ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2018.0254

ICTACT JOURNAL ON IMAGE AND VIDEO PROCESSING, AUGUST 2018, VOLUME: 09, ISSUE: 01

PERFORMANCE ANALYSIS OF AUDIO AND VIDEO SYNCHRONIZATION USING SPREADED CODE DELAY MEASUREMENT TECHNIQUE

A. Thenmozhi and P. Kannan
Department of Electronics and Communication Engineering, Anna University-Chennai, India

Abstract

The audio and video synchronization plays an important role in speech recognition and multimedia communication. The audio-video sync is a quite significant problem in live video conferencing. It is due to use of various hardware components which introduces variable delay and software environments. The synchronization loss between audio and video causes viewers not to enjoy the program and decreases the effectiveness of the programs. The objective of the synchronization is used to preserve the temporal alignment between the audio and video signals. This paper proposes the audio-video synchronization using spreading codes delay measurement technique. The performance of the proposed method made on home database and achieves 99% synchronization efficiency. The audio-visual signature technique provides a significant reduction in audio-video sync problems and the performance analysis of audio and video synchronization in an effective way. This paper also implements an audio- video synchronizer and analyses its performance in an efficient manner by synchronization efficiency, audio-video time drift and audio-video delay parameters. The simulation result is carried out using Matlab simulation tools and Simulink. It is automatically estimating and correcting the timing relationship between the audio and video signals and maintaining the Quality of Service.

Keywords:

Synchronization, Software, Hardware, Audio-Visual Spreading Codes, Temporal Alignment

1. INTRODUCTION

The audio and video synchronization is defined as the relative temporal distinction between the sound (audio) and image (video) during transmission and reception. It is also known as audio-video sync, A/V sync and audio/video sync. Lip synchronization (lip sync or lip synch) refers to the voice that is synchronized with lip movements. Human can able to detect the distinction between the audio and corresponding video presentation less than 100ms in lip sync problem. The lip sync becomes a significant problem in the digital television industry, filming, music, video games and multimedia application. It is corrected and maintained by audio-video synchronizers.

In multimedia technology, the audio and video synchronization plays an important role in synchronizing audio and video streams. With the advancement of interactive multimedia application, distinct multimedia services like content on demand services, visual collaboration, video telephony, distance education and E-learning are in huge demand. In multimedia system applications, audio-visual streams are saved, transmitted, received and broadcasted. During an interaction time, the timing relations between audio-video streams have to be conserved in order to provide the finest perceptual quality.

Claus Bauer et al. [4] suggested the audio and video signatures for synchronization. The signatures extracted from audio and video streams for necessarily maintaining synchronization between the audio and video signals. During transmission, the audio and video streams are recorded by combining audio and video signatures into a synchronization signature. At reception, the equivalent signatures are extracted and compared with the reference signature using a hamming distance correlation to estimate the relative misalignment between the audio and video streams. Finally, the calculated delays are recognized to correct the relative temporal misalignment between the audio and video streams. The synchronization efficiency is high for both audio and video streams. It is applicable for multimedia and networking technology.

Alka Jindal et al. [1] presented the overview of the various lip synchronization techniques in a systematic manner. First, speech assisted frame rate conversion approach is planned to extract information from the speech signal and apply image process to the mouth region to attain lip synchronization. This method is extremely helpful for video telephony and video conferencing applications. It is employed in the Meeting Transcriptions, Biometric Authentication and Pervasive Computing. It used in the dubbed foreign films and the cartoon animations.

Laszola Boszormengi et al. [10] presented the Audio Align synchronization of A/V streams based on audio data. It aims to modify the manual synchronization method. This approach presents code to align or synchronize multiple audio and video recordings overlap within the corresponding audio streams that eliminate the necessity of costly skilled hardware used in multitrack recordings. It's conjointly capable of synchronizing YouTube clips recorded at events like concerts and simply permits individuals to make long running continuous multicamera footage out of these clips.

Luca Lombardi et al. [13] discussed the automatic lip reading approaches. In this method, the automatic lip reading approach by using Active Appearance Model (AAM) and Hidden Markov Model (HMM). The AAM is used for detection of the visual features and the HMM is used for lip recognition. The visual features are extracted from the image sequences by the AAM and send to classifier where extracted features are compared with stored features in datasets to produce the final recognition result by HMM and for an improved lip reading. The AAM approach is more consistent with detection of non-speech section involving complex lip movements. The AAM visual feature extraction and HMM recognition model are analyzed sufficiently and appropriately.

