Object Tracking from Audio and Video data using Linear Prediction method

Thesis submitted in partial fulfilment
of the requirements for the award of the degree of
Master of Technology (dual)
In
Communication and Signal Processing
by

Nadakudity Sai Sita Anusha
(710ec4135)

Under the supervision of

Prof. Lakshi Prosad Roy

Department of Electronics & Communication Engineering

NATIONAL INSTITUTE OF TECHNOLOGY, ROURKELA

राष्ट्रीय प्रौद्योगिकी संस्थान,राउरके ला

May ,2015

Declaration

I hereby declare that

1) The work presented in this project is original and has been done by myself under the guidance of my supervisor.

2) The work has not been submitted to any other Institute for any degree or diploma.

3) The data used in this work is taken from only free sources and its credit has been cited in references.

4) The materials (data, theoretical analysis, and text) used for this work has been given credit by citing them in the text of the thesis and their details in the references.

5) I have followed the thesis guidelines provided by the Institute

N. SAI SITA ANUSHA

Rourkela, May 2015

Department of Electronics and Communication Engineering
National Institute of Technology Rourkela
ROURKELA-769 008, ODISHA, INDIA

CERTIFICATE

This is to certify that the Thesis entitled "**Object Tracking from Audio and Video data using Linear Prediction method**" submitted by NADAKUDITY SAI SITA ANUSHA bearing roll no. 710EC4135 in partial fulfilment of the requirements for the award of Mtech(dual) in Electronics and Communication Engineering with specialization in "Communication and Signal Processing "during session 2010-2015 at National Institute of Technology, Rourkela is an authentic work carried out by her under my supervision and guidance. To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University / Institute for the award of any Degree or Diploma.

Dr. Lakshi Prosad Roy

Assistant Professor

Department of ECE

National Institute of Technology

Rourkela-769008

Acknowledgements

First and foremost, I am very much indebted to my supervisor, Prof.Lakshi Prosad Roy , for his excellent guidance and valuable suggestions and discussions leading to fruitful work. From finding a problem to solving, his careful observation helped me a lot in my dissertation work and of course in due time. There are many people who are associated with this project directly or indirectly whose help, timely suggestions helped a lot in the successful completion of this project.

I am very much indebted to Prof. Sarat Kumar Patra and Prof. Kamala Kanta Mohapatra for teaching me subjects that proved to be very helpful in my work. My special thanks go to Prof. Ajit Kumar Sahoo, Prof. Manish Okkade for contributing towards enhancing the quality of the work and eventually shaping my thesis. My wholehearted gratitude to my parents Radha Krishna and Surya Kumari and my sister Prathyusha for their love and support.

N. SAI SITA ANUSHA

Rourkela, May 2015

Abstract

Microphone arrays and video surveillance by camera are widely used for detection and tracking of a moving speaker. In this project, object tracking was planned using multimodal fusion i.e., Audio-Visual perception. Source localisation can be done by GCC-PHAT, GCC-ML for time delay estimation. These methods are based on spectral content of the speech signals that can be effected by noise and reverberation. Video tracking can be done using Kalman filter or Particle filter. Therefore Linear Prediction method is used for audio and video tracking. Linear prediction in source localisation use features related to excitation source information of speech which are less effected by noise. Hence by using this excitation source information, time delays are estimated and the results are compared with GCC PHAT method. The dataset obtained from [20] is used in video tracking a single moving object captured through stationary camera. Then for object detection, projection histogram is done followed by linear prediction for tracking and the corresponding results are compared with Kalman filter method.

Table of Contents

List of figures:

List of algorithms:

Chapter 1

Introduction

Fusion of multi-modal information is a critical issue in multimedia. The main problem arises when distinctive modalities are combined for a synergistic effect. There has been substantial work done in tracking of using video. To estimate the position of a speaker, multiple microphones (i.e., by utilizing audio) have likewise been used .According to the position of the speaker, the sound reaches one microphone before the other and hence the signals received by the microphones are displaced by τ number of samples. However, the problem of using these modalities together is relatively new. Fusion of audio and video signals for tracking can improve the overall process performance by decreasing the uncertainties and the unreliability which are yielded by both the signals. For example, in scenarios like cocktail party where people sound and look alike challenges are faced because of video occlusions, clutter, reverberations, speech pauses and sharp movements for single modality trackers. Redundant information like target position is used for localization and tracking systems. These tracking systems are usually made up of microphones arrays and cameras whose data are processed using resource consuming techniques such as beamforming ,time delay estimation ,time difference of arrival ,kalman filter ,particle filters and jointly track two alternating speaker in a room . Normally, those applications are used in analysis of meeting room where person motion is either static or restricted to small trajectories. On the other hand, detection and recognition applications use different features of the signals, e.g. audio pitches and motion changes .These systems are often composed of just one or a few microphones and a pair of cameras, and are used mostly for speech recognition. The Kalman Filter and its extended version, the Particle Filter as well as hybrid approaches using Monte Carlo Markov chains have been all used to tackle the problem.

The fusion of multiple modalities can provide complementary information and increase the accuracy of the overall decision making process. Multimodal fusion benefits comes with a certain cost and complexity in the analysis process. This is because of the different characteristics of the involved modalities, which are stated in the following statements.

- Different media are usually captured in different formats and at different rates. For example, a video can be captured at a frame rate that is different from the rate at which audio is sampled, or even two video sources can have different rates of frame.
- Therefore this asynchrony problem has to addressed in fusion The strategy of fusion is affected by the processing time of different types of media streams.

The properties of both direct sound and reflections that are arrived later helps to find the auditorium acoustics. The following theory gives a combined idea regarding the background check done on video tracking and source localisation from different papers

1.1 Related Work

Generally people can decide the direction of a sound by using their ears. The combination of signals which are slightly different that arrive at ears helps us to conclude the direction of sound intuitively. Similarly we can build a sound localisation system with a system of microphones connected to a computer. This type of situation is similar to the cocktail party problem which states that a person trying to focus on a single speaker among a group of speakers in a party among noise and lot of conversations going on .Source localisation has a wide range of applications but here in this we use it for object tracking**.**

1.2 Audio measurements (Source localisation)

Localization of source is a very difficult task faced by acoustic engineers every day. The ultimate goal of any technique is to approximately locate the object. Resolution of space is defined as the capacity to separate two sources of sound and can be expressed in centimeters. For these two sources where they still appear separately and do not merge into a single source, it actually gives the closest distance between two sources. Source localization is better if spatial localization is low. Dynamic range gives the differences of sound level in decibel between actual sources of sound and surrounding mathematical artifacts inherent to the sound source localization techniques. The source localization is better if the dynamic range is higher.The lower the frequency, the higher will be the dynamic range in beamforming algorithms

1.2.1 Time Difference of Arrival (TDOA):

Given $m_1$and m_2 are the coordinates for the two spatially separated microphones and let *s* be the location of the source. If τ is the estimated delay which can be equal to the actual delay then it can be expressed as follows

$\tau = \frac{|s-m_1|-|s-m_2|}{c}$……..(11)

where *c* is the speed of sound in air

.

1.2.2 Near-Field Acoustic Holography :

NAH is a system in which the amplifier cluster is situated very close to the origin of the sound in the nearby field. It gives extraordinary results on the whole range that is repeated. The nearby field can be depicted as the domain that is near to the sound source than perhaps a few wavelengths of the most foremost repeat. NAH1 was exhibited in the middle 1980s and industrialized in the later 10 years. It has as of right now transform into a doubtlessly comprehended technique. NAH measures sound weight by engineering a couple of enhancers in a rectangular planar display. Amplifiers are reliably separated both on a level plane and vertically. The sound weight in the plane is then back-spread to the genuine surface of the thing. The separating between the microphones chooses the half-wavelength of the recurrence that is most extreme, and the degree of the group chooses the half wavelength of the base repeat. The isolating moreover chooses the spatial determination – a coarsely isolated bunch can't correctly limit sources made out of the fine mechanics of somewhat challenge. NAH has two important central focuses:

• Spatial determination is free of repeat. It parallels the beneficiary partitioning in the perception.

• Using the Dirachlet Green limit licenses causing from the purposeful weight field to a pace field. This technique sponsorships sound drive and sound power processings for unmistakable zones or fragments. The NAH framework is an amazingly correct outlining contraption for source constrainment. Regardless, it has a couple loads:

• NAH can simply multiply stable weight to a surface which is parallel to the conscious surface. The degree of the spread plane must be undefined to the conscious plane. To cutoff a source

on a complete vehicle, the estimation plane needs to compass the complete vehicle. For applications which are stationary and applications that are temporary and repeatable, for instance, moderate engine run-ups or entrance pounds, the data can be picked up in packs. In this way, it is possible to perform NAH estimations with a 20- to 30 data of the channel getting system.

• In every practical sense, NAH can be unbalanced for higher frequencies in view of the considerable measure of data expected to finish a fair examination.

