Research White Paper

WHP 176

May 2009

Factors affecting perception of audio-video synchronisation in television

A.J.Mason and R.A.Salmon

BRITISH BROADCASTING CORPORATION

BBC Research White Paper WHP 176

Factors affecting perception of audio-video synchronisation in television

A.J.Mason and R.A.Salmon

Abstract

The increasing complexity of television broadcasting, has, over the decades, resulted in an increased variety of ways in which audio and video can be presented to the audience after experiencing different delays. This paper explores the factors that affect whether what is presented to the audience will appear to be correct. Experimental results of a study of the effect of video spatial resolution are included.

Several international organizations are working to solve technical difficulties that result in incorrect synchronisation of audio and video. A summary of their activities is included. The Audio Engineering Society Standards Committee has a project to standardize an objective measurement method, and a test signal and prototype measurement apparatus contributed to the project are described.

This document is based on a paper that was published at the 125th Audio Engineering Society Convention, New York, October 2008. Some additional figures from the presentation that was made at the convention have been included as an appendix to the paper.

Additional key words: lip sync

White Papers are distributed freely on request.

Authorisation of the Head of Broadcast/FM Research is required for publication.

Factors affecting perception of audio-video synchronisation in television

Andrew Mason[1] and Richard Salmon[2]

[1] British Broadcasting Corporation, Tadworth, Surrey, KT20 6NP, UK
andrew.mason@rd.bbc.co.uk

[2] British Broadcasting Corporation, Tadworth, Surrey, KT20 6NP, UK
richard.salmon@rd.bbc.co.uk

ABSTRACT

The increasing complexity of television broadcasting, has, over the decades, resulted in an increased variety of ways in which audio and video can presented to the audience after experiencing different delays. This paper explores the factors that affect whether what is presented to the audience will appear to be correct. Experimental results of a study of the effect of video spatial resolution are included.

Several international organizations are working to solve technical difficulties that result in incorrect synchronisation of audio and video. A summary of their activities is included. The Audio Engineering Society Standards Committee has a project to standardize an objective measurement method, and a test signal and prototype measurement apparatus contributed to the project are described.

Some additional figures, used in the presentation given when this work was published by the Audio Engineering Society, have been included in Appendix A. References to these figures have been added to the text as Figure A1, A2, and so on.

1. INTRODUCTION

This paper aims to draw attention to some of the factors in television that affect the audience's perception of whether audio and video stimuli are correctly timed relative to each other. This property of the relationship between audio and video is commonly referred to as "lip sync". This term originates from observation of television pictures showing a person talking where a mismatch of more than a few tens of milliseconds between the sounds heard and the pictures seen can ruin the illusion that is television: the motion of the talker's lips should be correctly synchronised with the sounds heard.

The concept of correct synchronisation, and some of the technical and psychological factors that can affect our perception of the correctness a televisual presentation are discussed.

Several international organisations are actively working in the field of audio video synchronisation, with the aim of reducing the accumulation of errors that plague television broadcasts. The work of one of those organisations, the Audio Engineering Society Standards Committee, is aimed a producing standard ways of measuring audio video synchronisation. Some of the preliminary proposals are described.

Let us start by dismissing an obvious untruth: correct synchronisation does not mean that the sound caused by an action should reach the audience's ears at the same time as the light from the image depicting that action. Sound travels at about 340m/s in air, light at about 300,000,000 m/s. (A convenient "rule of thumb" for those familiar with imperial units of length, is one foot per millisecond for sound, one foot per nanosecond for light). So, sound caused by an event will always, in reality, reach an observer later than light from that event. We can ignore for now those special situations in physics laboratories where photons are slowed to a few miles per hour - the programmes are not very interesting and your eyes would probably freeze.

In reality, correct synchronisation is achieved by presenting the sound later than the image. How much later depends on a number of factors.

2. ACQUISITION EQUIPMENT

2.1 Sound

Sound for television is either synthesised, or it is acquired using a microphone. This discussion focusses only on microphones.

For recording the spoken word, we can characterise 6 different microphone techniques in common use. These have different distances between the talker's mouth and the microphone. The figures in the following table are intended only as simple examples.

Mic type	Distance	Delay
lip mic	2cm	~0.1ms
tie-clip/lavalier mic	30cm	~1ms
desk/stick mic	60cm	~2ms
boom mic	1m - 2m	~3ms - 10ms
shot-gun mic	1m - 7m	~3ms - 20ms
camera-mounted mic	1m - 7m	~3ms - 20ms

So, whilst the lowest delay might correspond to a situation where the talker is not usually shown in the picture (a commentator at a noisy sports event), there is still a significant range of delay between a talker making a sound and the sound actually reaching the microphone. An error of 20ms would make a significant contribution to the audience's perception.

The electrical signal produced by the microphone might be analogue or digital. The signal represents, in a simple way, the air pressure, or pressure gradient, at the microphone. Production of digital signals implies a sampling rate and a process of sampling and quantisation. For audio, these processes have an insignificant effect on perception of synchronisation.

2.2 Vision

The electronic representation of a single microphone signal is trivial compared to that of a video signal. An audio signal is a one-dimensional signal that varies with time, video is two- (or three-) dimensional and varies with time. Video is sampled both in time and in space, and both sampling processes have an impact on perception of synchronisation.

2.2.1. Frame rate and the electronic representation of moving images

Television is based on a frame rate; generally 25 frames per second (fps) in PAL/SECAM countries and 30 (actually 29.97) in NTSC countries.

Conventionally in television each frame is divided into two interlaced fields, essentially doubling both the capture rate and the display rate to 50 and 59.94 fields per second. For the purposes of this work, the nature of interlace is unimportant, and in terms of audio-video synchronisation the rate may simply be considered as 50 or 59.94 fps.

In addition, cinema material is often generated at 24 fps, but may be speeded up to 25 fps (and then each frame presented twice to give 50 fps) or changed to 60 (59.94) fps by repeating one frame three times and then the next frame twice in a 3:2:3:2 sequence (known as 3:2 pull-down). Another scheme which gives smoother presentation of motion which may occasionally be used is 3:2:2:3 presentation. The adaptation from 60 to 59.94 may be done by a 1/1000 slowing down, or by frame-dropping. The net result is that with the film-style presentation an instantaneous event is spread in time by, at best, a double presentation of the image sequentially in time.

Conventionally, a video signal is scanned downwards from the top of a frame or field. (Figure A1) The electrical representation of the segment of the image half-way down the picture appears on the wire (in digital or analogue form) half-way through the frame period. Hence, whilst delay (to one or both of the audio signal and the video signal) may be measured in milliseconds, it is difficult to ascribe meaning to synchronisation of less than field-period.

The change to high definition television (HDTV) does not alter the situation as far as frame rate is concerned. 1080-interlace HDTV is again a 50 or 59.94 field-per-second interlace format, whereas 720p is a progressive format where each frame is presented complete at a rate of 50 or 59.94 frames per second. However, high definition has brought about a new nomenclature for film-style presentation, that is psf (progressive segmented frame) whereby each frame (at 25 or 30 fps rate) is created as a sequential frame (as in film) but is segmented with alternate lines being carried as though in fields. This is essentially the same as conventional standard definition "film" material carried on TV.

2.2.2. Camera types

A range of types of camera is used to capture television images, with new types still appearing. Each has a different timing characteristic, and as described above,

the timing is hard to define due to the 3-dimensional nature of television, where time is only one of those dimensions.

A tube camera conventionally generates a signal representing the image by means of an electron beam scanned over the tube surface. The electron beam both "reads" the accumulated charge (a measure of the image brightness) at each point, and clears that charge ready for the next capture period. The charge accumulates during the entire period from one reading to the next (i.e. one field period, the integration time). Because the refresh may not be total, there is effectively a temporal smearing or lag, the characteristics of which are device specific, and for which there may be a measure of correction in the camera channel. Tube cameras are now becoming rare in broadcast use.

The now ubiquitous CCD (charge-coupled device) camera reads the complete image during a field-blanking, storing the resulting data for sequential read-out over the following field period. Thus the whole field is captured over the same time period (the integration time), compared to the tube camera where the data on each line is integrated starting at a different time from that of the next line.

The CMOS sensor is just starting to make an impact on the television camera market. Many have a rolling shutter that gives an integration that is at different times in different parts of the image, similar to a tube camera.

The image sensor may be exposed to light for the whole field period, but in order to cut down the light level (in bright conditions), or to reduce the smearing of motion due to the long integration time, a shuttered camera may be used. In this case the aperture is only open for a part of the field period, and thus the capture period is shorter and moving objects will be sharper in an individual frame. The timing of the shutter opening may be coincident with the start of end of the field period, or part-way through it.

A shutter may be a focal plane device, in which case the shape of the shutter, and its speed of opening and closing, will have different effects on timing in different parts of the image. (Figure A2) In an electronic camera, the shuttering may be electronic: charge from the sensor may dumped part-way through the field period, and thus the image is captured for only a portion of the field period.

