error detection. A parity bit is provided to permit the detection of an odd number of errors resulting from malfunctions in the interface. If set, it indicates an even parity.

AC-3 standard and its use in ATSC transmission is quite complex and beyond the scope of this article. You can download the entire ATSC audio standards document (A/52B) using the link given under Further Info. However, there are however some interesting details worth mentioning here.

ATSC Audio Details

Unlike analog NTSC, audio does not take a backseat to video in ATSC. Quite a bit of the standard is devoted to how sound will be delivered to the viewer. We've already seen how 5.1 surround sound can be transmitted with each DTV channel. Other parameters in the audio metadata can be used to enhance the viewing experience. One of these parameters is known as dialnorm.

The purpose of dialnorm is to equalize the sound levels when changing from one program to another. The value of this parameter — which is embedded within the audio stream — is meant to indicate the level of average spoken dialog within the complete audio program. This is then used to control the decoder compres-

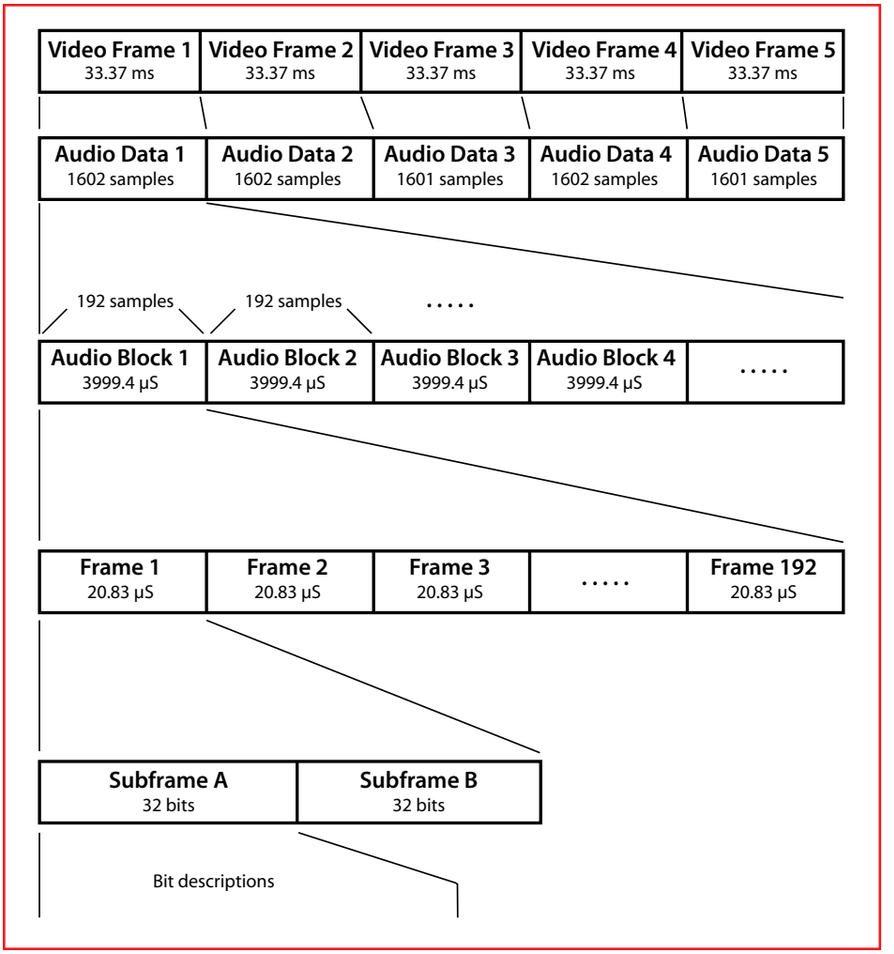


FIGURE 4. Packetization of data in digital audio streams.

sion gain within the HDTV receiver. If set properly, it will maintain a consistent dialog level between program elements and when changing from one channel to another, hence the abbreviation of "dialog normalization."