1.2.3 Beamforming

Beamforming is a strategy where the microphone set is situated in the far field. As a true principle, the far field is portrayed as being further a long way from the source than the bunch estimations or expansiveness. The distinguishing between close field and far field remains a hazy zone. In the nearby field, sound waves bear on like round or roundabout waves, while in the far field, they get the chance to be planar waves. Different amplifier setups are possible in beamforming methods. All things considered, the setup is by and large a trade off between component territory and source confinement precision. To exceed both fields, select a circuitous display with a pseudo-subjective microphone scattering. The beamforming system was at first made for submarines and natural applications. For the far field, sound waves hitting the group are planar waves. Under these conditions, it is possible to cause the intentional sound field clearly to the test thing. All beneficiary signs calculated by the beamforming display are incorporated, considering the deferral contrasting with the spread partition. The weight can be figured at whatever time before the display, allowing inciting to any kind of surface. Beamforming is now and again called "total and deferral," since it considers the relative delay of sound waves coming to assorted beneficiary positions. Beamforming obliges that all data be measured at the same time. It is regularly completed with an estimation game plan of 40 channels or more. Beamforming has the going hand in hand with good circumstances.

• Measure of the estimation group is not related to estimation group. The object for test can be greater than the group. With a mixture of half meter in estimation, it is possible to spread weight to an entire scenario. Because all the data are measured in the meantime, results can be seen minutes after data getting.

• Because of the by and large speedy getting and examination speed, beamforming lets modellers survey a couple of plans in a confined measure of time. This flexibility has some negative edges.

• The spatial determination is compared to the wavelength: where d is the partition between the source and show, D the group expansiveness, and l the wavelength. In an impeccable situation, when the radio wire is at a partition D to the source, the determination is identical to the wavelength. If the bunch is situated more removed from the structure, the determination ends up being more unpleasant. Overall, beamforming is only usable at frequencies more than 1000 Hz. Beamforming can't be used to discover sound drive and genuine source situating isn't possible.

1.3 Video Measurements (Video tracking):

Video tracking can be done by the following algorithms

- Mean shift tracking
- Background subtraction
- Particle filter tracking
- Kalman filter tracking
- Optical Flow algorithm

1.3.1 Mean shift tracking:

The clustering approach in image segmentation for joint spatial-shading space is Mean-Shift tracker(MS). Mean-shift division methodology is utilized to investigate complex multi-modular highlight space and ID of highlight groups. It is a non-parametric method. Its Region of Interest's (ROI) size and shape parameters are just free parameters on mean-movement process, i.e., the multivariate thickness kernel estimator. A 2-stage succession of irregularity safeguarding separating and mean-movement grouping is utilized for mean-movement picture. The MS calculation is introduced with an expansive number of conjectured bunch focuses arbitrarily looked over the information of the given picture. The MS calculation goes for finding of closest stationary purpose of fundamental thickness capacity of information and, in this way, its utility in recognizing the methods of the thickness. For this reason, every bunch focus is moved to the mean of the information lying inside the multidimensional ellipsoid

focused on the group focus. In the meantime, calculation develops a vector, which is characterized by the old and the new bunch focuses. This vector is called MS vector and registered iteratively until the group focuses don't change their positions. A few bunches may get converged amid the MS cycles . MS-based picture division (or inside the importance of feature outline's spatial division) is a direct expansion of the irregularity protecting smoothing calculation. After close-by modes are pruned as in the non specific highlight space investigation procedure, every pixel is connected with a huge method of the joint area thickness situated in its neighborhood,.

1.3.2 Background subtraction:

In this algorithm, the four main steps are

- Preprocessing
- Background modelling –(Recursive and Non–Recursive)
- Foreground detection
- Data validation

Background subtraction is thought to be superior to alternate methodologies as far as robustness is considered.

- Background subtraction is basic and computationally reasonable for constant frameworks, however are greatly delicate to element scene changes from lightning and incidental occasion and so forth. Consequently it is very reliant on a decent foundation upkeep model. The foundation subtraction ought to have the capacity to defeat the accompanying issues:
- Motion out of sight: Non-stationary background regions ought to be distinguished as a component of the foundation.
- Illumination changes: The background model ought to have the capacity to adjust, to continuous and quick change in light.
- Memory: Should not utilize much asset, as far as power and memory are considered.
- Shadows: Shadows cast by moving article ought to be distinguished as a feature of the background but not foreground

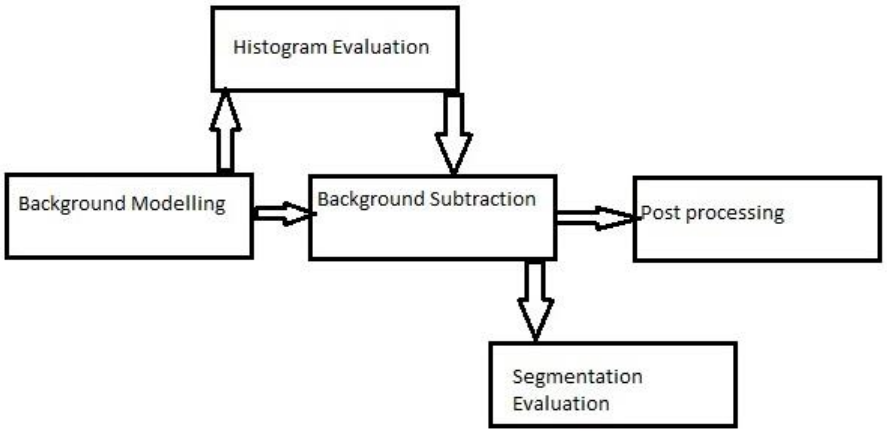


Fig 1.1 Block Diagram for background subtraction

- **Background subtraction**
- **Algorithm**
    1. Capture image containing background in the video and divide the video into frames
    2. Subtract image (= difference of motion)
    3. Remove noise using median filtering
    4. Threshold
    5. Boundary detection and highlight the blobs detected

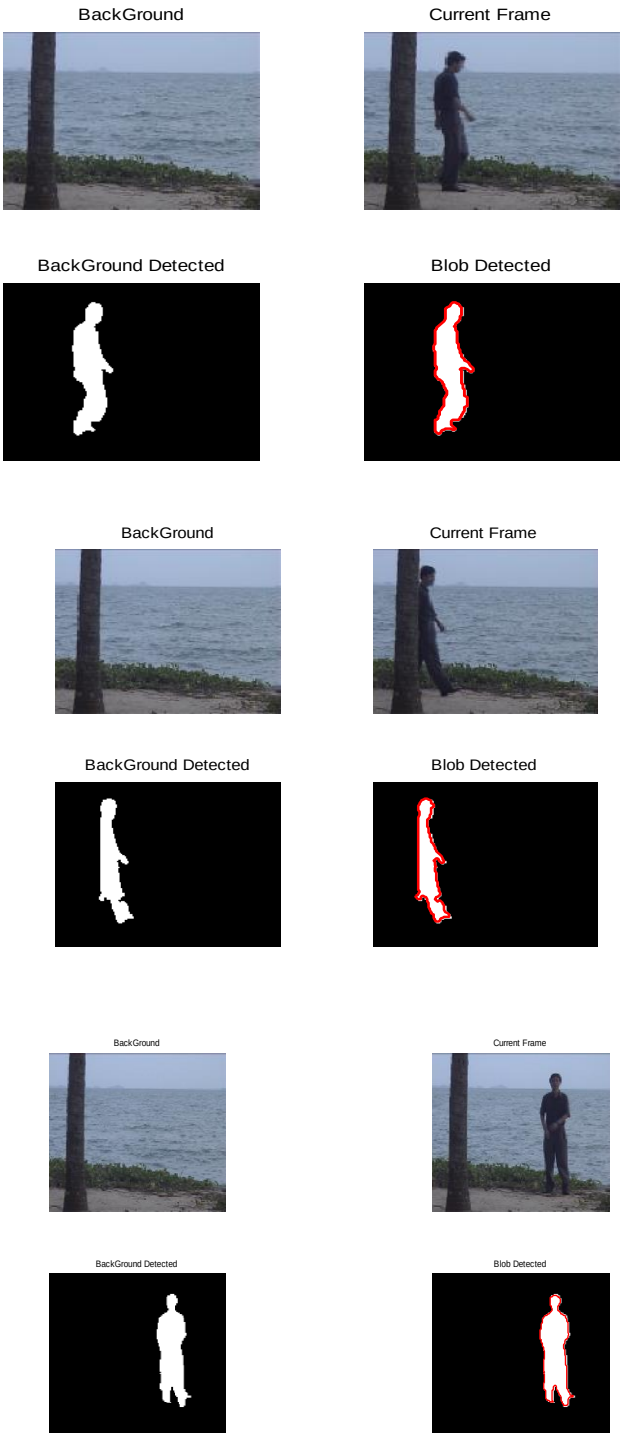


Fig 1.2 Results of Background Subtraction

1.3.3 Kalman Filter

The Kalman filter is a mathematical model that can measure the variables of a far reaching assortment of procedures. It surveys the states of direct framework. This sort of filter

works especially well before long and that is the reason it is as often as possible realized in embedded relations with control system and because we oblige an exact assessment of the strategy variables. The discrete Kalman filter is depicted by both a strategy model and an estimation mathematical articulation. The strategy model is depicted by the assumption that the current state can be related to the past state, as takes after

$x_k = \varphi_k * x_{k-1} + w_k$…….(1)

Where w_k is thought to be a discrete, white, zero-mean methodology clamor with known covariance grid, φ_k speaks to the state move framework which decides the relationship between the current state and the past one.