Another means of generating a TV signal is from film. Here again shuttering is employed, both to sharpen the capture of moving images, and to allow time for the film to be advanced in the gate to the next frame. As above, the exact form of a focal-plane shutter will have an effect on the relative timing within a frame.

Film material, whether shot specifically for TV production, or a transfer from a cinematographic production, is done by a tele-cine machine, scanning the film image. Except for the double-scanning (or 3:2 scanning) to create a TV frame rate from the film, the timing of audio relative to the scanning process should be well defined, taking its standards from the cinema industry, where the sound-track is also laid down on the film, although usually separated from the corresponding image by an offset to allow for the motion of the film both in the camera and the telecine as compared to the steady rate required at the audio-reading head.

3. PROGRAMME COMPOSITION

The definition of correct synchronisation is not that a sound corresponding to an image is presented at the same time as the image. The content of the images can imply different sound to vision delays:-

If the image depicts a person in close-up then the observer's expectation should be that the delay from picture to sound is short. This type of shot is typical for presenters in news programmes. A "head and shoulders" shot is quite common. (Figure A8)

At the other extreme, one might imagine a scene in a drama where an actor is shown in the far distance (though not so far that movements are invisible). Here the observer would expect a delay between picture and sound if an experience similar to reality were to be reproduced. Likewise, if a soccer match is being shown using an image from a camera giving a "wide shot", covering the whole pitch, then a delay between seeing the ball kicked and hearing the sound would be natural.

There are situations commonly seen in television productions where quick changes between close-up and wide shots are made. One example is the reporter in a crowded public space first seen talking in close-up , but then, as the camera "zooms out", revealed to be in a large space seen from a considerable distance. A second example is in sports television, there is frequently a change from wide shot to close-up and back.

In these cases there would be a changing perception of what delay from picture to sound is appropriate, if each shot were considered in isolation.

The nature of the sound can also suggest different context:-

If the sound accompanying the picture has been acquired using a microphone very close to the talker the character of the sound is different from that that would be obtained with a more distant microphone:

- The proximity effect of a directional microphone increases the low frequency content of the signal;

- The ratio of direct sound to reverberant sound is increased.

These factors give the sound a kind of intimacy, consistent with the observer being near to the talker. The observer might therefore expect the delay from picture to sound to be short. The reverse situation is where the sound is distant in character, and so a longer delay would be natural.

4. PRODUCTION EQUIPMENT

Within a normal television production environment, there are numerous pieces of equipment that will introduce differential delays to sound and picture. These are thought to be fairly obvious, and so will not be dwelt upon.

Vision mixer	Typically introduces a video frame delay.
Router	Should not introduce significant delay, unless asynchronous video sources are handled.
PAL 8 field synchroniser	Could introduce up to 160ms delay to video.
Audio mixer	Typically introduces only a small delay.

There is, however, the possibility to introduce significant change to audio or video signals during production. This is discussed in the next section.

5. PRODUCTION PROCESSING

During the production of a television programme some properties of the audio and video signals may need to be changed. There are many possible processes, and only some of the more common ones are described here.

5.1 Standards conversion (change of frame rate)

The process of temporally resampling a video signal is non-trivial, especially if interlace is involved. The problem arises largely because the frame rate is much slower than is required for accurate motion capture. Sophisticated motion compensation can be employed, and filters used that combine several frames. This introduces delay into the video signal path.

5.2 Standards conversion (change of aspect ratio)

Spatially resampling a video signal is conceptually simpler than temporal resampling. One may still need to consider that a raster scan means that different places in the picture have been sampled at different times. The process may require simpler processing than frame rate conversion, but in practice will introduce a frame or two of delay.

5.3 Digital video effects (DVE)

Any process involving mixes, wipes, changing the shape of the picture will typically introduce a frame or two of delay. The theoretical minimum delay for a mix is very small, but it's much easier to put in a frame buffer.

5.4 Low bit rate coding, including DVB

Both audio and video coding usually use some kind of "frame" (for example in MPEG audio) or "group of pictures" (MPEG video). This implies a delay because encoding cannot be completed until all the data of a frame or a picture required for prediction has been received by the encoder. Note that even software encoding and decoding of audio can introduce a "delay" (for example, in MPEG Layer II it is typically 481 samples).

Timestamps in MPEG programme streams are used to ensure that audio and video signals are decoded and presented at the same time. A clock signal is contained in the stream (system time clock - STC), and each frame (audio and video) has a time, according to that clock, at which it should be output (its presentation time stamp – PTS) (Figures A3 and A4). Incorrect generation of, or careless regard for, these bits of information can be disastrous for synchronisation. Many instances of perceived mis-timing in broadcast television reception are due to mis-handling of STC and PTS in the receiver. (Figure A5)

5.5 Sampling frequency conversion

The analogue of standards conversion for video, sampling frequency conversion of audio introduces a delay. The length of delay depends on the implementation of the converter. The filter length would be expected to be only a few, or a few tens of samples, and so a single conversion would not be significant.

5.6 Audio dynamic range compression

The "look ahead" required to subdue signal peaks unobtrusively can only be achieved by introducing delay. The longer the delay, the more slowly a gain change can be made.

6. REPRODUCTION EQUIPMENT

6.1 Display

Conventionally a cathode ray tube (CRT) display scanned an electron beam across and down the screen at the same rate as the incoming video signal (and indeed as the pick-up in a tube camera). The CRT phosphor exhibits some lag, but this is much less noticeable in a modern display than in early models.

Unfortunately, this simplicity no longer holds. The first break with convention was in the frame-doubling CRT display, where each field was presented twice (or possibly each frame twice).

Ignoring for the moment projection devices, which formed a small minority of the TV market, the main TV display devices are now flat panel devices, such as the LCD (the currently dominant technology) and Plasma (PDP) which has a significant minority market share.

A plasma display presents an image in a time-division multiplex of on-off pulses. The less significant bits in the digital signal are represented by extremely short pulses interpolated between the longer pulses, representing the more significant bits, which may be divided in time into a careful arrangement of sub-fields, aimed at minimising effects of colour fringing on moving edges.

LCD is in essence an "always on" device. Thus the image is maintained for a complete field period. This clearly has an effect on the visual perception of timing instants. In addition, the LCD is typically quite slow to change from the transparency required of each cell in one field period to that required for the next. This results in smearing or lag. In more recent products this is much improved, using techniques such as in-plane switching, vertical alignment or pin-wheel alignment, although the cheapest devices, using conventional twisted nematic LCDs, exhibit a very poor lag performance.

Flat panel devices are typically progressive, not interlaced, in their presentation of an image, and thus an interlaced signal must be "de-interlaced". This typically involves a processing delay, and together with other processing delays in the display, may add up to as much as 80ms. However, flat panel devices capable of displaying an interlaced signal without electronic de-interlacing exist, both as plasmas and LCDs, but are a small part of the market. In any case, any scaling undertaken on the signal to fit a fixed panel architecture, or to add overscan, inherently requires de-interlacing before the scaling.

Projection devices fall into four main groups. Those which use a CRT or LCD are not very different from a "direct view" device of that type. Digital micro-mirror displays may use a single device with a colour-wheel to display each colour sequentially. Such "colour sequential" devices thus present the red, green and blue components of the signal at different instants in time. More sophisticated ones may present a higher sub-field rate, or even have a white segment in the colour wheel in addition to the primary colours.

The frame rate that the display uses may differ from that of the video signal it is displaying. In this case the display performs a "standards conversion" operation. There are several ways in which this can be done, and the delay introduced by the processing can be different, and can vary (in much the same way that it does with "3:2 pull-down").

6.2 Loudspeakers and acoustic environment

An obvious factor affecting the delay between picture and sound is the distance between the loudspeaker (or, more generally, the electro-acoustic transducer) and the observer. The range of values can make a significant contribution:

Transit time of sound from loudspeaker to observer:

cinema	~50ms
home cinema	~15ms
living room	~10ms
bedroom	~5ms
handheld	~2ms
headphones	~0ms

Secondary effects of loudspeaker position are on the overall timbre of the sound and on the ratio of direct sound to reverberant sound at the observer. This could contribute to the impression of intimacy or distance leading to an altered expectation of picture to sound delay.

If reflections of the direct sound have a significant amplitude then these could provide misleading, or at least alternative, cues of synchronisation.

Headphones provide very good control in some ways: the delay introduced is very small, the ratio of direct to reverberant sound is very large. However, for in-ear headphones, so called "ear buds", the variation in timbre can be highly dependent on the goodness of fit.

In the discussion of programme composition, the issue of sound character, intimate contrasted with distant, was mentioned. The fidelity of the loudspeaker, or other electro-acoustic transducer, also affects the character of the sound heard by the observer.