The dialnorm parameter ranges in integer values from 31 (where decoder gain remains at unity) to a value of one (where decoder gain is reduced by 30 dB). Unfortunately, many producers and broadcasters currently do not provide a proper dialnorm value in their programs. This is partly due to the complexity and variability of actually measuring the dialog level properly. Thus, you may still find wildly varying levels between channels.

Other Audio Services

The ATSC standard also provides for alternate audio channels by allowing multiple AC-3 elementary streams within the full transport stream. As such, each alternate audio channel can have up to 5.1 channels of its own to provide a complete audio service. It is also possible for the alternate audio to consist of a single channel intended to be combined with other channels from a different stream (although not all HDTVs are capable of this).

One obvious use for an alternate audio channel would be to convey the dialog in a different language, much like the SAP (Secondary Audio Programming) service, currently available on NTSC channels. Because there can be any number of audio streams, this would allow multiple languages to be transmitted at the same time.

The ATSC standard also identifies several types of audio signals that can be transmitted. These are specified in Table 5.7 of the A/52 document (see Table 1).

A complete main (CM) channel represents the main audio service with dialog, music, and effects. This is the normal audio program which can be monaural (one channel), stereo (two channel), or surround sound (5.1 channel) where available. A music and effects channel (ME) contains only those respective portions of the audio, without dialog. This would be useful when supplying a program in multiple languages; the single ME service would be combined with various other streams containing only a dialog (D) service for each language.

The visually impaired (VI) service is designed to allow a separate audio channel to contain a narrative description of the program content. Also known as video described, this aids a person who is blind or otherwise visually impaired to comprehend what is happening on the screen. Likewise, the hearing impaired (HI) service is provided to aid those with slight hearing loss. Unlike captioning, which can provide audio content for those who are completely deaf, the HI service is designed to provide more intelligible audio by processing (compressing) the dialog channel and emphasizing it over the music and effects.

While the dialog service contains actual program dialog from the speaking actors, an additional commentary (C) service can be added to provide further information. This is like many DVDs which offer a special audio track to provide director's or actor's comments while you watch their movie.

The emergency (E) service is a special, high priority channel which can be used to convey vital announcements similar to the Emergency Alert System (EAS). Whenever an E service signal is present, it will automatically mute and/or replace the normal audio channels with the E channel audio.

The voice over (VO) and karaoke services allow an additional channel to be added to an existing AC-3 stream without requiring the audio to be decoded (i.e., uncompressed) back to baseband PCM audio data, mixed, and then re-encoded. Local stations could use this to add their own audio tags to programming supplied by their network.

Lip Sync

Because audio and video are processed separately by various circuits which can delay the signals significantly, special attention is needed to keep these parts of a presentation in sync. When they drift apart past a certain threshold, the discrepancy becomes very noticeable and objectionable.

Technically called audio/video sync, this quality is often referred to as lip sync (not to be confused with a Milli Vanilli performance). A/V sync errors are becoming a significant problem in the digital television industry because of the use of large amounts of video signal processing in television production and broadcasting and fixed pixel, progressive television displays such as Plasma, LCD, and DLP sets.

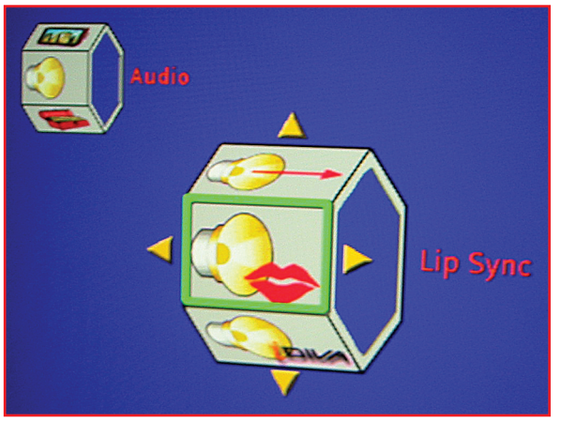
Studies have shown that "When audio precedes video by five video fields (83 ms), viewers evaluate people on television more negatively (e.g., less interesting, more unpleasant, less influential, more agitated, less successful). Viewers can accurately tell when a television segment is in perfect sync, and when it is five fields out of sync." See the Reeves and Voelker reference in the sidebar.