Fumei Liu et al. [7] propounded the lip reading technology for speech recognition system. This approach is based on the lip reading computer technology and integration of speech recognition technology. This method focused on location of the lip area, visual feature extraction and mouth shape classification. This method achieved satisfactory results on small and isolated

vocabulary. It is used in the hearing-impaired people perform almost perfect in the lip reading and understanding.

Anitha Sheela et al. [2] described the lip contour extraction using fuzzy clustering with elliptical shape information and active contour model. This method is the combination of both image and model based methods to improve the performance of the lip segmentation and lip contour extraction. This method provides accurate lip contours and accuracy of the visual speech recognition rate is improved. The lip contour extraction is used in visual speech recognition system and automatic speech recognition system in noisy environments. Speech recognition using visual features could be very helpful in lip reading, facial expression analysis and human machine interface applications.

Namrata Dave [14] presented the lip localization and viseme extraction method to segment lip region from image or video. At first detect the face region and lip region from input image or video frame in order to synchronize lip movements with input image. The objective of this method is to implement a system for synchronizing lips with speech. The visual features are extracted from video frame or image using YC_bC_r. The proposed algorithm works well in normal lighting conditions and natural facial images of male and female. This method provides high-quality accuracy. It is suitable for real time application and offline applications. Lip localization is used in lip reading, lip synchronization, visual speech recognition and facial animations.

2. PROPOSED METHODOLOGY

The proposed framework is automatically measuring and maintaining the perfect synchronization between audio and video using audio-visual spreading codes. The audio-visual spreading code conveys the relative timing of audio and video signals. The proposed framework is designed to be robust to modification of audio-video signals. The proposed method is guaranteeing that the audio and video streams are perfectly synced after processing. The Fig.1 shows the proposed frame work for audio and video synchronization based on audio-visual spreading codes.

During transmission, the audio and video signals are processed individually. The audio spreading code is extracted from the spectrograph of the input audio which is broken up into chunks. The spectrogram is the visual way of representing the spectrum of sounds and it can be used to display the spoken word phonetically. It is also called spectral waterfalls, voice grams or voiceprints. The video spreading code is computed by the absolute difference the consecutive video frames where the input video is broken up into video frames and finally to attain a coarse absolute difference image.

The audio-visual spreading codes or A/V sync spreading code based on content and don't change excessively. It is an authentication mechanism and formed by taking hash of the original audio-video streams. The robust hash filters the little changes in the signal processing and reduces the audio-visual spreading code sizes. It is based on the difference between the successive audio and video frames.

Within the communication network, the audio-video streams encounter different signal processing namely audio compression, video compression, format conversion, audio down sampling, video down sampling etc. and their relative temporal alignment between audio and video signals may be altered.

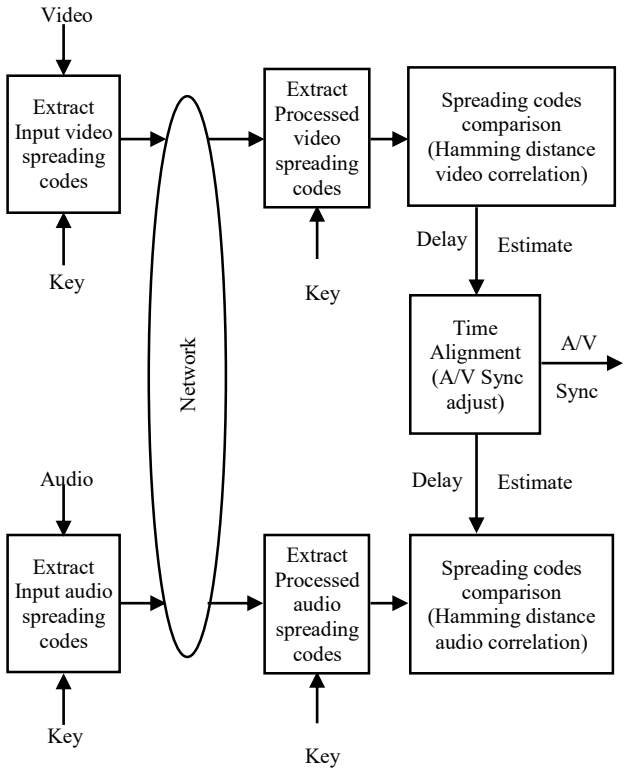

Fig.1. Audio and Video sync using audio-visual spreading codes

At the detection, the processed audio and video spreading codes are extracted from the processed audio and video streams. During synchronization, the processed audio and video spreading codes are compared with the corresponding input audio and video signatures using Hamming distance correlation. The output of the Hamming distance is used to estimate the temporal misalignment between the audio-visual streams. Finally the measured delays are used to correct the relative misalignment between audio-visual streams.