A poor reproduction of low frequencies, giving a "tinny" sound, reduces the excess of low frequencies associated with the intimate sound. This might therefore be expected to give an impression of greater distance, and so an expectation of longer delay from picture to sound.

6.3 Audio signal processing in the receiver

Many domestic receiver/amplifiers offer signal processing. This might be frequency response alterations for "midnight mode", stereo widening, matrix surround decoding, or even for digital bit rate reduction decoding (for example Dolby Digital or DTS decoding). Some of these processes can be implemented with insignificant extra audio delay, some with a larger, but well defined delay. However, the manufacturer is mostly free to do what they wish.

Some devices explicitly include audio delay compensation, although this can only be used to add more delay, not to remove it.

7. PERCEPTION PROCESSING

Human perception of images and sounds is a complex process. There are different stages involved, that do not happen instantaneously. To simplify the situation, we might consider that there is a reflex response and a cognitive response. One happens before the other, as demonstrated by the facts that we can blink to prevent impact by an insect before we "see" the insect, and we can be startled, and physically jump, in reaction to a sound before we "hear" the sound. Research, by other organizations, is still on-going into the physiological and neurological processes involved.

For the purposes of televisual perception, understanding these effects might be useful for providing insight into subjective measurement techniques for quality assessment.

7.1 Expectation based on distance from screen

Whilst the distance between the observer and a television screen has little effect on the delay of the image, it ought to introduce an expectation to the observer. Although the 2-dimensional image might represent an object at any distance, the observer can clearly see that the screen is at a particular distance. The objects being viewed are on the screen, but the viewer constructs a 3-dimensional illusion containing them, based on prior real-world experience. Sound made by those objects might be expected to arrive at a time determined by the distance from the screen, but this expectation is over-ridden by the distance implied by the content of the image. (Figure A6)

7.2 Interpretation of distance based on screen size

A "talking head" can be framed by the camera to occupy, for example, three quarters of the height of the image. Displayed on a screen of height 30cm this would appear to be approximately actual size (depending on the size of head). Domestic television screens, particularly flat panel types (plasma, LCD) are very often bigger than this. This image of a talking head would therefore be displayed bigger than actual size. The bigger the screen, the bigger the image.

The apparent distance of the object depicted, based solely on the size of the image, gets smaller as the screen gets bigger. In the case where the true size of the object is known, as is the case with the "talking head" replacing the 30cm high screen with a 60cm high screen would move the head half way between observer and screen. (Figure A7)

7.3 Distances represented in 3-dimensional television display

3-D television is still in its infancy and there are numerous problems still to be resolved. One of these is the problem of maintaining correct representation of the depth of the scene.

Consider a situation where the actual scene has dimensions H x W x D (height, width, depth) and the screen image has dimensions h x w x d.

For 2-D TV (with some dependence on screen size) the percentage of time for which a particular ratio of "actual" to "screen image" size could be as follows:

$H < h$	(5%)	(extreme close-up)
$H = h$	(30%)	(close-up)
$H > h$	(30%)	(normal)
$H >> h$	(35%)	(long shot)

There is an apparent depth, even in the 2-D image, because our brain has become accustomed to interpreting 2-dimensional representations of 3-dimensional realities.

For 3-D, image depth (d) could be scaled according to h/H,w/W. However, the depths (front to back distances) in a large scene (a soccer pitch, for example) would all be miniaturised in the same way that the heights and widths are. This would make real scenes look like scale models. So, the image depth should remain the same as it was in reality.

It is common practice to "zoom in" on a distant subject. When this is done in 2-D the viewer's interpretation is that the subject is close. When this is done in 3-D the usual action is to increase the separation of the two cameras capturing the scene to make the apparent depth of the image match that implied by the subject size - in other words the apparent depth is reduced. Not to do so could make the subject appear un-naturally large, rather than closer.

7.4 Speed of processing by the visual system

There is a 3-D television effect that makes use of a dependence on brightness of the speed of visual perception: The "Pulfrich effect" is that if one eye is covered by a dark filter, for example one lens of a pair of sunglasses, that eye perceives the visual scene later than the other eye. If there is lateral motion in the scene then the time difference in perception causes a spatial difference in perception. As a result, the two eyes see laterally moving objects in different places, and this is interpreted as depth information.

For large differences in intensity, the time difference could be 30ms. If this means that watching television in a dark room requires putting an extra 30ms of delay in the audio then this is significant. Note however, that the Pulfrich effect relates to a differential intensity between the two eyes, and that the effect is reduced at higher levels of illumination[1]. No work has been conducted as part of this study into the effect of screen brightness on audio-video synchronisation perception.

8. THE EFFECT OF IMAGE RESOLUTION ON AUDIO VIDEO SYNCHRONISATION PERCEPTION

With the introduction of digital high definition television services in the UK, interest was renewed in the problem of audio-video synchronisation. Together with the problems inherent with introducing a huge amount of new equipment into the system, a concern was raised that the current synchronisation tolerance limits might need to be revised for high definition video: it was hypothesised that increased image resolution might lead to increased perceptibility of errors.

EBU Technical Recommendation R37 was revised in 2007 to include references to high definition and to compressed audio decoding delay.

Experiments have been conducted specifically to identify whether there is a need for a more stringent requirement for synchronisation for high definition than for standard defintion. These took the form of asking human subjects to correct deliberately introduced synchronisation errors in an extract from a television programme. By examining the accuracy with which they were able to perform the task for the different picture definitions, we can determine whether there is a more stringent requirement.

8.1 Experimental arrangement

The task given to the subjects was to correct a series of arbitrary synchronisation errors introduced into an extract from a television programme played in a loop. (Figure A9). They did this with standard definition pictures and with high definition pictures. The process is explained with reference to Figure 1.

The television programme extract is played from a server, with the audio advanced by a fixed offset of 300ms. The server plays out via an HDSDI interface, with the linear PCM audio embedded in the HDSDI stream. The format of the video is uncompressed high definition 1080i25 (1080 lines, interlaced, 25 frames per second). For standard definition replay this stream is immediately passed through a hardware standards converter (not shown) and converted to 625i25 (on SDI).

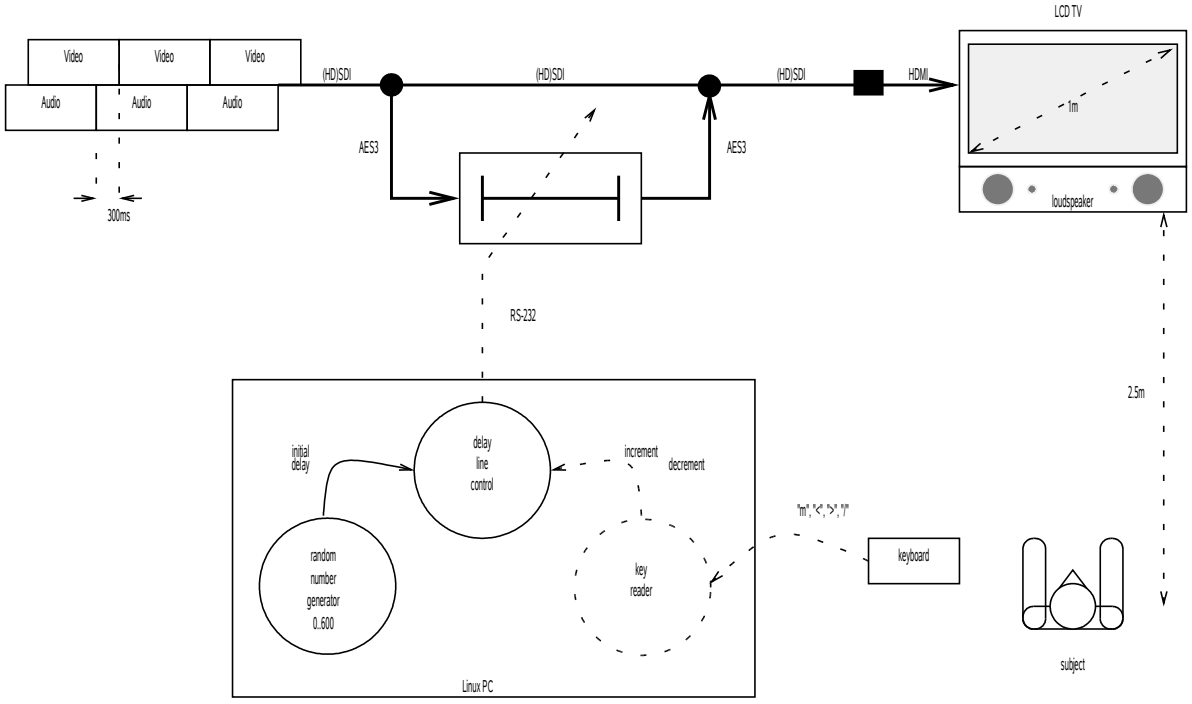

Figure 1 Schematic diagram of audio-video synchronization subjective test experiment

The audio is extracted from the HDSDI or SDI stream and routed through a digital audio delay line. After the delay line, the audio is embedded back into the stream, and the stream is converted to HDMI to connect to the display. The HDMI signal contains the audio and video and was presented using a domestic-style LCD TV with built-in loudspeakers.

