Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-11, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

A Survey – Audio and Video Synchronization

A.Thenmozhi[1] & P.Kannan[2]
[1]PG Scholar, ECE Department, PET Engineering College, Anna University, India.
[2]Professor and HOD, ECE Department, PET Engineering College, Anna University, India.

Abstract— *The audio and video Synchronization is extremely necessary. The synchronization loss between image and sound continues to disturb observers and irritate telecasters. The demand is to assure synchronization without adjusting content at the same time as still retaining price low. The objective of the synchronization is to line up both the audio and video signals that are processed individually. This paper describes the summary of the numerous audio-video synchronization techniques and its performance evaluation in an efficient manner. The various techniques namely audio and video signatures for synchronization, lip localization method, lip contour extraction method, lip reading approach, speech recognition, and audio-visual recognition and methods.*

Keywords— *Lip Localization, Lip Contour Extraction, Lip Reading, Speech Recognition, Audio-visual Recognition*

I. INTRODUCTION

The Audio-video synchronization or lip synchronization defines the temporal difference of the sound and vision elements during production, combining, transmitting, reproduction and reception process. The Synchronism is defined as the process of exactly coordinating or adjusting two or more activities at the same time or rate. Humans are very sensitive to discover the distinction between visual illustration and corresponding audio. The Audio-Visual delay or Transcription error is the voice isn't synchronic to the lip movements. The audio and video data are handled individually and therefore the adjusting between audio and visual streams is very dependent on the implementation of the software package and hardware environment. The synthesis and coordination between audio and video signals is very essential in multimedia applications due to severe timing restraints. The synchronization is guaranteeing that the audio and video streams matched after processing.

II LITERATURE SURVEY

Claus Bauer et al (2008) [3] suggested the audio and video designation for synchronization. The audio and video signature or mark extracted from audio and video signals for necessarily maintaining coordination between the audio and
video signals. The audio signature is extracted from spectral waterfalls or voice prints and then the video signature is also
made on absolute distinction image using two consecutive video frames. During transmission, the audio and the video streams are recorded with the aid of joining audio and video mark into a synchronization mark. At reception, the equivalent mark are derived and correlated with the stored mark using a Manhattan distance correlation to measure the improper alignment amid the audio-visual streams. Finally, the calculated delays are recognized to correct the relative timing misalignment between the audio-visual streams. The synchronism efficiency is high for both audio and video signals. It is applicable for multimedia and networking technology.

Alka Jindal et al (2008) [1] presented the review of the different lip syncing techniques in a precise manner. First, speech aided frame frequency reduction approach is planned to obtain information from the speech signal and apply image process to the mouth region in order to attain lip integration. This method is extremely helpful for videophone and visual collaboration applications. Then the detection of soliloquy methodology in video treasury identifies discourse in video frames. It is employed in the realistic authentication. In audio-visual animation approach, speech is employed as an input to achieve miming facial animation. This method may be used in the re-recording foreign films, cartoon and movie animations. Lastly the automated lip synching for artificial faces is employed for the speech analysis and the speech reading techniques to significantly determine lip shapes for a given speech sequences. These approaches are essentially centred on multimedia presentation and estimate the synchronism for distinct multimedia applications. The artificial face methodology are often effectively used to make naturalistic speaking person lip and conjointly used for teaching related to recorded audio and simple approach.

Laszola Boszormengi et al (2012) [6] presented the Audio Align synchronicity of Audio-Video signals supported on audio stream. It targets to

modify the hand operated synchrony method. This technique presents code to adjust or coordinate multiple audio and video registries from diverse sources overlay within the analogous audio streams that eliminates the necessity of costly skilled hardware and used for multitrack recordings. It takes variety of files as input, evaluates their audio content and detects equivalent data that are detected overlays as coordination points. It provides high efficiency and speed. The mechanism will be applied to several real-world things like synchronistic multitrack sound recordings generated with low cost shopper devices or synchronism multicamera video shots. It's conjointly capable of integrating the web camera clips recorded at events and simply permits individuals to make extended multicamera footage.

Josef Chalaupka et al (2013) [5] suggested the Visual element extraction for voice Recognition of Vietnamese. This method conferred on visual feature extraction for handling with Vietnamese dialect. The visual features will be improved for automatic lip-reading 1-Stage Fisher's Linear Discriminant Analysis visible fore end and Hierarchical Fisher's Linear Discriminant Analysis visible fore end. It uses a strong and skilled methodology known as Constrained by a statistical Local Models to extract face and facial feature boundaries. With the path of Constrained Statistical Local Models the contour-based feature like lip and face boundaries, active appearance-based feature and classical-based feature will be totally extracted. The Hierarchical Fisher's Linear Discriminant Analysis visual forepart outperforms the Stage-1 Fisher's Linear Discriminant Analysis visual forepart provides maximum and moderate recognition efficiency. It greatly improves the voice identification performance particularly in noise condition.