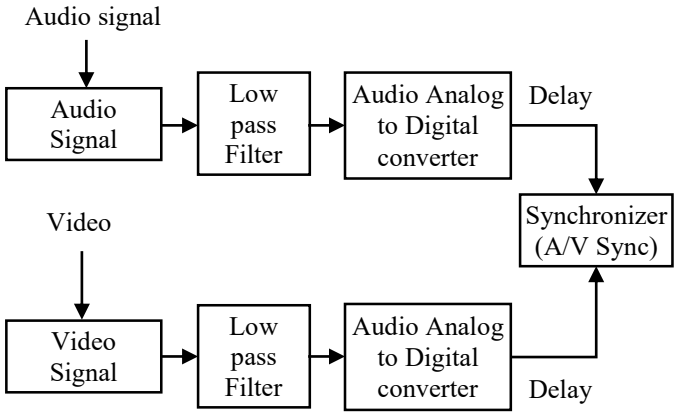

Fig.2. Audio - Video Synchronizer

The test content used for the performance assessment of the system consisted of A/V clips of a variety of content types such as scripted dramas; animation program, music concert, news programs, sports and lives music. The input audio and video is taken from the recorded dataset for the audio and video synchronization. The input video is divided into frames.

A low-pass filter (LPF) is a filter that passes low frequency signals and attenuates high frequency signals by the cutoff frequency. It prevents the high pitches and removes the short-term fluctuations in the audio - video signals. It also produces the smoother form of a signal.

An analog-to-digital converter (ADC) converts an input analog voltage or current to a digital magnitude of the voltage or current. It converts a continuous time and continuous amplitude analog signal to a discrete time and a discrete amplitude signal.

The delay line produces a specific delay in the audio and video signal transmission path. The Synchronizer is a variable audio delay used to correct and maintain the audio and video synchronization or timing.

3. RESULTS AND DISCUSSION

The proposed audio and video synchronization methodology discusses the audio-visual content, frame conversion, audio and video signature extraction, hamming distance measurement, the audio-video synchronization and its performance measures.

3.1 A/V CONTENT

The test content used for the performance assessment of the system consisted of 5s A/V clips of a variety of content types such as scripted dramas, talk programs, sports and live music. The Fig.3 shows the input audio and video is taken from the recorded dataset for the audio and video synchronization. The frame number is given as the input for synchronization purpose.

Fig.3. Input Audio and Video

3.2 VIDEO FRAME

The input video is divided into frames for generating the video spreading codes. The input video is divided into 30fps. There are totally 74 frame are available in the input video frame. The Fig.4 shows the frame conversion of the input video.

3.3 AUDIO SPREADING CODES GENERATION

The audio signature is primarily based on the representation of the spectrograph. The Fig.5 shows the audio spreading code generation using spectrograph.