The LCD TV automatically adjusts itself to the resolution of the signal on its HDMI interface. It displays the pictures at the same size independent of definition. It was observed that when given standard definition pictures there was an additional delay to the rendering of the video of 20ms.

The delay line is controlled by a Linux PC. Software on the PC sets an initial delay between 0ms and 600ms and then gives control to the subject. The subject can increment and decrement the delay in large and small steps using four keys on the keyboard. When the subject is happy with the synchronisation they press another key and the PC records the final delay setting, and sets another, random, delay.

Subjects were requested to repeat this process for a period of 30 to 40 minutes for one definition (SD or HD) and, during a separate session at a different time, for the other definition.

As well as recording the final result of each trial, the software records each key press. This allows a study of individual strategies, or biases in the system.

Eleven subjects participated in the experiment.

8.2 Summary results

The complete study is, at the time of writing, unfinished. The tests were started with an LCD TV of a type and size that is representative of a typical domestic installation. However, more accurate portrayal of motion is to be expected from a cathode ray tube (CRT) display because of the much shorter persistence of the image. The tests are being continued with a high definition CRT. Test material has also been recorded using a high frame rate camera (300 frames per second), so a third phase of testing will focus on the influence of temporal resolution on audio-video synchronisation perception.

8.3 Mean value of final time offset

Note that in this test method, the absolute time offset between sound and picture on the server need not be known to any great precision. It is sufficient that the time offset in the replayed signal is constant. For measuring the sensitivity of subjects to synchronisation it is only necessary to look at the width of the distribution of the test results.

The mean value of the test results for the SD pictures was 18ms greater than for the HD pictures. By objective measurement (using an oscilloscope and light pen) it was established that the LCD TV introduced a extra video delay of 20ms working in SD.

8.4 Distribution of final time offset

A histogram showing the distribution of the results from HD and SD tests is shown in Figure 2.

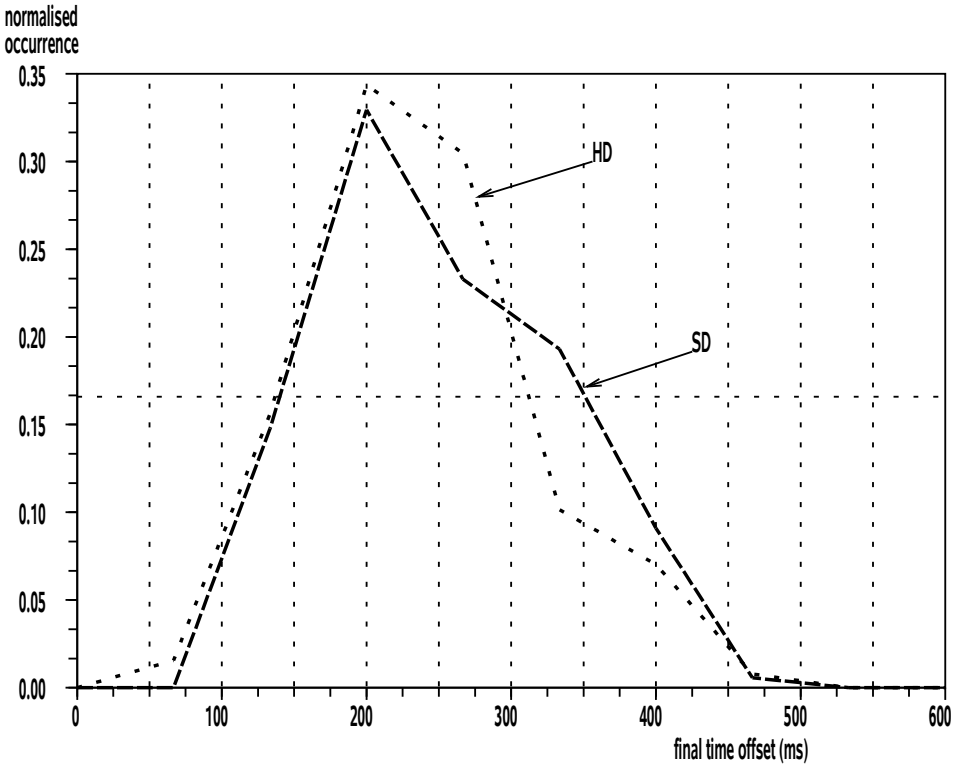

Figure 2 Histogram of audio-video synchronization subjective test results

Note first the assymetry of both distributions. This is consistent with the observation that sound in advance of picture is unnatural and therefore less tolerated than picture in advance of sound.

Secondly, it can be seen that the width of the distribution of the SD results is wider than that of the HD results. By inspection, about half-way up the distribution, the SD curve is 200ms wide, while the HD curve is 170ms wide. It should be noted that the apparent shape of these distributions is sensitve to the number of bins in the histogram. The difference between the standard deviations of the two data sets is 2.3ms (stdvHD=77.8ms, stdvSD=80.1ms)

There is considerable variation in the final offsets. Inspection of the records of key presses and discussion with the subjects reveals something about this. Some subjects are not very sensitive to synchronisation errors - they adjust the time offset but stop well before the "correct" value. Some subjects have difficulty deciding whether the sound is before or after the picture, oscillate either side of the "correct" value, and give up. Subjects could initially tell that the synchronisation was not correct, but after trying to correct it, eventually could no longer reliably tell what was correct and what was not.

Overall, subjects were consistent in their opinion that the task was difficult. Tests with a CRT may be found to be easier. Other test methods are also being considered. (Figure A10).

9. GLOBAL INTEREST IN AUDIO VIDEO SYNCHRONISATION

A number international bodies are currently working in the field of audio video synchronisation. The reason for this is that the problem, decades old, is not solved, but getting worse. The introduction of digital broadcasting, high-definition broadcasting, set-top boxes with audio and video outputs to separate audio and video transducers, displays with significant signal processing inside, have all contributed to make it easy for huge errors to arise.

The International Electrotechnical Commission has been working in this area for a number of years. There have been two publications from this: IEC 62312-1 [2] and IEC 62312-2 [3]. The first of these defines some terminology for characterising the delays to audio and video that a device under test (DUT) might cause, specifies some environmental conditions, and describes measurement procedures in the most general of terms. The second again characterises the DUTs and describes some ways in which timecode, or the equivalent, available on certain interfaces such as IEC 60958 and IEC 61883, can be used to maintain correct synchronisation.

A third publication is imminent at the time of writing this paper. IEC 62503 "Multimedia quality - Method of assessment of synchronization of audio and video" describes a subjective test method for measuring subjective quality as a function of audio-video synchronisation error, and a means of deriving an appropriate correction delay for the DUT from the subjective measurements. The test method uses the familiar 5 grade impairment scale - "imperceptible", "perceptible but not annoying", and so on.

The Advanced Television System Committee also has a specialist technology and standards group working on audio-video synchronisation. The focus of its work is within broadcasting, on the coding and multiplexing, and set-top box end of the problem. A working group within the Consumer Electronics Association has been created as a result of this with the intention of producing recommendations for the use of MPEG timestamps to maintain correct synchronisation. Once this task is complete, the ATSC specialist group will continue its work.

The Society of Motion Picture and Television Engineers (SMPTE) is also working on audio video synchronisation, together with the European Broadcasting Union (EBU). This has already produced a "lip-sync cookbook" intended to be a short guide to best practice in preventing synchronisation errors.

There is also the activity of the "Joint EBU - SMPTE Task Force on Time Labelling and Synchronization", that indirectly could help achieve synchronisation. In essence, its aim is the production of a "next generation" replacement for timecode. A process of drawing up user requirements, followed by a request for technologies, should produce results that can then be formally standardised.

Within the AES, standards committe working group SC-02-01 has a project to define standard measurement methods for assessing audio video synchronisation. This will go further than the current IEC 62312-1-1 in that it will define audio and video test sequences that can be used subjectively or objectively for measurement. It is also intended to provide simple hardware examples to encourage production of more, simple, measurement systems. The early stages of the work are described in the next section.

There is, clearly, an overlap in the work areas of these organisations. However, everyone agrees that the problem needs to be addressed, and there is willingness from all parties to improve the current situation!

10. OBJECTIVE AND SUBJECTIVE MEASUREMENT OF AUDIO VIDEO SYNCHRONISATION

A simple test signal, an electronic "clapper-board" that may be used subjectively to estimate, or objectively to measure, audio video synchronisation is as follows:

Figure 3 shows a single frame from the video sequence. The horizontal green bar grows from left to right. The frame shown is from the middle of the sequence so the bar is half-way across. At the start of this frame period a click starts in the accompanying audio track (actually the sound of two blocks of wood being banged together). A red flash appears at the top of this frame and is present for one frame only.