In this kind of situation we endeavour to track the state of a contact in light of its last known state. Here, the state vector includes a two-dimensional position conveyed in Cartesian coordinates, a two-dimensional speed and a two-dimensional accelerating. By considering a predictable speeding up, The state transition matrix can be determined from the keen kinematic scientific proclamations as follows

$s_k = s_{k-1} + v_{k-1} * t + 0.5 * a_{k-1} * \mathrm{t}^2$….(2)

$v_k = v_{k-1} + s_{k-1}$*t ……..(3).

$a_k = a_{k-1}$……..(4)

Where s_k is characterized to be the contact's position, v_k is its speed, a_k is the contact's quickening and t is the testing period. In a framework shape, the above comparisons can be composed as follows

Measurement update

$z_k = H_k x_k + v_k$……(5)

$\check{x}_k = \hat{x}_{k} + k_k(z_k - H_k * \hat{x}_{k})$……(6)

$p_k = (I - k_k * H_k)\check{p}_k$……..(7)

Time update

$\check{x}_{k+1} = \varphi_k * \check{x}_k$…….(8)

$p_{k+1} = \varphi_k * p_k * \varphi_k^{T} + Q_k$……(9)

k_k is kalman gain, x_k is state vector, z_k is measurement vector, H is relation between z_k andx_k , $\check{x}_k$ is apriori estimate , p_k is state covariance matrix,v_k is the zero white Gaussian noise. The main role of the Kalman filtering block is to assign a tracking filter to each of the measurements entering the framework

1.3.4 Optical flow algorithm

It is utilized as a part of conjunction with the Kalman filter estimation for following a contact in a surveillance scene is executed. With a specific end goal to make a calculation that has the capacity to track a contact in a scene, three distinctive, expansive scale assignment must be achieved. In the beginning, the algorithm have to take an incoming surveillance video signal and segment it into a number of frames where contacts are distinguished from the background of the scene. The following step is the following of the contact all through the video sequence. At last, the subsequent track must be prepared with a specific end goal to examine the object's motion. The algorithm developed by horn and Schunk is used for segmentation of the input video stream. The optical stream calculation approximates the development of the contact in the present casing as referenced to the past edge. By deciding the movement of objects, one can recognize the contact and the background of the scene. After watchful tuning and preparing, the yield of the segmentation process is gone to the Kalman filter calculation for further transforming

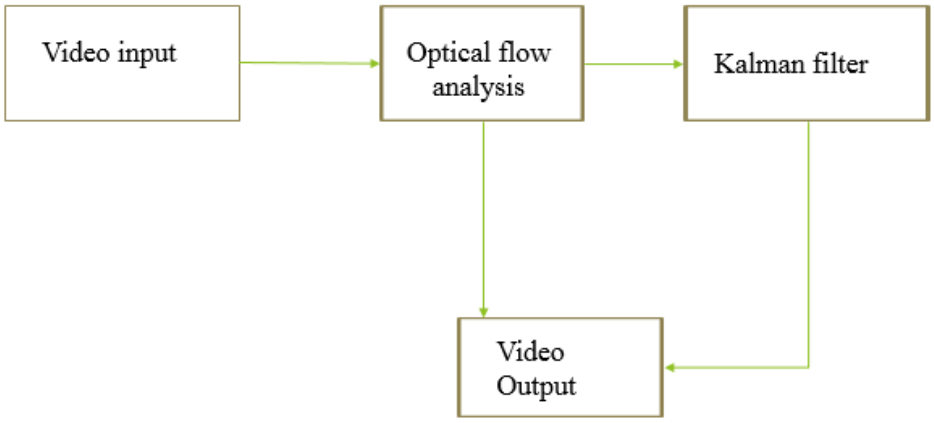


Fig 1.3 The structure of the surveillance system

The optical flow analysis is one of the important block from the above theory . The principle reason for this block is to focus the presence of conceivable contacts in the incoming video sequence and to apply a methodology to them in such way that the Kalman filter will have the capacity to track them with insignificant mistake. Optical flow is calculated by using Horn schunk algorithm .This algorithm is based on a technique computed by using constancy of brightness with a global smoothness to obtain an estimated velocity field .There are two main processes for the implementation of the HS algorithm. The first one is an estimation of partial derivatives, and the second one is a minimization of the sum of the errors uan iterative process to present the final motion vector. The following formulae are used to implement optical flow algorithm.

Equation for the rate of change of image brightness:-

$I(x,y,t) = I(x+u, y+v, t+1)$……(10)

$\varepsilon = u\, Ix + v\, Iy + It = 0$

where *u* and *v* are the horizontal and vertical motion vectors of optical flow.

Examples of the motion vector of the HS algorithm from a part of AKIYO sequence

Examples of the motion vector of the HS algorithm from a part of soccer sequence

Fig 1.4 Examples of motion vector of the HS algorithm

1.4 Audio- Visual Fusion:

Audio and video modalities capture different information of the same scene. The video signal contains the information about the appearance (color, texture, shape) and distribution of the objects in the scene, while the sounds (speech, music, noise) are only available in the audio signal. As discussed before, humans combine in a natural way the information in audio and video modalities. For example, we can easily understand the relationship between an object that is falling and the sound of the crash, we intuitively link moving lips to the presence of speech, and we know the kind of music that we will hear when we see a guitarist's arm moving.

Thus, through the joint processing of audio and video signals we can better understand a scene than when considering each modality separately. Again in the human case, we can use lip-reading to detect the speaker between two persons that move the lips, and it is possible to assign the sounds to the corresponding music instrument when we are in a concert. In general, audio-visual fusion methods integrate the information that is present in the video signal captured with one or more video-cameras and the audio signal recorded with one or more microphones. When considering the video domain, two or more video cameras allow a 3D understanding of the observed scene, where a depth can be associated to each object. Regarding the audio domain, microphone arrays are commonly used to localize and separate the sound sources in the scene. Several examples of audio-visual fusion approaches using multiple video cameras and/or microphones can be found in [7]. However, those configurations (with several audio or video sensors) need some calibration and they cannot be applied to general situations, but rather to controlled environments such as previously prepared meeting rooms. Even though many electronic devices integrate video cameras and microphones to their hardware, in most cases only one sensor of each modality is available. Notice that two microphones are present in some devices, but they are located so close that the recorded audio signals are very similar. Here we planned to consider the simplest but also the most common audio-visual configuration, where the content of the scene is captured by one microphone and one video-camera.

Audio and video signals share a temporal axis, but the resolution of this axis is different. Typically, we have much more audio samples than video frames since the sampling rate of the audio signal is much higher. Then the challenge lies in efficiently combining the information in both channels in order to extract a maximum of knowledge about the scene that we observe. The information present in audio and video modalities has a very diverse nature. Furthermore, audio and video signals have different dimensionality and temporal resolution. Thus, some assumptions need to be made in order to combine both modalities. Several works in audio-visual perception have demonstrated the correlation between audio and video modalities in the speech case. Specifically, they showed that the correspondence between the speaker lips movements and the produced sounds can be exploited by the listener to better understand speech, especially in noisy environments. This is particularly evident when we think about lip-reading, i.e. the technique to understand speech by visually interpreting the lips movement (and also the face and tongue in a minor degree). Another example that demonstrates the relationship between hearing and vision in speech perception is the McGurk effect described in [17]. This effect may be experienced when we combine a video of a person uttering one phoneme with a

soundtrack corresponding to a different phoneme. In this case, the perceived phoneme is often an intermediate phoneme different from the video and the audio ones. For example, a visual /ga/ combined with an audio /ba/ can be heard as /da/ because of the effect associated to the video interference. In this thesis, we do not restrict the analyzed audio-visual sequences to movies containing speakers. As a result, we need to base our fusion methods on an assumption that applies to all kind of audio-visual sources. The assumption that we use in this thesis is common in all applications in the joint audio-visual processing domain. It states that the presence of a sound is approximately synchronous with the movement that has generated it. Thus, related events in audio and video channels happen at approximately the same time (small lags can appear due to the different arrival times for each sensor). Several studies on audio-visual perception even support the idea that when there is a small temporal shift between events in audio and video modalities. Introduction perceive them as being synchronous. Music instruments represent some other examples of the synchrony between motion and sounds. The fingers movements are coherent with the piano sounds, the rhythm is controlled by the periodicity with which the drumsticks hit the drums, and the hands movements are also correlated with the guitar sounds. In fact, audio and video modalities are observing the same scene and they only share the temporal axis .Thus, synchrony is the only way to link both channels if we do not have previous knowledge about the characteristics of the sources in the scene. Let us now discuss the main challenges in the combined processing of audio and video signals according to the assumption of synchrony between related events. A first challenge in this domain is to distinguish the distracting motion from the motion related to the sounds, since the distracting motion can also be sporadically synchronous with the soundtrack. Another important challenge is the presence of multiple audio-visual sources: in this case some sounds are related to some movements in a part of the image while other sounds are synchronous with the motion in other locations. Furthermore, not all the sounds are generated by moving objects, i.e. a hi-fi equipment playing music represents a clear example of audio-only source. As a result, complex backgrounds composed of several moving objects and/or acoustic noises difficult the audio-visual fusion task when only one video-camera and one microphone are available. Finally, other challenges are common in most signal processing applications and are given by the video camera limitations (low quality, resolution and frame rate) and the microphone limitations (such as directivity patterns, internal noise and wind noise).