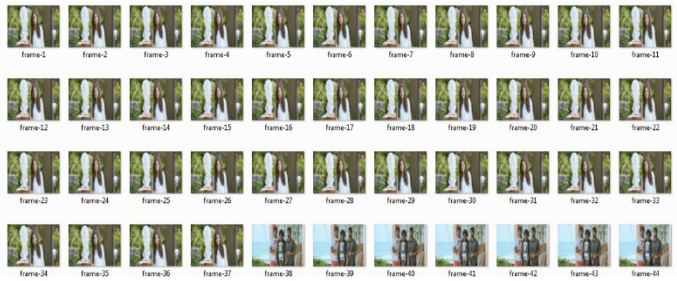

Fig.4. Frame conversion

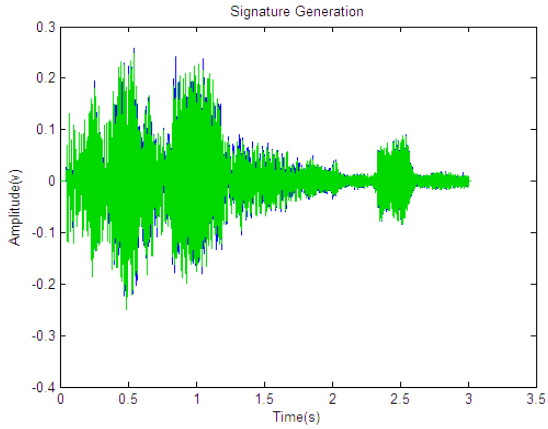

Fig.5. Audio spreading codes extraction using spectrogram

3.4 VIDEO SPREADING CODES GENERATION

The video signature is based on the illustration of the distinction image between two consecutive frames. The Fig.6 shows the video signature generation.

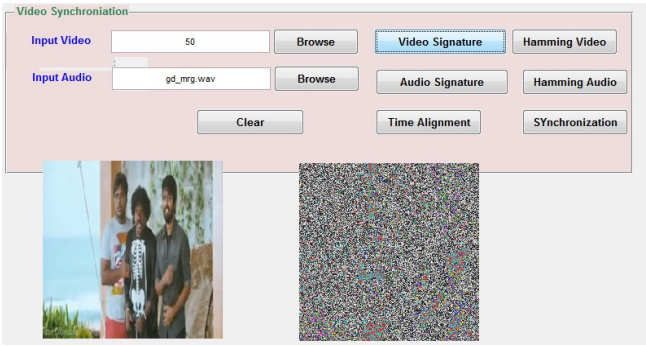

Fig.6. Video Spreading Codes Generation

3.5 HAMMING VIDEO AND HAMMING AUDIO

The Hamming distance correlation is used to calculate the temporal misalignment between audio-visual streams and the quality of the audio-video synchronization can be measured. The Fig.7 shows the hamming video and Fig.8 shows the hamming audio.

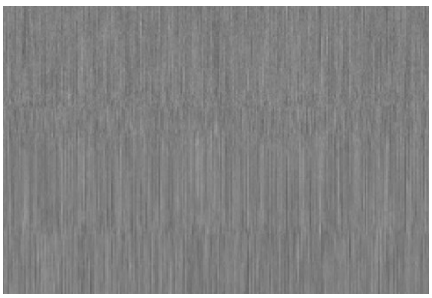

Fig.7. Hamming Video

Fig.8. Hamming Audio

Table.1. Hamming distance for video and audio

Input Image	Encode Image	Hamming Video	Hamming Audio
1010	1010	0	1
1010	1010	0	1
1010	1010	0	1
1010	1010	1	1
1110	1110	0	1
1010	1010	0	1
1010	1010	0	1
1110	1110	0	1
1110	1101	2	1
1010	1010	0	1
1110	1110	0	1
1110	1010	1	1
1001	1001	0	1
1001	1001	0	1
1010	1100	2	1
1001	1001	0	1
1001	1101	1	1
1110	1110	0	1
1110	1110	0	1
1011	1001	1	1
1000	1000	0	1
1000	1000	0	1
1011	1010	1	1
1011	1011	0	1

From the Table.1, it is inferred that the hamming distance for the video and audio. Hamming code is a set of error-correction codes that can be used to detect and correct bit errors. It is used to find the misalignment between the audio and video streams. The output of hamming distance based correlator is estimated delays are relative to each of the corresponding input spreading codes.

3.6 TIME ALIGNMENT

The estimated relative misalignment is used to achieve the same alignment between the audio and video streams that was present before processing. It aligns the audio and video frame in an appropriate manner. The Fig.9 shows the relative time alignment between the audio and video stream. It decodes the corresponding video frame that is given as input in Fig.10 with proper time alignment between the input and processed video frames. The Fig.10 shows the decoded input video frame.

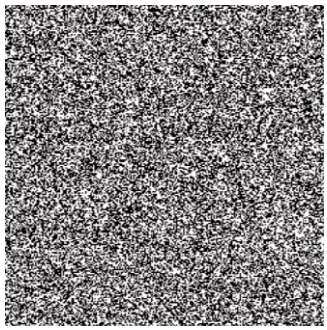

Fig.9. Audio-video stream time alignment

Fig.10. Decoded video frame

3.7 AUDIO - VIDEO SYNCHRONIZATION

The Fig.11 shows the audio and video synchronization using signature. The A/V sync using spreading code provides perfect synchronization between the corresponding audio and video streams. Finally, the reliability measures along with the estimated delays can be used to detect or correct the relative misalignment between the A/V streams. It can detect and maintain the audio - video sync accuracy.