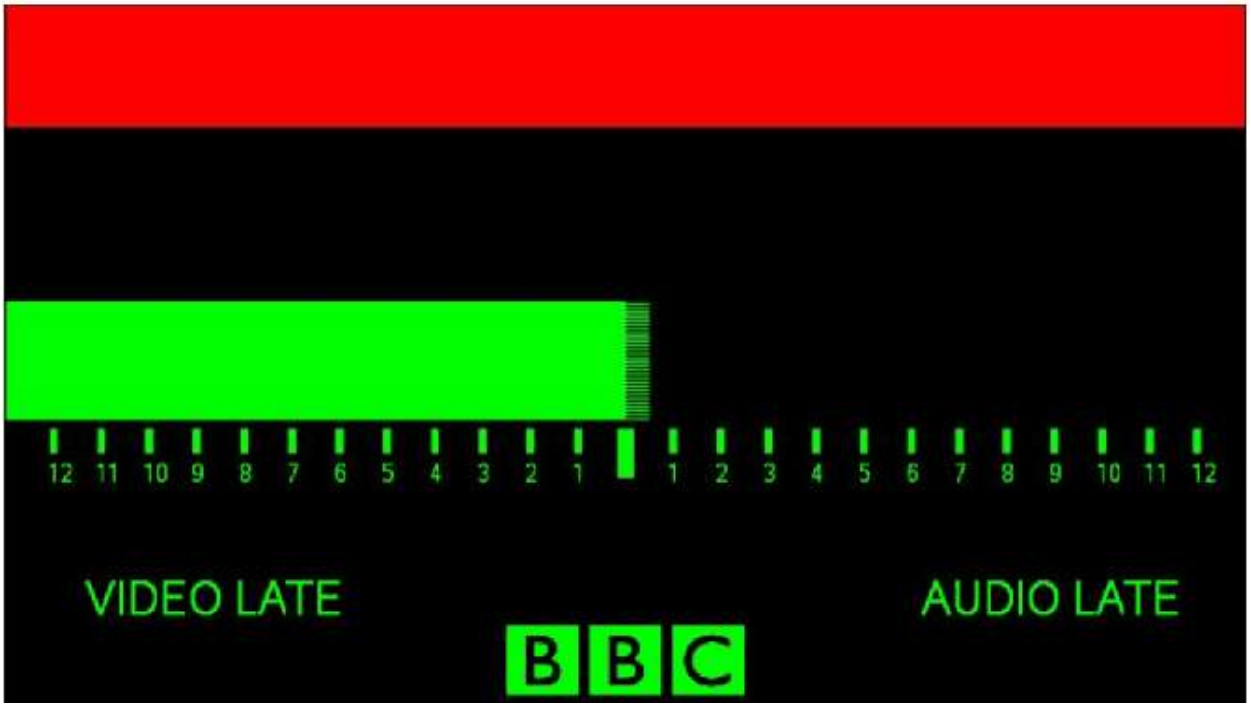

Figure 3 Central frame from clapper-board video sequence

Note that this is a frame consisting of two fields. In the first field of the frame the green bar just reaches the central (unnumbered) tick mark. In the second field the green bar reaches half-way to the next tick mark to the right (numbered 1). Since this particular display is interlaced only half the lines are illuminated in the second field, appearing as a feathered edge.

To simplify objective measurement when using analogue component video (RGB), the red channel contains only the flash. All other video in the sequence is green. No white (or pink, or red), captions should be added to the sequence. The red channel may therefore be diverted to an oscilloscope, while the green channel can remain connected to the display. Green was chosen as the dominant colour because video syncs are often put on the green channel.

The wood block sound lasts for several frames, but this was found to be subjectively the most acceptable sound from many tested.

10.1 Using the clapperboard signal subjectively

Start by looking at the rightmost tick whilst listening for the click. If you can say that the click has not happened by the time the bar reaches the tick, look at the next tick to the left.

Continue listening and working leftwards and stop when you can no longer say that the click has not happened before the bar has reached this tick. The offset is that indicated by the last tick where you could not say that the click had not happenned.

10.2 Using the clapperboard signal objectively

For measuring a channel, the red video and audio signals can be examined using an oscilloscope. For measuring using a display, an example of some hardware that can be used is shown here.

The hardware uses a "Microchip Picdem 2 Plus" evaluation board with a small amount of additional audio and video interface circuitry. Figure 4 shows the circuit diagram of the interface.

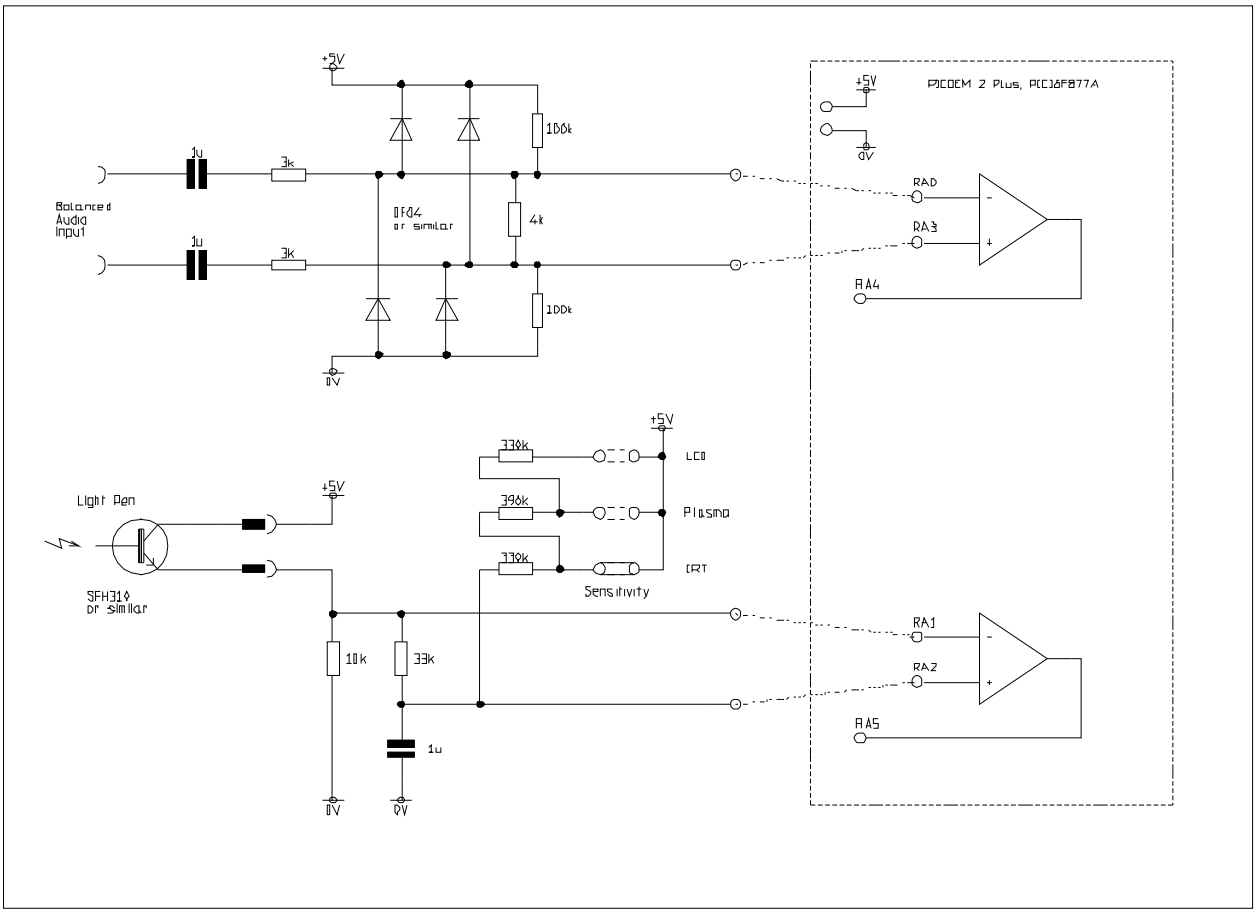

Figure 4 Schematic diagram of audio and video interface hardware for Picdem 2 Plus evaluation board

The audio signal is received using a balanced input circuit, while the red flash of the video is detected using a "light pen" (a phototransistor) held on the screen.

The software running on the processor on the evaluation board simply measures the time between events detected on audio and video inputs. There are constraints built into the timing mechanism to distinguish between audio late and audio early assuming a 1 second repetition rate of the clapper-board sequence.The measured time offset is displayed, to the nearest millisecond, together with a message showing audio or video late on the LCD panel. This information is also sent out of the RS-232 port on the board so that it can be seen, or logged, remotely (for example by using a PC and terminal application). It is planned to make the software available under the GNU GPL.