1.4.1 Audio-visual tracking algorithm:

The system uses audio and video measures of the target position on the ground plane to strengthen the single modality predictions that would be weak if taken on their own as occlusions, clutter, reverberations and speech pauses happen in the test environment. In particular, the inter microphone signal delays and the target image locations are input to single modality Bayesian filters, whose proposed likelihoods are multiplied in a Kalman Filter to give the joint AV final estimation. Despite the low complexity of the system, results show that the multi-modal tracker does not fail, tolerating video occlusion and intermittent speech (within 50 cm of accuracy) in the context of a non-meeting scenario. The system evaluation is done both on single modality than multi-modality tracking, and the performance improvement given by the AV fusion is discussed and quantified i.e., 24 % improvement on the audio tracker accuracy .Fusing audio and video signals for tracking can improve the overall process performance by decreasing the uncertainties and the unreliability which both signals individually yield. For example, in large open spaces where people sound and look alike (cocktail party scenarios) video occlusions, clutter, reverberations, speech pauses and sharp movements still represent challenges for single modality trackers. In general, localisation and tracking systems exploit redundant information e.g. target position .They are usually made up of arrays of microphones (i.e.,.16) and cameras (i.e.,4) whose data are processed using resource consuming techniques, such as beamforming and particle filters and jointly track two alternating speaker in a room .Normally, those applications are used in meeting room analysis where person motion is either static – when people are talking around a table - or restricted to small trajectories - when they are in front of a board. On the other hand, detection and recognition applications use different features of the signals, e.g. audio pitches and motion changes .These systems are often composed of just one or a few microphones and a pair of cameras, and are used mostly for speech recognition and automatic scene analysis. Bayesian inference is the bedrock of most joint tracking schemes. The Kalman Filter and its extended version, the Particle Filter as well as hybrid approaches using Monte Carlo Markov chains have been all used to tackle the problem. To the best of our knowledge this work cannot be compared directly to others work as we do not do AV localisation info fusion in meeting analysis contexts (where targets usually face cameras and microphones). The closest work we could find is the one from, but yet it is different in thatit consists of a larger sensor network (2 cameras and 16 microphones). In contrast with the prior work in this area our approach:

a) Uses only four pairs of microphones and one camera;

b) Uses low complexity audio and video measurements extraction algorithms;

c) Does not place any constraint on speaker movements.

Passive acoustic source localisation can be obtained by triangulating the time differences of arrival (TDOAs) across pairs of microphones. Knowing the asynchrony across the signals gathered at each microphones pair, together with the microphones positions, allows the triangulation algorithm to infer the target position. This approach is based on the Generalised Cross Correlation (GCC) function computation, as the asynchrony information between signals is given by the maximum energy peak of their cross correlation signal. GCC is known to be sensitive to room reverberations, however it is still effective under moderate reverberant environments. Furthermore, combining it with a position estimator like a Bayesian filter, rather than a triangulation algorithm, provides it to be valid even in more echoic environments as the filter allows for false alarms, which can be imagine like "ghost" sources generated by reverberation. With reference to the upper branch of Fig. 1, it can be seen that before extracting the delays of arrival, we detect the speech/no speech audio segments using a voice activity detector (VAD). VAD evaluates for each short segment of speech (64 ms), which the signal was subdivided in, its SNR value, and compares it to a pre-calculated threshold which is set on the basis of the static room noise power. The segment is labelled as speech if its SNR is bigger than the threshold. Speech segments are then processed using a GCC Phase Transform (PHAT) step, for the signal to be more robust to reverberations. The signal vector obtained z_a can be written as

$$z_{a=}(\tau_j(t): t = 1,2, \dots Na) \dots\dots\dots\dots\dots\dots\dots\dots (12)$$

are the TDOAs collected at the microphone pair j at each time step t and Na represent the total number of audio measurements. Note that as TDOAs are not linear in the speaker position x,

we input them to an Extended Kalman Filter. This means the EKF observation equations are approximated by a first order Taylor expansion about each latest position estimation of the filter.

As the two data feeds are recorded and processed at different rates an interpolation is carried out to ensure synchrony between the vector measurements. Moreover, since they also have to be temporally aligned from the moment in which a speaker starts to when they stop talking, we automatically computed such instants using the VAD for both the streams. The camera is calibrated offline and, with respect to the results obtained, microphones positions are manually found for the audio measurements and video tracks to be registered to the same ground plane coordinates. Once the audio and video observations have been collected and aligned, they are input to their respective Kalman trackers which estimate the target trajectory. At this point a data association step, nearest-neighbour based, chooses between the two video observation inputs available the one that is closest to the concurrent audio measure. Hence, the selected video filter likelihood is multiplied by the audio likelihood in the AV fusion node to give a global AV likelihood estimation; finally the central node state estimation$x_{av}(t/t)$ is fed back to each filter input ($x_{av}(t/t)$ = $x_a(t/t)$= $x_v(t/t)$ as the best estimate of the previous time step. In particular, under the standard KF assumptions, the data likelihoods at each time step are:

$$p(^{z_a}/_{x}) \propto \exp(-(z_a - \tau(x))^T \sum_a^{-1} (z_a - \tau(x))/_2 \text{.....(13)}$$

$$p\left(^{z_v}/_{x}\right) \propto \exp(-(z_v - \tau(x))^T \sum_v^{-1} \left(z_v - \tau(x)\right)\Big/_2 \text{.....(14)}$$

1.5 Motivation

After doing a survey on the chosen topic, it is found that audio visual tracking of object is a challenging task with object tracking from video sequence and source localization separately are also difficult tasks. Object tracking is a prolonged process because of amount of

data that is contained by audio and video. Synchronization of these two modalities is a rare topic of research. From the literature review, all the video tracking algorithms, source localization methods are studied .From these methods, Linear Prediction method is used. Among the existing methods of source localization, most of them use the features related to spectrum which corresponds to the information regarding the vocal tract system in speech most of the time. The spectral features are corrupted because of the medium, noise and the room reverberation during transmission. But features that corresponds to source of exciation information are robust to such degradations.

1.6 Objective

This thesis main goal is to improve object tracking using audio-visual modalities. It consists of 3 components namely 1)Audio (source localization) 2)Video tracking 3)Fusion(audio-visual)

Tracking of objects is foundation for many important applications. Associating objects that need to be tracked in consecutive video frames is the objective of tracking objects. This association is difficult if object are moving fast compared to frame rate. An accurate and efficient tracking capability is needed for building higher level vision-based intelligence. Applications such as video-conferencing or meeting will have scenarios in which speaker may be moving continuously. The data about this speaker that is in motion can be acquired from a speech signal and can then be fed to a system for moving camera to keep focus on speaker. This helps to provide a significant improvement for communication using audio-visual scenarios for the far-end listeners

The main goal of the work in this thesis is threefold:

1. To set up a system to capture video scenes that is used for segmentation and tracking of a object that is in motion using a stationary camera which can also be used for higher level applications and to make significant improvements in object tracking using different method (linear prediction)instead of using commonly used algorithms. Finally, the aim is to show how to do detection and tracking of moving objects based on motion in a video from a stationary camera.
2. To set up a system that contains 24 element microphone array with required hardware to capture data in a room of known dimensions and reverberation time and source localization

i.e., speaker who is moving is tracked in this project using again linear prediction other than using common methods with help of excitation source information

3. For audio video fusion of the data i.e., the results obtained from the above two steps are fused or synchronized in order to develop a more efficient algorithm for object tracking.

1.7 Conclusion:

1. From this chapter video, audio and audio-visual tracking algorithms are studied .It is inferred that among video tracking algorithms, Kalman filters have much lower computational requirements than particle filters, but are less flexible and they calculate for linear systems. But if system is nonlinear, then particle filter is used by discretizing the problem into individual particles each one is basically one possible state of the model and collection of particles can be modelled by probability distribution.

2.Beamforming methods is easy to compute and use and understand and are simple but they have less resolution. Time delay estimation methods have higher resolution and robust to noise and are more commonly used methods for source localisation.

3. Audio visual fusion increases the performance of object tracking by 24 percent .When there is a visual clutter, audio is used for tracking source. When audio is not clear, video tracking is used in this technique.

Chapter 2

Source localisation

2.1 Time Difference Of Arrival (TDOA):

For two spatially separated microphones with noise present , the time delay estimation is used in many applications such as in acoustics, microphone array processing systems, radar communication and recognition of speech , receiver cluster transforming frameworks and discourse acknowledgment.

