REPORT 1204

AUTOMATIC SYNCHRONIZATION OF VIDEO AND AUDIO AFTER TRANSMISSION

(Question 47/10)

(1990)

1. Introduction

Whenever the video and audio components of a television signal are transmitted over separate paths, there is the possibility that their transmission times will be different. When the difference is such that the viewer perceives a separation between vision and sound, enjoyment of the programme can be impaired. The subject has been studied in detail in Australia and is reported in [CCIR, 1986-90a].

A guide to the amount of allowable time difference, may be taken from the standards for film, where the picture can lead the sound by no more than two frames (83 ms) or lag behind it by no more than one frame (42 ms). The reason for the asymmetry in these allowable time differences is that sound arriving after vision is a familiar experience while sound arriving before vision is not.

Frame synchronizers are used increasingly in long distance transmissions because they confer so much operational flexibility. However, large delay differences can easily be introduced unless steps are taken to have the sound delayed equally with the vision within the synchronizer.

Delay differences are often compensated manually by an operator trying to judge the correction before other viewers can perceive it, but this is a difficult process. A method is under study by which the relative delays between vision and sound can be corrected automatically.

An audio signal in digital form is easily delayed be feeding it into a shift register and the delay can be varied by changing the address from which the output is taken. The scheme shown in Figure 1, only provides an increased delay of the audio component.

A similar delay of video is more difficult to achieve. However there is little need for such a video delay because sound following vision is more easily tolerated, and is less likely to occur. The only foreseeable cause of increased delay in the audio path is the insertion of a digital FIR low-pass filter with a sharp cut-off. Such a filter might be needed, for example, in converting the audio sampling rate from 48 kHz in the studio to 32 kHz for transmission, but this would only introduce a delay of 5.3 ms.

2. Generation of time codes

No correction method is proposed for sound in analogue form as a method has been proposed already for this purpose [Cooper, 1988]. However, it is anticipated that sound and vision will be transmitted almost entirely in the future by digital means. So long as sound and vision are transmitted in the same bit stream, no problems of synchronization need arise. Once the video signal is processed, though, for frame synchronization or whatever other purpose, a re-synchronization method will be needed.

In the digital audio interface for broadcasting studios given in Recommendation 647, § 3.5 the channel status data carries, in bytes 18-21, a 32 bit binary code which conveys the number of the first sample of the current block. Even at a 48 kHz sample rate, the 2^{32} states of this code can convey unique numbers of samples over a period of more than 24 hours. Thus, in the channel status data of the audio interface, a time indication is present, every block of 192 samples, i.e. every 6 ms with a 32 kHz sampling frequency or 4 ms with 48 kHz, to a precision of 31.25 μs.

However, the digital audio interface is primarily designed to carry monophonic or stereophonic programmes in a studio environment and is unlikely to be used for transmission between studio centres. Thus any time-code information would have to be transcoded to take its place in the transmitted bit stream.

It is proposed therefore that a separate Audio Vision Synchronization (AVS) code be generated and transmitted with each signal, AVS(A) code in the audio bit stream and AVS(V) code in the vertical interval of the video. It is estimated that this would require 11 bits per block, i.e. an additional 1.83kbit/s with 32kHz sampling frequency or 2.75 kbit/s with 48kHz sampling.

These 11 bits would describe the timing of the first sample of the current block to the nearest millisecond. This would be done by taking the binary sample address for each block already present in bytes 18-21 of the channel status data of the Digital Audio Interface, as described earlier, and dividing it by the sampling rate in kilohertz, e.g. by 32, 44.1, or 48 as appropriate.

The resulting binary quotient would time the start of each block of digital audio in precise units of 1ms with 32kHz or 48kHz sampling and to within a maximum error of ±½ms with 44.1kHz sampling. With 11 bits, a total of 2048 separate states can be specified, i.e. the time code would have a complete cycle time of 2.048 seconds, from a count of zero to a count of 2047 and back to zero again.

The EBU time code specified for video indicates a count of pictures only, i.e. one every 40 ms in PAL. For NTSC, the count would be one every 33.37ms. Thus there seems no way of relating the audio time code to the conventional video time code. It is proposed therefore that an additional 11 bit signal, similar to that inserted into the audio signal, be inserted into every vertical interval of the video, indicating the time, in units of 1ms, at which it was inserted. It should be noted that this audio video Synchronizing time code is entirely independent of the conventional video time code.

It is not necessary to carry information on the sampling frequency within the AVS code as this has been taken care of already in the division of the binary address number of the Digital Audio Interface by the sampling frequency in kHz. However information about the sampling frequency can be recovered from the AVS code by counting the difference between the AVS time code numbers of successive blocks.

If the difference is taken between the numbers two blocks apart, the time will be 12.0 ms for 32 kHz sampling frequency, so that the number difference will be 11, 12 or 13. For 48 kHz sampling, the period will be 8.0 ms, producing a difference of 7, 8 or 9. Thus logic recognizing one or other group of difference numbers will identify the sampling frequency. A similar procedure applies for digital audio sampled at other rates.

The proposed procedure depends on high clock stability in the time codes. The stability of television timing signals is specified in Report 624 as ±5 Hz in 4.433 MHz for PAL and that of digital audio in Recommendation 646 as ±1 part in 10^5. This comparatively wide tolerance was allowed so as to provide for audio signals originating from remote portable equipment.

However, Recommends 3 of that document states - " when an item of digital audio equipment is operating in a free-running mode, the maximum tolerance for the internal sampling frequency should be ±1 part in 10^5. When items of digital audio equipment are interconnected, in sound broadcasting or television applications, provision must exist for locking the internal sampling frequency (e.g.: television synchronizing signals, broadcasting house master clock, high accuracy clock from a telecommunication network); - "

The proposed scheme should work satisfactorily within these tolerances.