Furthermore, there is a larger tolerance for audio that is delayed in comparison to the video. This is a phenomenon that we are all used to when we watch a fireworks display or, to a larger degree, an electrical storm. We see the effect before we hear it. Of course, this is due to a totally different reason: the difference in velocity between light and sound waves. But if you've ever had to watch a program with significant A/V

Table 1. Bit Stream Modes

bsmod	acmod	Type of Service
000	Any	Main audio service: Complete main (CM)
001	Any	Main audio service: Music and effects (ME)
010	Any	Associated service: Visually impaired (VI)
011	Any	Associated service: Hearing impaired (HI)
100	Any	Associated service: Dialog (D)
101	Any	Associated service: Commentary (C)
110	Any	Associated service: Emergency (E)
111	001	Associated service: Voice over (VO)
111	010 - 111	Main audio service: Karaoke

FIGURE 5. Lip sync adjustment on an HDTV.



sync error, you know how annoying it can be.

Good engineering practices specify that the audio should never lead the video by more than 15 milliseconds or lag by more than 45 milliseconds. To keep the audio and video signals in sync, Presentation Time Stamps (PTS) are added to the transport stream packets. This allows the MPEG decoder in the receiver to re-assemble the packets correctly and keep the audio and video (and captions, etc.) in sync.

When the audio and video packets are multiplexed together, they can be sent up to one second apart. Fortunately, most of the other delays in the transport stream affect audio and video together. However, if you consider the delays encountered in encoding, buffering, multiplexing, transmission, demultiplexing, decoder buffering, decoding, and presentation, there can be over five seconds of delay between the broadcast input and your TV display. You can easily see this by switching between one of your local station's analog and digital channels.

Even if the receiver in an HDTV decodes a perfectly synchronized signal, there still can be a difference in the picture and sound when viewed. This is because TVs now have lots of computing power and use it to enhance HD, as well as SD pictures. They have large video buffers and DSP (Digital Signal Processing) chips to perform resolution changes (mapping the incoming video resolution to the native resolution of the display device) and correction for progressive display of interlaced sources (de-interlacing and 3:2 pull-down removal). They can also perform image enhancement to reduce specific artifacts of the display (e.g., Sony's Digital Reality Creation).

Some of these processes add considerable delay, especially when they need to examine multiple video fields to perform their function. This can cause noticeable A/V sync errors. Some HDTVs now have user adjustments to compensate for this (see Figure 5). **NV**

Further Info

Digital Audio Compression Standard (AC-3, E-AC-3) Revision B
www.atsc.org/standards/a_52b.pdf

"Effects of Audio-Video Asynchrony on Viewer's Memory, Evaluation of Content and Detection Ability"
by Reeves and Voelker
www.lipfix.com/file/doc/reeves_and_voelker_paper.pdf

Glossary of Useful Terms

ATSC Advanced Television System Committee — The organization and name of the digital television standard adopted in the US.

DTV — Digital Television

DAT — Digital Audio Tape

HDMI: High-Definition Multimedia Interface — A method of connecting components using a single cable that carries digital video signals along with multichannel digital audio.

HDTV: High Definition TeleVision — Part of the new Digital Television standards, those formats that have either 720 or 1080 lines of vertical resolution.

MPEG :Motion Picture Experts Group — Standard for transmitting compressed audio and video.

NTSC: National Television System Committee — The organization and name of the analog television standard currently used in the US.