The two microphones received signal can be shaped according to

$r1(t)=s(t)+n1(t);$ $r2(t)=s(t-D)+n2(t);$…….(15)

Where r1(t),r2(t) are the outputs of two spatially divided amplifiers, s(t) is the source signal, and n1(t) ,n2(t) are Gaussian additive noises ,D is the time delay between the two signals, t is the observed time and helps to find the difference of time between the two received signals. The noises and signals are thought to be uncorrelated which are Gaussian in nature and have zero mean .There are numerous calculations to gauge the delay of time. The accompanying are different methods for TDE

1. Cross correlation
2. Phase transform
3. Maximum likelihood estimator

2.1.1 Cross correlation:

The most commonly used method for calculating the delay of time i.e.,D is by computing the cross-correlation function of the signals that are received at ends of the two microphones and then to find the maximum peak of the output which gives the TDE. The Cross correlation method can be performed by the using the following statements

$$R_{r1r2}(\tau) = E(r1(t)r2(t-\tau)) \ldots\ldots\ldots\ldots (15)$$

$$D_{cc} = \arg\max[R_{r1r2}(\tau)] \ldots\ldots\ldots\ldots\ldots (16)$$

2.1.2 Phase Transform(PHAT):

An approach to hone the peak of the cross-correlation is by whitening the information of input signal by utilizing weighting function, which prompts the generalized cross-correlation method (GCC). The PHAT is a GCC methodology which has gotten extensive consideration because of its capacity to abstain from creating correlation function peak from spreading out. This can be communicated numerically by

$$R_{r1r2}(\tau) = \int_{-\infty}^{\infty} \varphi(f) G_{r1r2}(f) e^{-j2\pi ft} \ldots (17)$$

$$\varphi(f) = \frac{1}{G_{r1r2}(f)} \ldots\ldots\ldots\ldots\ldots\ldots (18)$$

where $G_{r1r\,2}(f)$ is the cross-spectrum of the received signal, $\varphi(f)$ is the PHAT weighting function

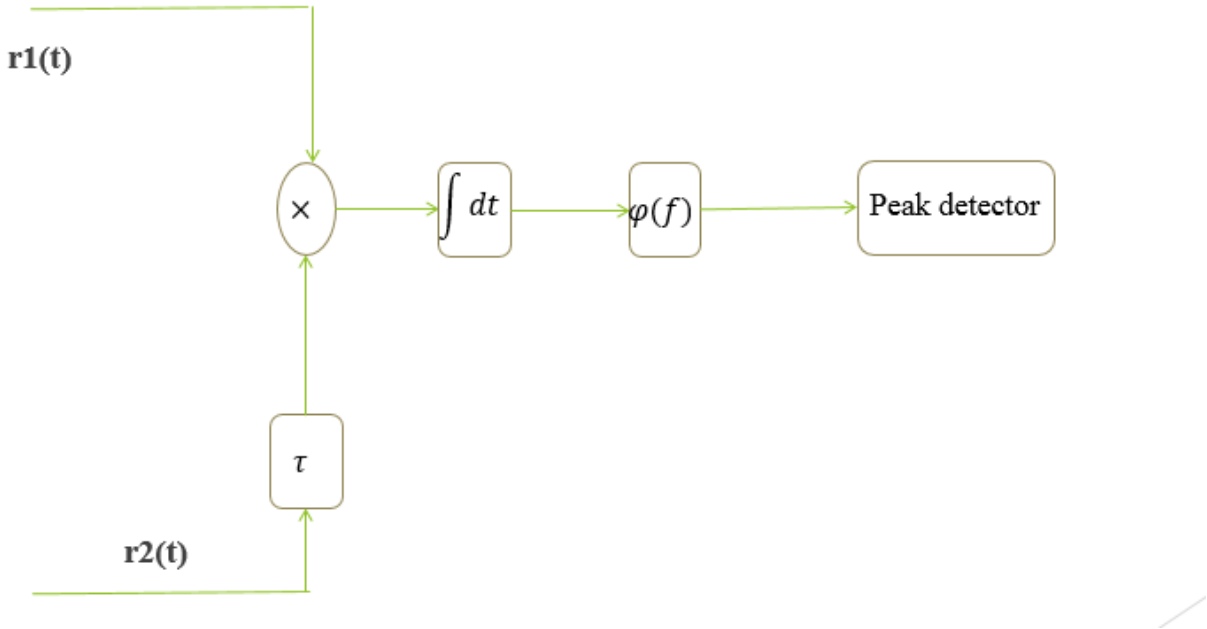


Fig 2.1 Block diagram for Generalised cross correlation- PHAT

2.1.3 Maximum likelihood (ML) method

The Maximum Likelihood (proposed by Hannan and Thomson) is another important method among all the GCC methods because it gives the extreme probability answers for TDE problems. The ML weighting capacity φ_{ML}(f) is decided to enhance the precision of the assessed defer by lessening the signs sustained into the correlator in otherworldly district where the SNR is the most minimal. The notoriety of ML estimator originates from its relative easiness of execution and its optimality under suitable conditions. Clearly, for Gaussian signal that is not correlated and have no noise and doesn't echo (i.e. no resounding), the estimator for ML of time delay is not symptotical reasonable and gainful in the purpose of repression of intervals that are long recognition .

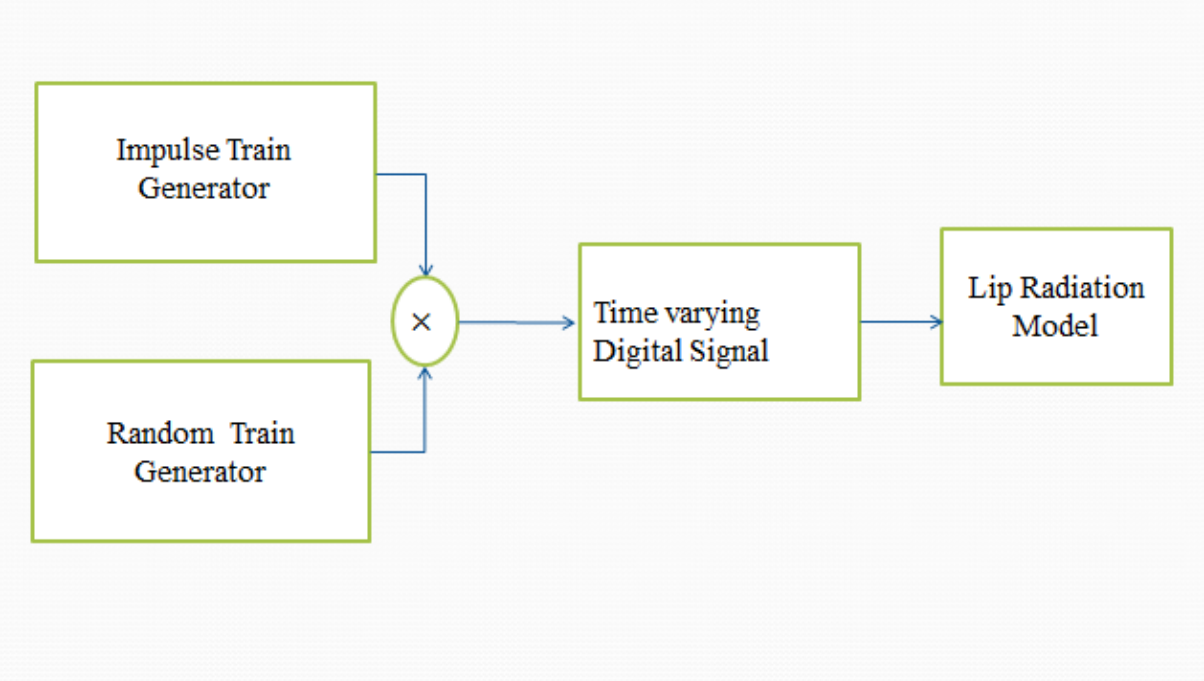


Fig 2.2 Block diagram of simplified model for speech production

2.2. Estimation of Time-Delay using Information of Excitation Source:

Arrays of microphones are largely used for applications like speaker recognition of speaker /identification of speaker, surveillance of acoustics, teleconferencing, and acquisition of speech in automobile environments, capture of the sound with good quality and so on. An array of microphone have more favourable circumstances compared to a single microphone. Above all else using arrays of microphone display we can confine a source of sound and will be able to track the speaker's position absolutely. The next point of preference is that if the area of the source is known the array of microphone will be controlled by the source giving spatial separating. So the key necessity for each one of all these applications is the limit of the array of microphone to find a discourse or source of sound precisely. Widely two sort of s methodologies are present for constraining a source of sound.

Focalization utilizing a directed beamformer, High determination unearthly estimation routines and Time Difference of Arrival (TDOA) based strategies.

Most ordinarily employed technique for localization of source is the TDOA based technique and the headway frameworks are explored and other investigative figuring methods that can be associated with this strategy. The TDOA system shall be studied clearly. In this project we will be assuming that the origin of sound to be a point source and the microphones have a directional pattern that is omni-directional (although sensible showing could be conceivable). Hence there is a time delay between the sound waves because of several microphones. So localization of source is a two stage process. Initially the sign got by a few receivers is transformed to get data about the time-defer between sets of amplifiers. Different systems are there for assessing the time-delay that are taking into account a cross relationship between the signs. The time-delays that are estimated for pairs of microphones shall be utilized for calculating the area of the speaker.