3.8 AV SIGNALS

The test content used for the performance assessment of the system consisted of 5s A/V clips. The Fig.12 shows the input audio is taken from the recorded dataset for the audio and video synchronization and every 10msec for audio.

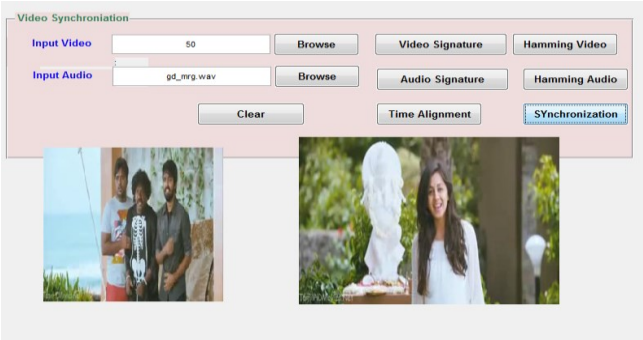

Fig.11. Audio - Video synchronization

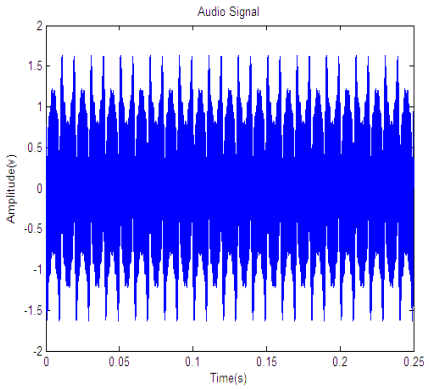

Fig.12. Input audio signal

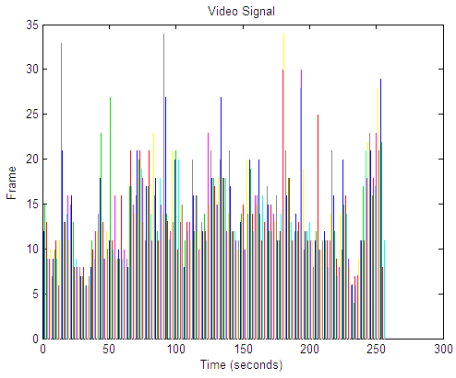

Fig.13. Input video signal

The Fig.13 shows the input video is taken from the recorded dataset for the audio and video synchronization. Every video frame plays 3ms The input video is divided into 50fps. There are totally 74 frame are available in the input video.

3.9 NOISE REMOVAL

The Fig.14 shows the audio low pass filtered output. The filter allows the frequencies below the cut off frequency but the high frequencies in the input signal are attenuated.

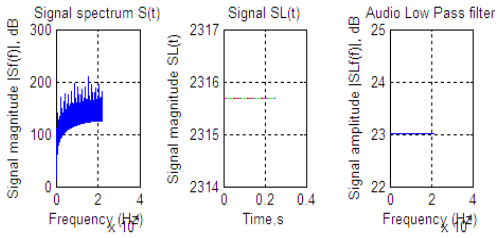

Fig.14. Audio Low passes filter output

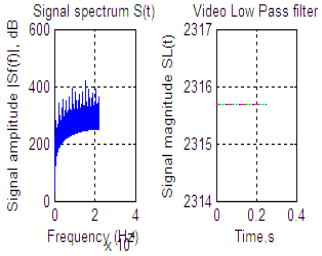

Fig.15. Video Low passes filter output

The Fig.15 shows the video low pass filtered output. The filter allows the frequencies below the cut off frequency but the high frequencies in the input signal are attenuated.

3.10 ANALOG TO DIGITAL CONVERSION

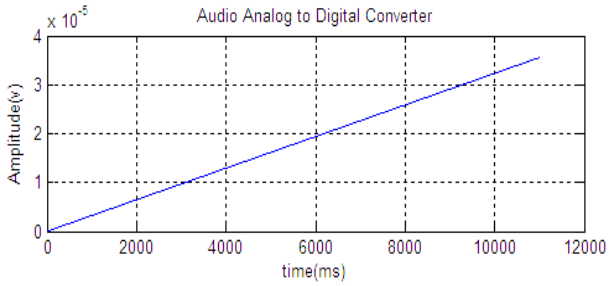

Fig. 16. Audio Analog to Digital converter

The Fig.16 shows the audio analog to digital converter output. ADC converts the analog audio signal into digital representing the amplitude of the voltage. The Fig.17 shows the video analog to digital converter output. ADC converts the analog video signal into digital representing the amplitude of the voltage.