3. Correction of Delays.

On reception, each of the two AVS time codes, which arrive only at intervals, of a block of audio or a field of video, is used to synchronize its own continuously-running time code, as in Fig.1. These two derived time codes are then compared and their difference used as the correction signal to vary the audio delay until that difference is zero.

It might seem that, with a cycle time of 2048 states, the system could correct delay differences up to 2.048 seconds. However, as the output difference, N(D), is taken as N(V) - N(A). So long as the video lags behind the audio, N(D) is positive, but if the sound ever lags behind the video, N(D) will be negative and be indicated by a large number, N(D) + 2048. Thus if the sound lags by 100 ms, then the indication will be 1948.

It is therefore proposed that if the difference between the two derived time codes lies between decimal numbers 1536 to 2047, it should be interpreted as lying between -512 and -1, and therefore used to set the audio delay to zero, along with an indication "Video leads audio". Thus the system can correct errors of audio leading video by a maximum of 1.535 seconds, sufficient for the largest errors that can be foreseen.

Because the video syncs operate asynchronously from the audio-derived time code, the indicated times of the video can be in error by a maximum of ±½ ms, so the indicated time difference N(D) between audio and video will also vary by this amount. It is therefore proposed that hysteresis be built into the system so that the delay correction number N(C) does not change until N(D) differs from it by more than ±2ms, and that once change does occur, it then proceeds until the correction N(C) is equal to N(D). This feature is not shown in Fig.1.

It is also proposed that initial errors greater than 3 ms be corrected immediately, but that errors of ±2 ms be corrected at the rate of 1 in 500, i.e. over a period of one second.

The variable delay can be implemented using already known techniques.

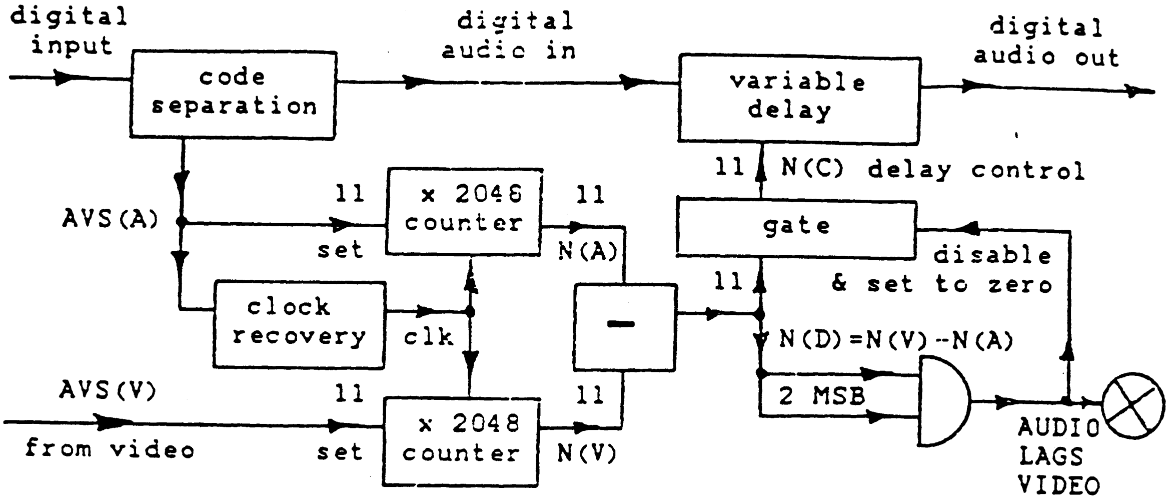

Figure 1

Schematic of automatic audio/video synchronizer

4. Transmission of A/V synchronization code

A possible method for transmitting the AVS code within each of the component signals is given in the following sections.

4.1 Audio

[CCIR 1986-90b] proposes a method by which high quality sound programme signals comprising a stereo pair sampled at 48 kHz can be carried in the ISDN, either at 1920 kbit/s (for H12 access) or 1536 kbit/s (for H11 access).

In the method proposed in the document, the channel status data of the digital audio interface, which include the 32 bit binary time of day clock, are transmitted with 48:1 time expansion. In other words, the data are transmitted transparently, but one complete channel status block is transmitted every 192 ms, instead of every 4 ms as in the case of the studio interface. The 11 bit A/V synchronizing code can be derived from these data as described earlier, the one disadvantage being that any change in transmission delay is detected only after a delay of up to 192ms. However this would not seem an important consideration.

Once the timing difference between the recovered time code and the initial time code, which might be up to 192ms, has been established, provision must be made to make any necessary fixed adjustment to it before the AVS code is derived. This aspect is not addressed in any detail in [CCIR 1986-90b] beyond the observation in section 10, Channel Status, that "Because only one data block out of 48 is transmitted, the two counters (local sample and time of day address code) must be incremented in the decoder by the appropriate amount".

4.2 Video. Before an 11 bit code can be inserted into the vertical interval, agreement would be needed on where it is to be put. The best place would be somewhere such as in the Programme Delivery Code proposed in [EBU, 1989].

5. Conclusion.

It is believed that the method proposed can synchronise the audio and video components transmitted over different paths, without occupying more than a minimum of "real estate" in either signal.

REFERENCES

COOPER, J.C. [1988] - Video-to-Audio Synchrony Monitoring and Correction - JSMPTE, September, pp 695 - 698.

EBU [1989] - Specification of the domestic video programme delivery control (PDC) EBU SPB 459 rev., March 1989.

CCIR Documents

[1986-90]: a. 10/315 (Australia), IWP CMTT/4-8 (IWP CMTT/4).