2.3 Linear Prediction Algorithm

Let us assume that no of pairs of microphones present are M. m_1^i and m_2^i] are the vectors that represent the spatial coordinates of the two microphones for the ith pair of microphones(i.e., x,y,z) given i = [1 2,..M]. Let s give the location of source. The actual delay related to a source that is at s and for the ith pair of microphones can be calculated by the given formula

$$\tau_i\,(s)=\frac{|s-m_1^i|-|s-m_2^i|}{c} \quad \text{.....(18)}$$

Speech happens because of excitation of a signal for a period fluctuating vocal tract framework with time-differing excitation .A basic and noteworthy method of vocal tract excitation framework is the excitation of the voiced signals because of the vocal folds that are vibrating at the glottis, which to a first close estimation can be dealt with as comprising of a

succession of driving forces. The vocal tract framework data will be represented in terms of spectral features, which may be assumed to be superimposed on the glottal excitation pulses. The otherworldly highlights because of the vocal tract get tainted because of the transmission medium, noise and the room reaction. Although the epoch locations i.e., the moments of noteworthy excitation are not influenced by the exchange qualities of the microphones and the medium. By using the given speech signal, required information can be extracted by using the linear prediction (LP) analysis. In the analysis of Linear Prediction each sample prediction can be expressed as a combination of the p samples that are taken from past and which are linear , where *p* is the order of prediction. If given speech sequence is termed as s(n), then its value that is predicted can be given by,

$s'(n) = -\sum_{k=1}^{p} a_k\, s(n-k)$……(19)

where LP coefficients are given by {ak}. The error between the given speech sequence and that of its predicted one is given by

$r(n)=s(n)-s'(n)=s(n)+ \sum_{k=1}^{p} a_k\, s(n-k)$

In the Linear Prediction residual a large portion of the envelope of the spectrum data is uprooted. So the spectral corruptions because of noise and reverberation are dispensed with. The peaks that are present in the LP residuals cross-correlation are because of the epochs as epochs locations have to be robust to degradations. However the Linear Prediction residual signal amplitude varies relying upon the phase of the signal. Consequently on the off chance that we specifically use to Linear Prediction residuals, it may bring about a poor relationship peak. Along these lines, as opposed to utilizing the LP remaining specifically, another feature called the Hilbert envelope of the Linear Prediction residual is utilized.

Among a pair of microphones, the time-delay is assessed by calculating the cross-correlation function of the linear prediction residual's Hilbert envelope. The cross correlation function is calculated for every frame, and the displacement of the peak with respect to the center is the

time-delay of signal .Once the delays of time are assessed, the localization of source issue can be figured as takes after: Let us assume there are M sets of microphones. Let m1i and m2i for i ∈ [1,M] be the vectors representing the coordinates of the two microphones in the ith pair of microphones. Let the source be situated at s. The actual delay connected with a source at s and the ith pair of amplifiers is given by, where, c is the velocity of propagation of sound in the acoustical medium. By and by, for a microphone pair that is given, the assessed delay τi and the actual delay Ti(s) will never equal because of presence of noise in estimated time delay

T_i(s)=$\frac{|s-m1|-|s-m2|}{c}$…………………(20)

Let the no of sensors be M pairs , the estimated delays their spatial coordinates and we can find an estimate $\check{s}$ of the location of the source.

Given $m_1$and m_2 are the coordinates i.e. ((*x; y; z*) of the two microphones and assume the location of the source be s . If τ_i is the delay estimation and suppose is to be equal to the actual delay then

$\tau = \frac{|s-m_1|-|s-m_2|}{c}$………………(21)

where *c* is the speed of sound in air. The above equation with $\frac{m1+m2}{2}$ as the center with m1 and m2 being the two focal points and the line joining the two microphones as the axis of symmetry represents only one half (since we have to take into account the sign of) of a hyperboloid of two sheets. The above equation represents a half hyperboloid from a pair of microphones. So two microphones cannot determine the source location in 3D space uniquely. Hence to get the 3D location of source, the minimum number of microphones required is 3.

There are 3 possible pairs of microphones if we are given 3 microphones. One half of a hyperboloid represents for a pair of microphones. To specify a unique point in 3D space, three such type of hyperboloids intersects. Since they intersect to give a curve and not a unique point. τ_i, two pairs are not sufficient . With a known variance var(τ_i), the estimated time delay is corrupted by zero-mean additive white Gaussian noise. This variance is usually the result of the particular time delay estimation method. τ_i is normally distributed with mean Ti(s) and variance var(τ_i). With a source at s, Ti(s) is the actual delay associated for the ith pair of microphones which is given by

$$\tau_i \sim N(T_i(s), var(\tau_i))$$

Each of the time delays are independently corrupted by zero mean additive white Gaussian noise is assumed and the likelihood function can be written as:

$$P(\tau_{1,}\tau_{2,\ldots..}\tau_M; s) = \prod_{i=1}^{M} \frac{1}{\sqrt{2\pi var(\tau_i)}} \exp[\frac{(\tau_i - T_i(s))^2}{2var(\tau_i)}] \ldots\ldots\ldots\ldots\ldots\ldots(22)$$

The log-likelihood ratio is:

$$Ln(P(\tau_{1,}\tau_{2,\ldots..}\tau_M; s)) = \sum_{i=1}^{M} \ln(\sqrt{2\pi var(\tau_i)}) + [\frac{(\tau_i - T_i(s))^2}{2var(\tau_i)}] \ldots..(23)$$

$\hat{s}_{ML}$ is the position which maximizes the log likelihood ratio or equivalenlty one which minimizes , the Maximum Likelihood(ML) location estimate:

$$J_{ML}(s)) = \sum_{i=1}^{M} \frac{(\tau_i - T_i(s))^2}{2var(\tau_i)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(24)$$

This is same as the previous case except that the variance term comes into picture. Therefore

$$\hat{s}_{ML} = \arg(\min(J_{ML}(s)) \ldots\ldots\ldots\ldots(25)$$

Algorithm 2 :

- Capture the speech signal , apply a known delay .
- Find the 10^{th} order Linear prediction residual
- Find the Hilbert envelope
- Find the cross-correlation of the speech signal, GCC-PHAT,CC of Hilbert envelope

2.4 Nonlinear Least Square Methods (Gauss Newton):

The Gauss–Newton method is a strategy used to fathom non-linear least squares problems. It is an adjustment of Newton's strategy for calculating a function's minimum. Not at all like Newton's system, the Gauss–Newton calculation must be utilized to minimize an aggregate of squared capacity values, however it has the point of interest that second derivatives, which can be challenging to process, are not needed

Problems because of the non-linear least squares ensue in non-linear regression, where parameters in a model are sought such that the model is in good agreement with available observations. Given n variables $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ for m functions $\mathbf{r} = (r_1, \ldots, r_m)$, with $m \geq n$, the Gauss–Newton algorithm finds the minimum of the sum of squares iteratively .

$$s(\beta) = \sum_{i=1}^{m} r_i(\beta)^2 \text{.....(26)}$$

The method proceeds by the iterations, with an initial guess $\beta^{(0)}$ for the minimum at the beginning ,

$$\beta^{s+1} = \beta^s - (J_r^T J_r)^{-1} J_r^T r(\beta^{(s)} \text{.....(27)}$$

Are the entries of the Jacobian matrix , where, if $\mathbf{r}$ and $\boldsymbol{\beta}$ are column vectors and the symbol T denotes the matrix transpose.

If $m = n$, the iteration simplifies to

$$\beta^{s+1} = \beta^{s} - (J_r)^{-1}(\beta^{(s)})$$

Given model function $y = f(x, \boldsymbol{\beta})$, to find the parameters $\boldsymbol{\beta}$ such that a best fits some data points (x_i, y_i) is the goal, the functions r_i are the residuals

$$r_i\,(\beta) = y_i - f(x_i, \beta) \ldots (28)$$

Convergence properties:

If the algorithm converges, it can be shown that the increment is a descent direction for S, and then a stationary point of S is the limit., Not even local convergence as in Newton's method is guaranteed in this method.

For the Gauss–Newton algorithm, the rate of convergence of can approach quadratic. If the initial guess is far from the minimum or the matrix $J_r^T J_r$ is ill-conditioned, the algorithm can converge slowly or not at all For example, consider the problem with m=2 equations and n=1 variable, given by

$$r_1\,(\beta) = \beta + 1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (29)$$

$$r2\,(\beta) = \lambda\beta^2 + \beta - 1 \ldots\ldots\ldots\ldots\ldots\ldots (30)$$

The optimal is at $\beta = 0$. (for $\lambda = 3$ **but a**ctually the optimal is at $\beta = -2$, because S(0)=$1^2+(-1)^2$ =2, but S(-1)=0**;** If λ=0; and the method finds the optimum in one iteration, then the problem is in fact linear If |λ| < 1 ,error decreases asymptotically and then method converges that is linear with a factor |λ| at every iteration. However, then the method does not even locally converge if |λ| > 1.

In Gauss–Newton method convergence is not guaranteed in all instances. The approximation

$$r_i \frac{\partial^2 r_i}{\partial \beta_j \partial \beta_k} \ll \frac{\partial r_i}{\partial \beta_j} \frac{\partial r_i}{\partial \beta_k} \text{.....(31)}$$

that needs to hold to be able to ignore the second-order derivative terms may be valid in two cases, for which convergence is to be expected. The function values r_i are small in magnitude, at least around the minimum. The functions are only "mildly" non-linear, so that $r_i \frac{\partial^2 r_i}{\partial \beta_j \partial \beta_k}$ is relatively small in magnitude.

2.5 Conclusion:

Because of flattening of the spectrum, GCC PHAT resulting peak corresponds to a dominant delay but it works well when the noise is low. For ML function, as the room reverberation increases, its performance decreases. So room reverberation affects ML estimation function. But these methods do not utilize methods of speech production to get more robust estimate. TDE using source's excitation information performs production of speech technique to estimate TDE .It is robust to room reverberation and noise. It takes into consideration only the epochs of the speech.