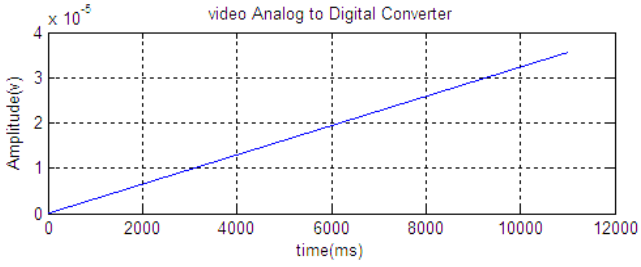

Fig.17. Video Analog to Digital converter

3.11 A/V SYNCHRONIZATION

The objective of the synchronization is to line up both the audio and video signals that are processed individually. The Fig.18 shows the A/V signal synchronization. It aligns both audio

and video signals. The synchronization is guaranteeing that the audio and video streams matched after processing.

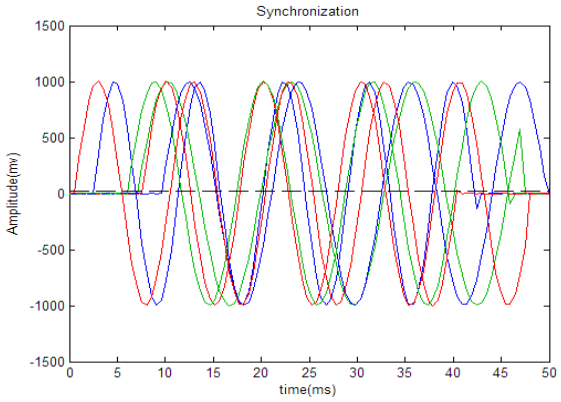

Fig.18. A/V signal Synchronization

4. PERFORMANCE ANALYSIS OF AUDIO-VIDEO SYNCHRONIZATION

4.1 SYNCHRONIZATION EFFICIENCY

The Synchronization efficiency is the process of establishing consistency among A/V content from a source to a processed or target A/V content storage in percentage. If the synchronization efficiency is high, then the audio and video are perfectly synchronized. Otherwise the audio and video synchronization will be poor. It is expressed as

$$\eta = P_{out}/P_{in} \quad (1)$$

where,

η = synchronization efficiency,

P_{out} = synchronized A/V stream and

P_{in} = unsynchronized A/V stream.

4.2 AUDIO-VIDEO TIME DRIFT

The Time drift is defined as the amount of time the audio departs from perfect synchronization with the video where a positive number indicates the audio leads the video while the negative number indicates the audio lags the video. The Audio - Video time drift can be represented as

$$t_{A/V} = t_r - t_p \quad (2)$$

where,

$t_{A/V}$ = A/V Time drift,

t_r = Source time and

t_p = deviation time.

4.3 AUDIO TO VIDEO DELAY

The Audio to Video Delay is referred as the relative time alignment delay between the audio and video streams. The amount of the visual data is much bigger than audio data and the delays which are generated to the audio and video streams are typically unequal. The solution to audio to video delay is to add fixed delays to match the video delay. Finally, the estimated delays are used to correct the relative misalignment between the audio and video streams. The A/V delay is given as,

$$D = t \pm t_0 \quad (3)$$

where,

D = audio - video delay,

t = audio/video time and

t_0 = extra audio/video time.

Table.2. Performance analysis for audio and video synchronization

Parameters	A/V Content
Synchronization Efficiency	99 %
Audio and video sync time drift	16ms
Audio to video delay	16ms

From the Table.2, it is inferred that the audio and video synchronization parameter. The synchronization efficiency is very high. The audio - video sync time drift and audio to video delay are very less.