The prototype system is shown in Figure 5. The audio and video interface board, with the light pen, is in the foreground, and the Pic evaluation board is resting on the laptop that is logging the results.

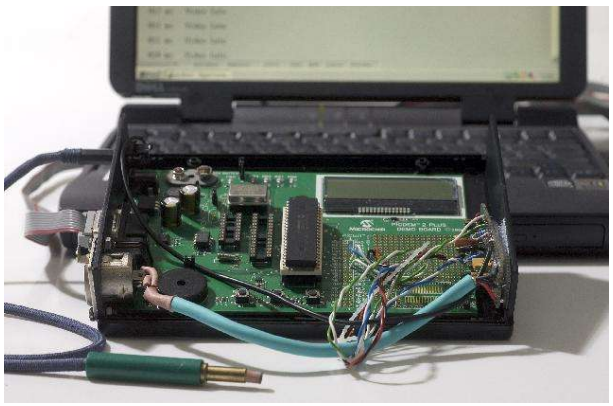

Figure 5 Prototype A/V measurement system, with laptop PC logging the measurements

11. THE FUTURE

In the battle for bit rate on digital broadcast networks, audio constantly has to defend itself against video. Audio data reduction ratios are typically between 5:1 and 10:1, whilst video could be 100:1. From the analysis of the factors affecting video and audio perception on television, one can argue that the portrayal of audio is very close to reality, whilst the portrayal of video is far different from reality (for so long as images are presented as 2-dimensional representations, with a subtended angle of significantly less than 1 steradian, and with a very low temporal sampling rate). The relative compressibility of the signals is consistent with this.

As the quality of the video experience improves, with higher spatial resolutions, higher frame rates, and the inclusion of the third dimension the visual experience will become much more like "being there". It will approach the illusion of reality that audio can already achieve. As a result, it seems inescapable that our expectations of the accuracy of audio-video synchronization can only increase, along with the technical complexity of maintaining it.

12. REFERENCES

[1] Lit, A., "The magnitude of the Pulfrich stereophenomenon as a function of binocular differences of intensity at various levels of illumination" The American Journal of Psychology, Vol LXII, April, 1949

[2] IEC/TS 62312-1-1, "Guideline for synchronisation of audio and video - Part 1-1: Measurement methods for synchronisation of audio and video equipment and systems - General", March 2008

[3] IEC/TS 62312-2, "Guideline for synchronisation of audio and video - Part 2: Methods for synchronisation of audio and video systems", June 2007

APPENDIX A: FIGURES FROM CONFERENCE PRESENTATION

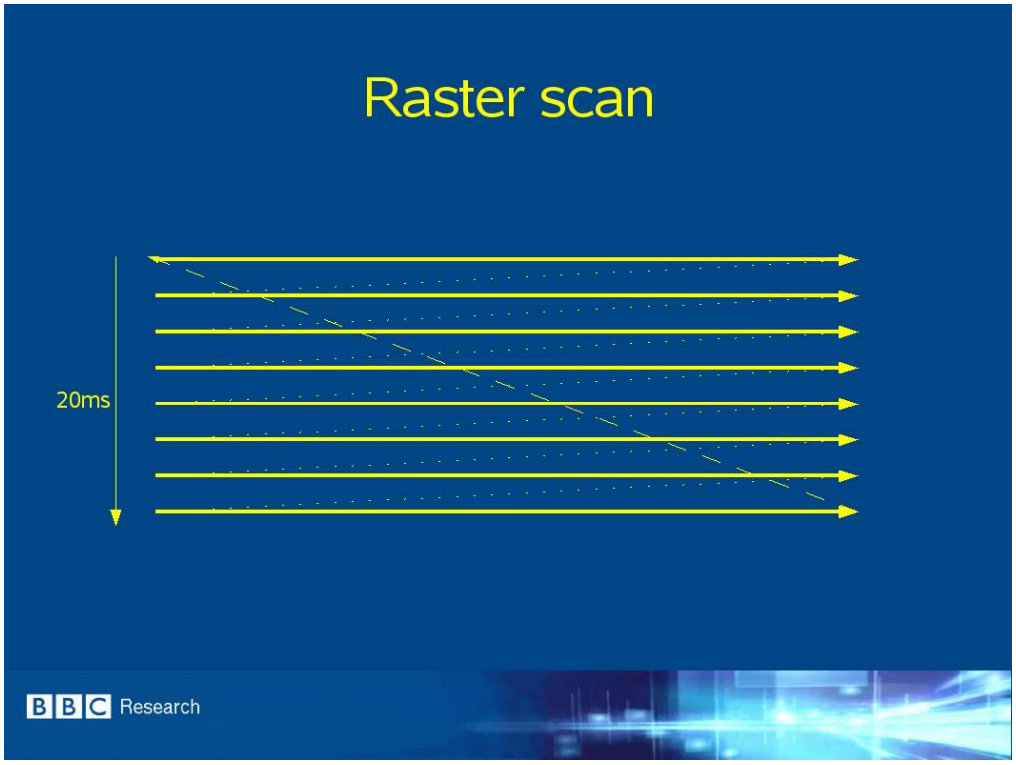

Figure A1 Example of top-to-bottom, left-to-right, television picture scan for one field at 50 fields per second

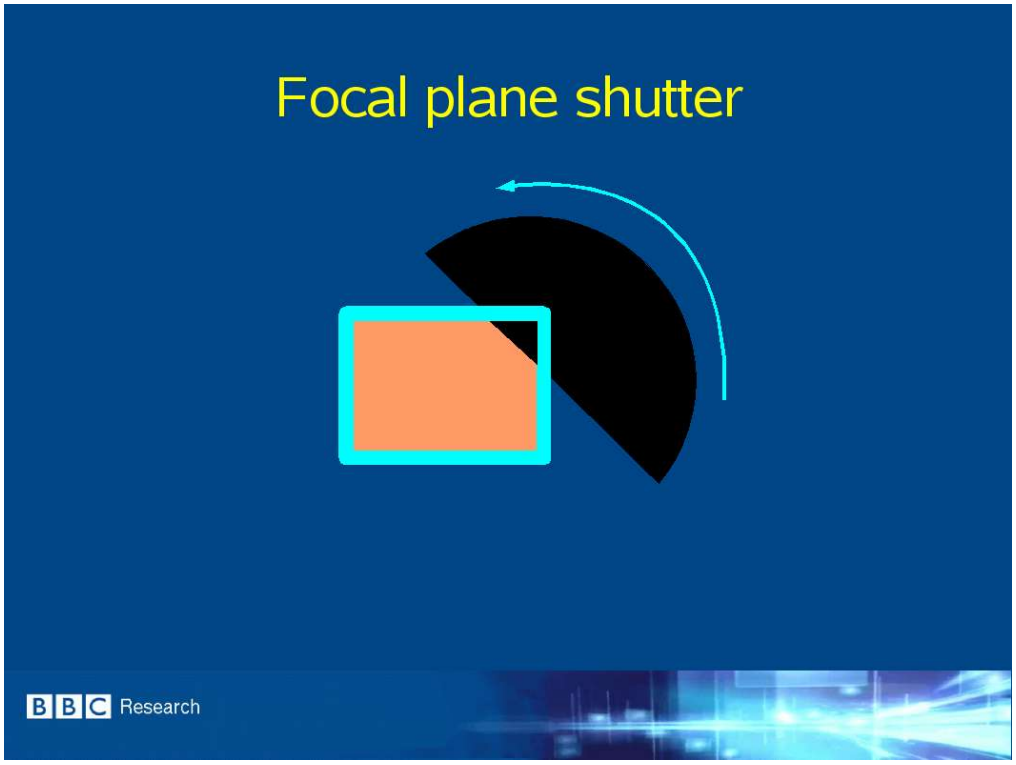

Figure A2 Cine film focal plane shutter, showing different exposure time of different film areas