Chapter 3

Video Tracking

As a rule, existing methodologies detailed to manage the tracking of a solitary moving article or multiple objects can be subjectively grouped into a few classes, i.e. feature-based approach, template-based method, gradient-based method, statistical model and prediction approach. The highlight based methodology uses certain qualities of the target picture, (for example, line sections, corners, contours and curvilinear) for creating ID in the following frame. A proficient representation model for the target is essential for this technique and some of the time the definite element models are either hard to get or need complex numerical portrayals. Then again, the format based technique takes the layout overall. Both of these systems are less vigorous to changing state of the followed target and transitory impediments. Both of these methods are less robust to changing shape of the tracked target and temporary occlusions.

The gradient-based strategy performs following tracking based on spatio- temporal gradient of the image intensity or the depth and motion data of the moving object with respect to the camera. The measurable model includes acquiring some former data about the flow or certain highlight attributes of the target and in this way subsequently analyzing the corresponding posterior distribution. As a rule, this technique is computationally requesting and in a matter of seconds scarcely helpful continuously frameworks with high frame rate.

The prediction approach includes determining the consequent direction of the tracked object taking into account past estimations. The current expectation procedures are for the most

part Kalman Filter (KF) and the Extended Kalman Filter (EKF), where a state model that well represents the dynamics of the target is produced, and the state parameters are evaluated in view of the procedure and estimation comparisons by considering the additive stochastic noise. Then again, KF can be a lapse, error-prone inclined if the supposition of neighborhood linearity is disregarded or when the followed target effectively moves. Predictive tracking methodology used in the work manages the movement estimation from regression analysis, where the movement parameters are fitted by the optimal requests of the time-depending polynomials models. Numerous ordinary position based following methods as a rule characterize the flow models for position forecast on target. Hence, the actualized framework gets to be unyielding and less powerful. The following calculation used in this project is in view of the forecast approach as it is more vigorous to interim impediments and it empowers quicker following pace to be accomplished.

3.1 Motion Detection

The motion detection scheme employed is based on the differencing of frame technique, from the static-background image sequences to extract edges of the moving object. At this stage, the detection of edge using hexagonal detector is applied to the raw image pixel values in order to obtain the edge map for the current frame. It has been selected, instead of other traditional gradient edge operators (such as Sobel, Prewitt, Roberts) due to its significant performance in computation time, as well as accuracy in generating edge pixels as verified in the project. Comparison of edge maps (frame differencing) is carried out by subtracting the edge pixels of the previous frame from the edge pixels at the corresponding spatial location of the current frame. If the pixel difference exceeds the threshold value for binarization, its value will be marked as ‘255’ to indicate availability of motion. Otherwise, the pixel will be assigned with value ‘0’ which implies static background. The resulting object mask contains information

about the location, size as well as the shape of the moving object. The moving edges extraction process is shown in the fig 5.3

3.1.1 Projection Histograms

The Projection Histograms method is implemented on the edge detected image to find out the actual area (centroid) of the object that is moving in the current frame, by using the projections in the horizontal and the vertical directions.

For tracking the area of the detected moving object, initially, the values of both the horizontal and vertical projections are obtained. Horizontal axis projection can be obtained by summing up all the pixels column-wise, i.e.,

$p_v(i) = \sum_j f(i,j)$….(32)

$p_h(i) = \sum_i f(i,j)$……(33)

where: i is column, i = 0, 1.0, 2.0, 3., …, 255

j is row, j = 0, 1., 2., 3., …, 255

f (i, j) is the pixel value at column i and row j

			255					$p_v(j)$ =
			255					
		255	255	255	255	255	Vs →	1275
		255	255	255				765
			255	255				510
			255					255
			255				Ve →	255
		Hs ↓				He ↓		
$p_h(i)$		510	1785	765	255	255		

Fig 3.1 Summations of rows and columns to obtain Vertical and Horizontal projections

These projection values are used next and a bounding box that encompasses all these binary edges of the object that I being tracked is drawn. Cx and Cy give the center the bounding box and will then be identified a centroid of the moving object. These centroid values can be calculated by using the following formula

Coordinate of x for the centroid, $Cx = (Hs + He) / 2$

Coordinate of y for the centroid, $Cy = (Vs + Ve) / 2$

Where:

Hs = coordinate of x for the horizontal perimeter at the starting

He = coordinate of x for the horizontal perimeter at the end

Vs = coordinate of y for the vertical perimeter at the starting

Ve = coordinate of y for the vertical perimeter at the end

After the above data is calculated and made available, prediction (based on Maximum Entropy Method and the Linear Prediction method) is executed to calculate the centroid of the object that is in the next frame. The number of data required depends on the order of Linear Prediction being employed. Calculation of predicted values of Cx,Cy is done independently. The respective predicted values for Cx and Cy are stored in the separate database for prediction.

3.2 Tracking Algorithm

The tracking algorithm which is used here combines the Linear Prediction with the Projection Histograms technique and hence the Linear Prediction algorithm will be helpful to enable the implemented tracking system to achieve real-time performance.

3.2.1 Linear Prediction

Linear Prediction is used to predict the next centroid location of the tracked target denoted as Cn , based on its finite past centroid measurements. In the developed tracking system, the second order Linear Prediction is adopted. This second order Linear Prediction has been identified as the optimal order for the employed Linear Prediction, by analysing the trade-off between the computational time and accuracy based on the acquired image sequences. From our empirical experiments as shown in , it has been found that the second order Linear Prediction enables the resultant predictor to achieve minimum mean error as well as fast computation time.

From the past 3 centroid coordinates, prediction is obtained for the second order predictor that is being used, namely {Cn-i ; i = 1,2,3}. The linear predictor of order 2 can be written in the form as shown

$\check{c}_n = -\sum_{i=1}^{3} a_i c_{n-i}$.....(34)

a1, a2 and a3 are the prediction coefficients and these are selected by minimizing the mean-squared error of prediction :

$\varepsilon = E[e_n^2] = min$.....(35)

where e_n is the prediction error

$e_n = c_n - \check{c}_n$.....(36)

The best linear predictor of order 2, Cn can be obtained efficiently by using Levinson's algorithm, where the lattice realizations of best linear prediction filters for orders p = 0, 1 and 2 are determined. In the algorithm that is used, the Maximum Entropy Method (MEM) is employed to solve the prediction coefficients, due to its better performance as compared to the autocorrelation and the covariance method. It is competent to guarantee that the predictor does not keep running off the piece of information, and dependably brings about a minimal-phase filter. Additionally, its minimization rule has the capacity create more exact expectation (contrasted with the covariance and autocorrelation technique), by minimizing the aggregate forward and backward squared prediction errors, i.e.:

$\varepsilon = E[e_n^2] = min$……..(37)

The iterative procedure of Levinson recursion is applied to determine the prediction error filter of order p:

$$\begin{pmatrix} 1 \\ a_{p,1} \\ a_{p,2} \\ \vdots \\ a_{p,p} \end{pmatrix} = \begin{pmatrix} 1 \\ a_{p-1,1} \\ a_{p-1,2} \\ \vdots \\ a_{p-1,p-1} \\ 0 \end{pmatrix} - \gamma_p \begin{pmatrix} 0 \\ a_{p-1,p-1} \\ a_{p-1,p-2} \\ \vdots \\ a_{p-1,1} \\ 1 \end{pmatrix} \text{……..(38)}$$

The below mentioned lattice relationships are validated for n that falls in the range of $p \leq n \leq 2$ to make sure that the filter does not waste the useful data.

$e_p^+(n) = e_{p-1}^+(n) - \gamma_p e_{p-1}^-(n-1)$

$e_p^-(n) = e_{p-1}^-(n-1) - \gamma_p e_{p-1}^+(n)$……..(39)

Elaborate explanation on the Maximum Entropy Method and the Linear Prediction method used in the implemented tracking algorithm can also be referred in. Once the predicted

centroid is obtained from the second order Linear Prediction, the search region for the subsequent frame will be constrained to certain confined region based on the predicted centroid location, as shown in Figure 4. This inevitably helps to speed up the tracking process, as both the motion detection and tracking modules can be applied to a smaller area for the next time interval, rather than covering the whole frame. Subsequently, the centroid error *Ce* is computed. Due to simplicity, we have chosen to determine the error based on the Euclidean distance between the predicted centroid and the actual centroid can be given by

$$c_e = \sqrt{(c_x - \check{c}_x)^2 + (c_y - \check{c}_y)^2} \quad \text{.....(40)}$$

Where: (Cx,Cy) = actual centroid coordinates

(Cx, Cy) = predicted centroid coordinates

On the off chance that the error exceeds a pre-fixed threshold value, the actual value is updated to the centroid value in database of prediction utilizing the Projection Histograms technique. Something else, the anticipated centroid quality is viewed as reliable and the area of the moving object or person for whenever moment will be resolved just by this value.

The used algorithm followed by calculations has been executed on a Pentium IV PC. In diverse scenes that contain moving vehicles or persons the above algorithm is performed for video image sequences. The edge detection is done in the grayscale arrangement and the span of every envelope is adjusted to a size of 256 * 256 pixels. The average tracking velocity accomplished is roughly 5 to 20 fps. From there results that are obtained, it is clear that the Linear prediction algorithm following calculation has the capacity to track the recognized moving object precisely at an ongoing rate. The normal following pace attained to is roughly 5 to 20 fps, for a 256 * 256 pixel picture.