5. CONCLUSIONS

Thus the audio and video synchronization using spreading code technique was implemented and their performances were analyzed sufficiently and appropriately. The proposed system would automatically estimate and preserve the perfect synchronization between the audio and video streams and it would maintain the perceptual quality of audio and video. This method provides high-quality accuracy and low computational complexity. It can detect and maintain the audio - video sync accuracy. The audio - video sync time drift and audio to video delay are very less. The audio-video synchronizer provides a guarantee and perceptual quality of audio-video signals synchronization without generating any artifacts. The experimental test results were shown the guarantee and quite simple process applicable for the real world multimedia application and offline applications. This method is suitable for film production, content distribution network, communication network, mobile devices and traditional broadcast networks. Hence the audio-video synchronizer requirements are standardized very accurately. In future work, the proposed framework will be developed with modified structures to provide vast improvement in real time application. We have planned to use results obtained to use for live music concerts and meetings as future enhancement. Improvement of future work may also include improvement of the signature matching and thus increase the synchronization rate.

REFERENCES

[1] Alka Jindal and Sucharu Aggarwal, “Comprehensive Overview of Various Lip Synchronization Techniques”, *Proceedings of International Symposium on Biometrics and Security Technologies*, pp. 23-29, 2008.

[2] K. Anitha Sheela, Balakrishna Gudla, Srinivasa Rao Chalamala and B. Yegnanarayana, “Improved Lip Contour Extraction for Visual Speech Recognition”, *Proceedings of International Conference on Consumer Electronics*, pp. 459-462, 2015.

[3] N.J. Bryan, G.J. Mysore and P. Smaragdis, "Clustering and Synchronizing Multicamera Video via Landmark Cross-Correlation", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2389-2392, 2012.

[4] Claus Bauer, Kent Terry and Regunathan Radhakrishnan, "Audio and Video Signature for Synchronization", *Proceedings of IEEE International Conference on Multimedia and Exposition Community*, pp. 1549-1552, 2008.

[5] N. Dave and N.M. Patel. "Phoneme and Viseme based Approach for Lip Synchronization", *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol. 7, No. 3, pp. 385-394, 2014.

[6] Dennis Laurijssen and Erik Verreycken, "Low-Cost Synchronization of High speed Audio and Video Recording in Bio-Acoustic", *International Journal of Experimental Biology*, Vol. 8, No. 2, pp. 1-22, 2017.

[7] Dragan Sekulovski, Hans Weda, Mauro Barbieri and Prarthana Shrestha, "Synchronization of Multiple Camera Videos using Audio-Visual Features", *IEEE Transactions on Multimedia*, Vol. 12, No. 1, pp. 79-92, 2010.

[8] Fumei Liu, Wenliang and Zeliang Zhang, "Review of the Visual Feature Extraction Research", *Proceedings of IEEE 5th International Conference on Software Engineering and Service Science*, pp. 449-452, 2014.

[9] Josef Chalaupka and Nguyen Thein Chuong, "Visual Feature Extraction for Isolated Word Visual only Speech Recognition of Vietnamese", *Proceedings of IEEE 36th International Conference on Telecommunication and Signal Processing*, pp. 459-463, 2013.

[10] K. Kumar, V. Libal, E. Marcheret, J. Navratil, G. Potamianos and G. Ramaswamy, "Audio-Visual Speech Synchronization Detection using a Bimodal Linear Prediction Model", *Proceedings of Computer Vision and Pattern Recognition Workshops*, pp. 54-57, 2009.

[11] Laszlo Boszormenyi, Mario Guggenberger and Mathias Lux, "Audio Align-Synchronization of A/V Streams based on Audio Data", *Proceedings of IEEE International Symposium on Multimedia*, pp. 382-383, 2012.

[12] Y. Liu and Y. Sato, "Recovering Audio-to-Video Synchronization by Audiovisual Correlation Analysis", *Proceedings of 19th International Conference on Pattern Recognition*, pp. 1-2, 2008.

[13] C. Lu and M. Mandal, "An Efficient Technique for Motion-based View-Variant Video Sequences Synchronization", *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 1-6, 2011.

[14] Luca Lombardi and Waqqas ur Rehman Butt, "A Survey of Automatic Lip Reading Approaches", *Proceedings of IEEE 8th International Conference Digital Information Management*, pp. 299-302, 2013.

[15] Namrata Dave, "A Lip Localization based Visual Feature Extraction Methods", *An International Journal on Electrical and Computer Engineering*, Vol. 4, No. 4, pp. 23-28, 2015.

[16] P. Shrstha, M. Barbieri and H. Weda, "Synchronization of Multi-Camera Video Recordings based on Audio", *Proceedings of 15th International Conference on Multimedia*, pp. 545-548, 2007.