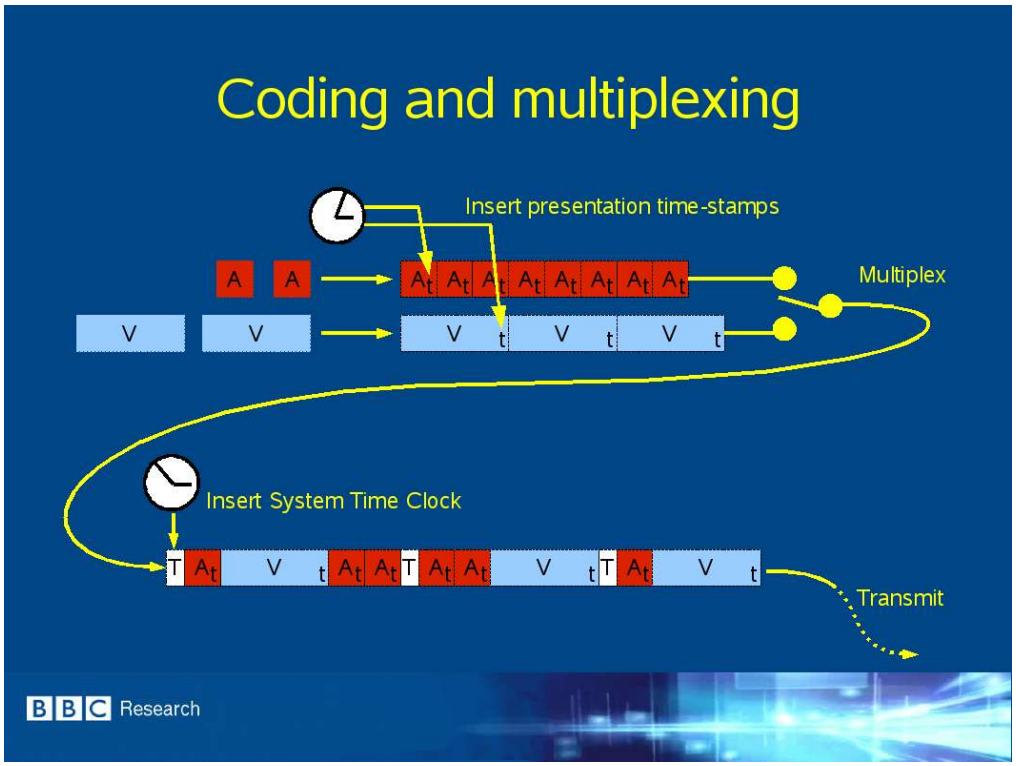

Figure A3 Insertion of timestamps and multiplexing audio and video packets, typical of MPEG

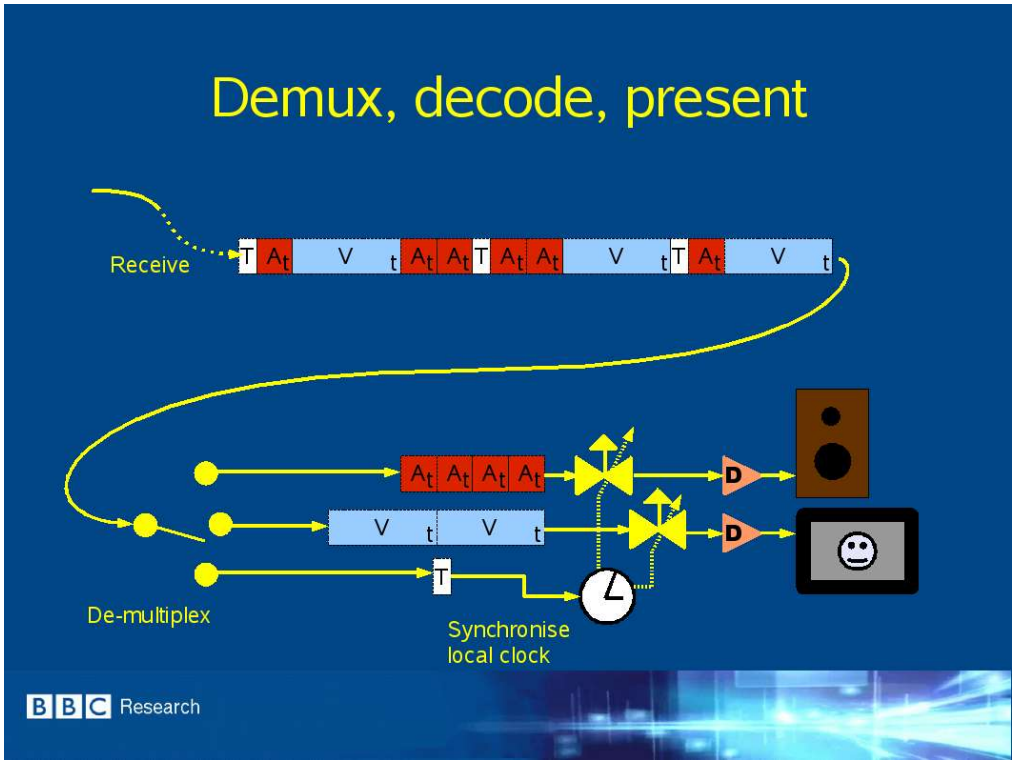

Figure A4 Demultiplexing and use of timestamps to regulate decoding and presentation of audio and video

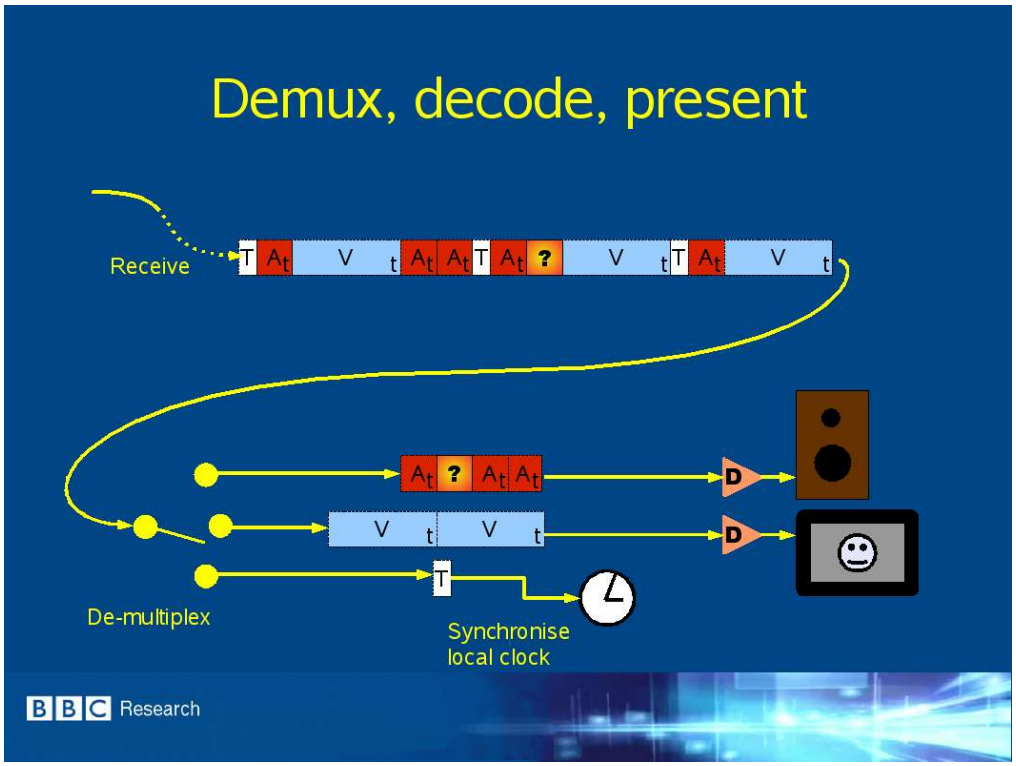

Figure A5 Sources of error: lost packets and failure to use timestamps or recovered system time clock

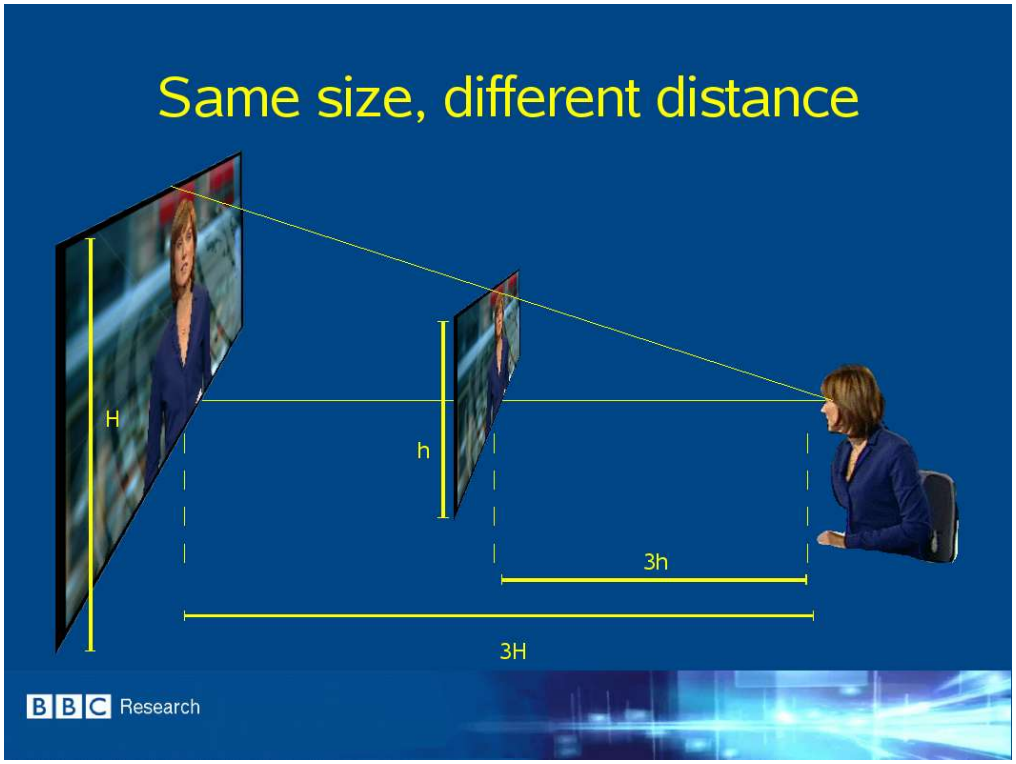

Figure A6 Example showing how larger screen size can give same apparent image size at greater distance