Luca Lombardi et al (2013) [7] discussed the automatic lip reading approaches. In this methodology, the automated speech reading approach by using Active shape Model or smart snakes and statistical Markov Model. The smart snake is employed for detection of the visual elements and therefore statistical Markov Model is employed for speech recognition. The visual features are extracted from the image sequences by the smart snake and deliver to classifier wherever derived features are correlating with reference features in datasets to provide the ultimate perception result by statistical Markov Model for an improved lip reading. The smart snake approach is more persistent with detection of silence section involving complex and advanced lip reading. The smart snakes visual feature extraction and statistical Markov Model are analyzed adequately and befittingly.

Fumei Liu et al (2014) [4] propounded the lip reading technology for speech recognition system. This approach is based on the speech reading computer mechanism and synthesis of voice identification technology. This technique targeted on position of the lip area, visual component abstraction and mouth model categorization. The lip region is primarily relied on the conventional image processing method and statistical methodology. The derived visual components depend on the conventional methods are appearance based method, model based method, integration of the appearance and model based method and motion based approach. The mouth shape classification is completed by visual speech unit. This method achieved satisfying results on small and isolated vocabulary. It can be employed for the deaf people achieves excellent in the voice reading and recognizing.

Anitha Sheela et al (2015) [2] described the lip border or outline extraction using fuzzy or soft clustering with elliptical shape information and snake model. This technique is the merging the pair of image and model based strategies to enhance the performance of the lip localization and lip border extraction. Visual voice elements are depicted by shape information of the lips and grayscale digital image of the lip area. This methodology provides correct lip border extraction, high efficiency and the visual speech to text rate is increased. The lip border extraction is employed in audio-visual identification system and automatic voice identification system specifically in noisy environments. The speech recognition using visual features may be terribly useful in the voice reading, facial features analysis and man-machine interface applications.

Namrata Dave (2015) [8] 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 given 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. A phoneme is the smallest unit of sounds and viseme is a visual speech element or generic facial image used to describe particular sound. 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

III. CONCLUSIONS

In this paper, the focus is to implement an audio and video synchronization for the automatic speech recognition system and audio-visual interaction in multimedia. The conclusion of this paper provides the lip localization based on lip contour extraction method. It can detect and maintain lip-sync accuracy. It also provides high recognition accuracy and the synchronization efficiency is high for both audio and video streams. It is suitable for real-time and non-real-time file-based systems including web-based media. The system supports other application such as content verification, content ID, and metadata distribution. Computational complexity is suitable for both professional and consumer applications. With the advent of interactive multimedia system and network automation, distinct multimedia services like content on demand services, visual collaboration and distance education, E-learning are in huge demand. In multimedia system applications, audio-visual streams are saved, broadcasted and prompted. Throughout exposition time, the timing relation between audio and video have conserved and offered the most effective perceptual quality.

ACKNOWLEDGMENT

At first, I thank Lord Almighty to give knowledge to complete the survey. I would like to thank my colleagues, family and friends who encouraged and helped us in preparing this literature survey.

REFERENCES

[1] Alka Jindal, Sucharu Aggarwal Lee, (2008) "Comprehensive overview of various lip synchronization techniques", International Symposium on Biometrics and Security technologies.

[2] Anitha Sheela.k, Balakrishna Gudla, Srinivasa Rao Chalamala, Yegnanarayana.B, (2015) "Improved lip contour extraction for visual speech recognition", IEEE International transaction on Consumer Electronics, pp 459-462.

[3] Claus Bauer, Kent Terry, Regunathan Radhakrishnan, (2008) "Audio and video signature for synchronization", IEEE International conference on Multimedia and Exposition Community (ICME), pp 1549-1552.

[4] Fumei Liu, Wenliang, Zeliang Zhang, (2014) "Review of the visual feature extraction research" IEEE 5th International Conference on software Engineering and Service Science, 2014, pp 449-452.

[5] Josef Chalaupka, Nguyen Thein Chuong, (2013) "Visual feature extraction for isolated word visual only speech recognition of Vietnamese", 36th International conference on Telecommunication and signal processing (TSP), pp 459-463.

[6] Laszlo Boszormenyi, Mario Guggenberger, Mathias Lux, (2012) "Audio Align-synchronization of A/V streams based on audio data" IEEE International journal on Multimedia, pp 382-383.

[7] Luca Lombardi, Waqqas ur Rehman Butt, (2013) "A survey of automatic lip reading approaches" IEEE 8th International Conference Digital Information Management (ICDIM), pp 299-302.

[8] Namrata Dave, (2015) "A lip localization basedvisual feature extraction methods" An International journal on Electrical and computer Engineering, Vol. 4, No. 4.