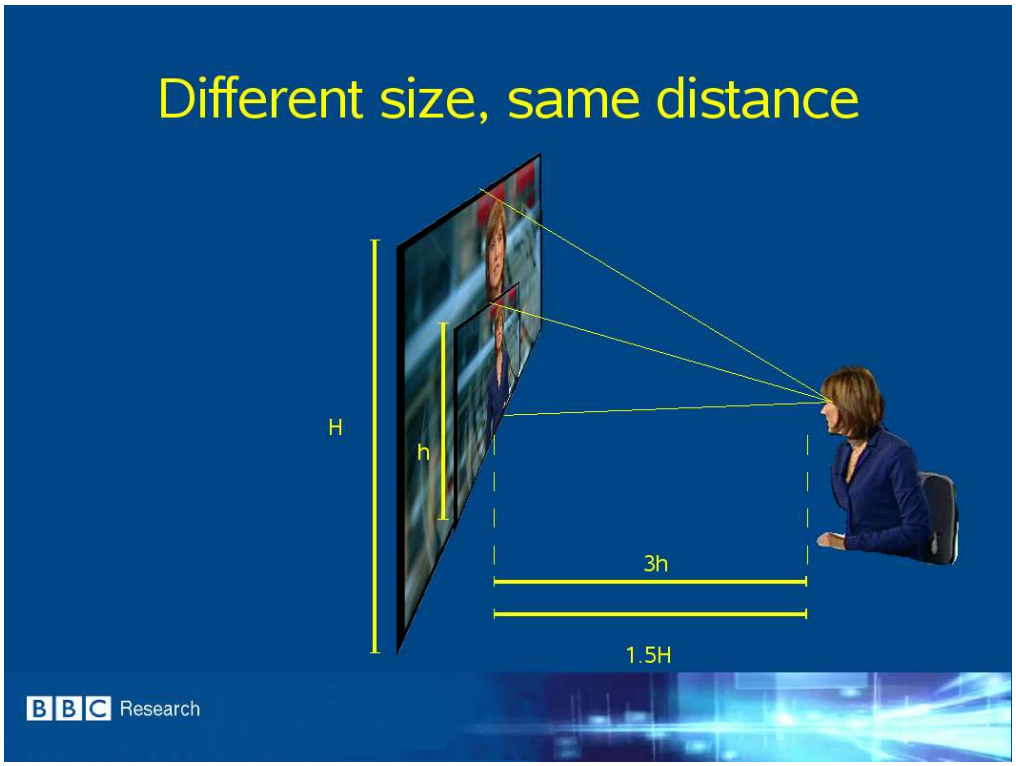

Figure A7 Example showing how larger screen size gives larger image at same distance

Figure A8 Different image compositions : close-up, and combined middle distance and further away.

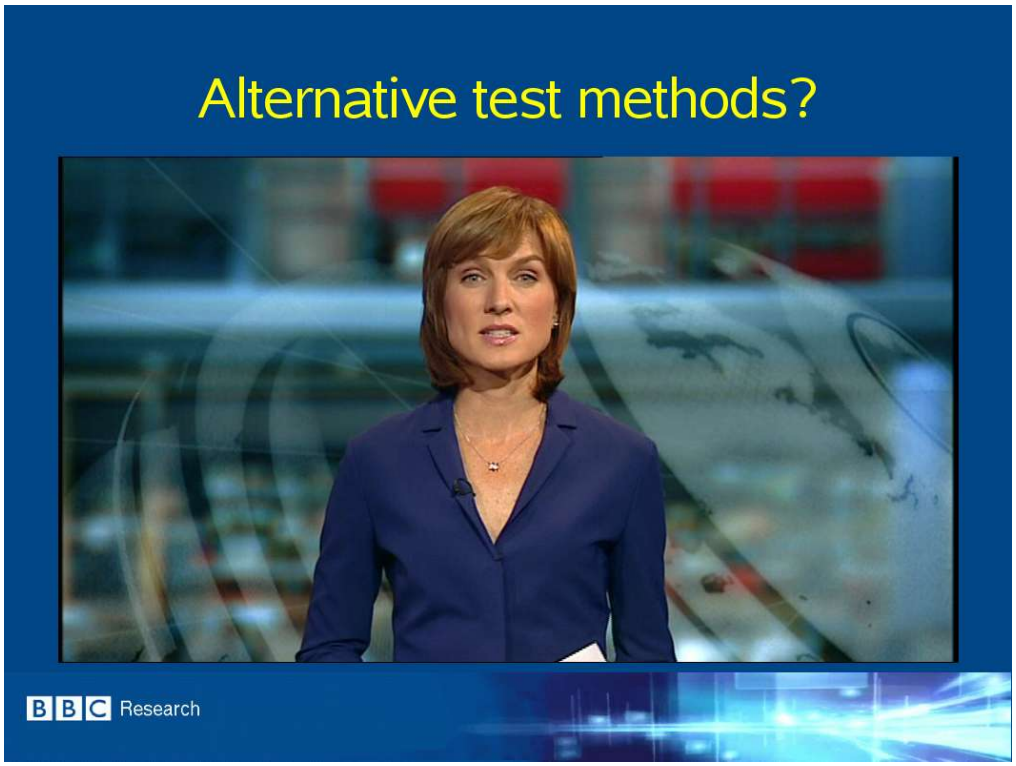

Figure A9 Example of image presentation typical of that used in the subjective test

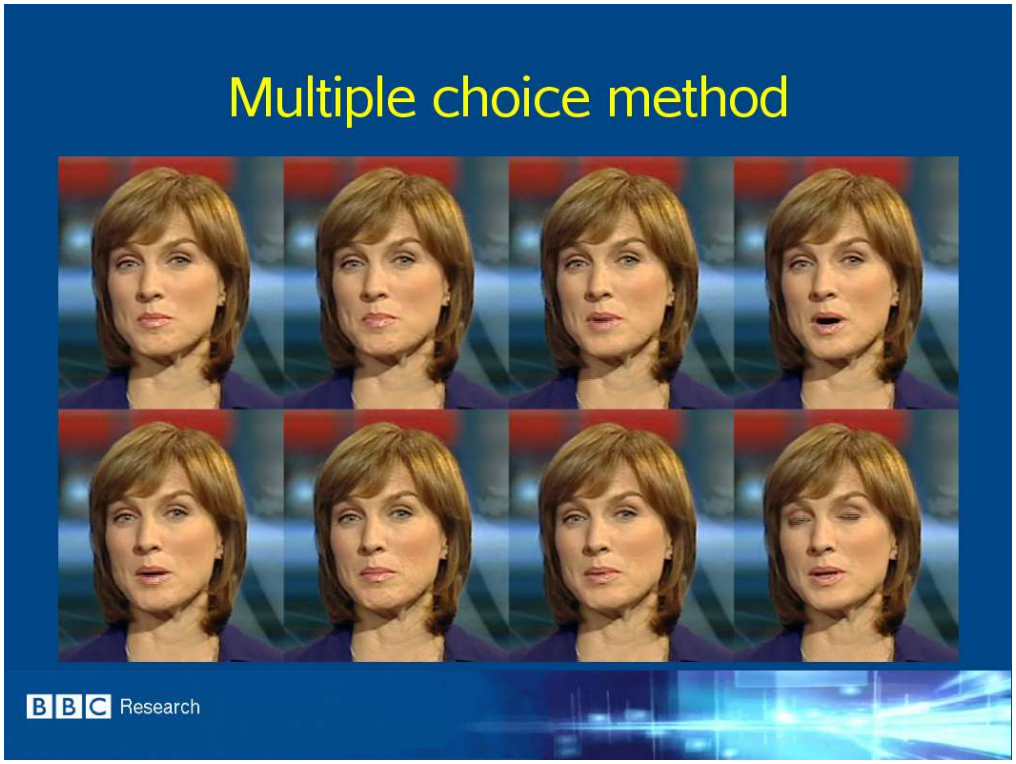

Figure A10 Alternative image presentation, for a “multiple choice” subjective test where the subject is asked to indicate which tile shows the correct synchronisation.