3.3 Maximum entropy principle

Another way of mentioning about this: Take precisely stated prior data or testable information about a probability distribution function. Consider the set of all trial probability distributions that would encode the prior data. Of those, the one with maximal information entropy is the proper distribution, according to this principle. Take precisely expressed prior information or testable data around a PDF. Consider the arrangement of all trial likelihood functions that would encode the earlier information. Of those, the one with maximal data entropy is the best possible distribution, as indicated by this rule. Given testable data, the most extreme entropy strategy comprises of looking for the likelihood dispersion which amplifies data entropy, subject to the requirements of the data. This obliged enhancement issue is ordinarily unraveled utilizing the strategy for Lagrange multipliers.

The principle of most extreme entropy is valuable expressly just when connected to testable data. Testable data helps to give information regarding a probability distribution whose truth or false is very much specified. For instance, the announcements the desire of the variable x is 1.37 and p4 + p5 > 0.6 (where p4 + p5 are probabilities of occasions) are explanations of testable data. Given testable data, the maximum entropy strategy comprises of looking for the likelihood appropriation which boosts data entropy, subject to the limitations of the data. This constrained optimization issue is commonly tackled utilizing the technique for Lagrange multipliers. Entropy maximization with no testable data regards the widespread "requirement" that the whole of the probabilities is one. Under this limitation, the most extreme entropy discrete likelihood conveyance is the uniform circulation

$$p_{i=} \frac{1}{n} \qquad \text{for all} \quad i = 1,2,3 \ldots n \ldots\ldots\ldots (41)$$

3.3.1 Discrete case

Some testable data I around an amount x taking values in {x1, x2,..., xn} is present . This data that has m type of constraints on the expectations of the function f_ki.e., the likelihood distribution that is required to satisfy is given by

$\sum_{i=1}^{n} \Pr(x_i)\, f_k(x_i) = F_k \quad \{K = 1,2 \ldots . m\}$……..(42)

And, the probabilities should sum up to one, giving the constraint

$\sum_{i=1}^{n} \Pr(x_i) = 1$……..(43)

With maximum information entropy subjected to these constraints, the probability distribution is

$\Pr(x_i) = \frac{1}{Z(\lambda_1,\lambda_{2..},\lambda_m)} \exp(\lambda_{1,} f_1(x_i) + \cdots .. \lambda_{m,} f_m(x_i))$……..(44)

It is sometimes called the Gibbs distribution. The normalization constant is given by

$z\left(\lambda_{1,}\lambda_{2..,}\lambda_{m,}\right) = \exp\left(\lambda_{1,} f_1(x_i) + \cdots .. \lambda_{m,} f_m(x_i)\right)$……..(45)

And, is customarily called the partition function. (Interestingly, the Pitman–Koopman hypothesis expresses that the fundamental and necessary condition for the sampling distribution to admit sufficient statistics of bounded measurement is that it have the general type of maximum entropy distribution.)

The λk parameters are Lagrange multipliers whose specific values are determined by the constraints as indicated

$F_k = \frac{\partial}{\partial \lambda_{1,}} \frac{1}{Z(\lambda_1,\lambda_{2..},\lambda_m)}$……..(46)

These *m* simultaneous equations do not generally possess a closed form solution, and are usually solved by numerical methods.

3.3.2 Continuous case

For distributions which are continuous, the Shannon entropy will not be used, because it is only defined for probability spaces that are discrete. But Edwin Jaynes has given the following formula that is close to the relative entropy .

$$H_c = -\int p(x) log \frac{p(x)}{m(x)}\, dx \text{.....(47)}$$

m(x) is called the "invariant measure", which is relative to the restricting thickness of discrete focuses. But m(x) is taken to be known till any further mathematical statements are given for any calculations.

A firmly related amount, the relative entropy, is normally characterized as the Kullback–Leibler difference of m from p .The Principle of Minimum Discrimination Information is the derivation standard of minimizing this,

Some testable data I around an amount x which takes values in some interim of the genuine numbers is present. This data that have m type of imperatives on the desires of the capacities fk, i.e. probability density to fulfil is required .

$$\int p(x) f_k(x) dx = F_k \quad \{k = 1,2,..m\} \text{.....(48)}$$

And, the probability density should integrate to one, which gives the constraint

$$\int p(x) dx = 1 \text{ .....(49)}$$

The probability density function with maximum H_c s which is subjected to these constraints is

$$p(x) = \frac{1}{Z(\lambda_1, \lambda_{2..}, \lambda_m)} \text{m(x)} \exp\left(\lambda_1 f_1(x_i) + \cdots .. \lambda_m f_m(x_i)\right) \text{.....(50)}$$

with the partition function determined by

$$Z(\lambda_1, \lambda_{2..}, \lambda_m) = \int \text{m(x)} \exp\left(\lambda_1 f_1(x_i) + \cdots .. \lambda_m f_m(x_i)\right) \text{.....(51)}$$

As in the discrete case, the values of the λ_k parameters are determined by the constraints according to

$$F_k = \frac{\partial}{\partial \lambda_{1,}} \log(Z(\lambda_{1,}\lambda_{2..,}\lambda_m)) \ldots\ldots\ldots\ldots\ldots\ldots (52)$$

3.4 Hexagonal edge detection

As the domain under which computer vision is applied is extending, the image characterization of the progressively relies on upon data gathering on the structures that are contained inside of an image. These structures are frequently comprised of unpredictable bends. It is hence of central significance to discover effective methods to represent these structures. One such technique that is regularly utilized is edge detection. Be that as it may, the accomplishment of edge detection is regularly constrained with regards to curved structures. This is principally because of the effectiveness of algorithm being restricted by the nature of the image given as input. Subsequently, matrices which have resolution that is high are regularly utilized as a part of request to represent the information with adequate constancy. For the domains under which computer vision is applied extends, the characterisation of the image progressively relies on upon gathering data about the structures contained inside of an image. Another approach to beat this trouble is through the utilization of option inspecting grids. Hexagonal lattice sampling is a very good solution that has received some attention and been shown to have better efficiency and less aliasing. The Hexagonal sampling grid offers less capacity time, less processing time, expanded coding effectiveness, less quantization blunder, equidistant property and predictable integration, and so on

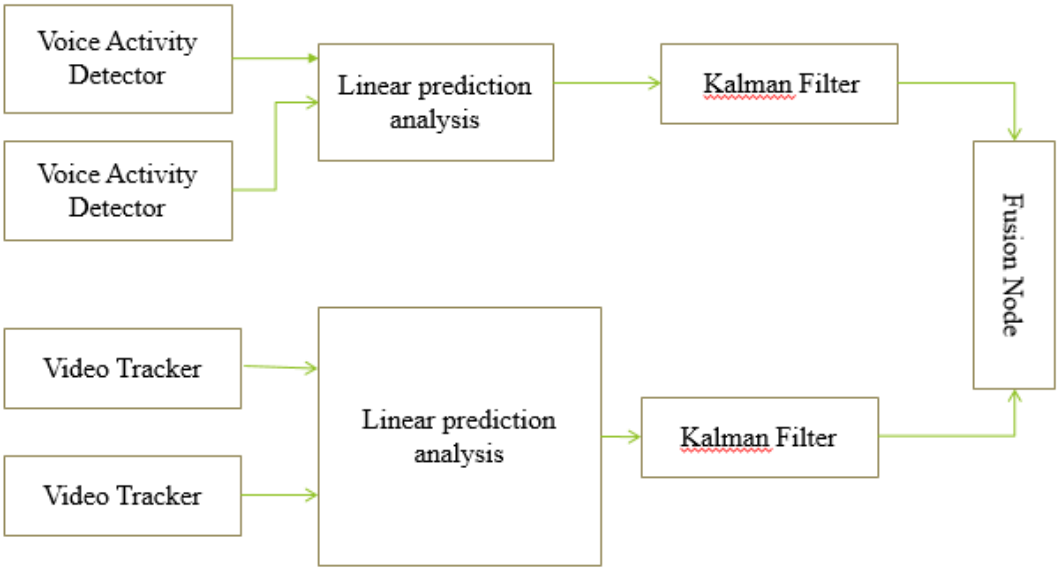


Fig 3.2 Schematic of Video Tracking Algorithm

Algorithm 3

- Input the video frames
- Hexagonal Edge Detector
- Differencing and Binarization
- Output of Moving edge detection process
- Frame Differencing
- Projection Histogram
    - Calculate Hs,He,Vs,Ve
- Centroid Prediction and confine search area
- Error computation and Data updating
- Object tracking

3.5 Conclusion

For tracking, it is found that linear prediction of second order is enough for getting tracking results that are good . The system used for tracking in each sequence and hence succeeds to track the detected moving object accurately and real-time speed is achieved. Though accuracy of tracking using Kalman Filter is efficient for linear estimates but it is not fast in real time. Tracking using linear prediction is faster in real time. For each frame of image, the result obtained from the Linear prediction method and the Kalman Filter is compared. It is clear from the Linear Prediction algorithm that it exhibits better accuracy. The results show the tracking results of different frame numbers.

Chapter 4:

Results and Discussion

4.1 Source Localisation Results

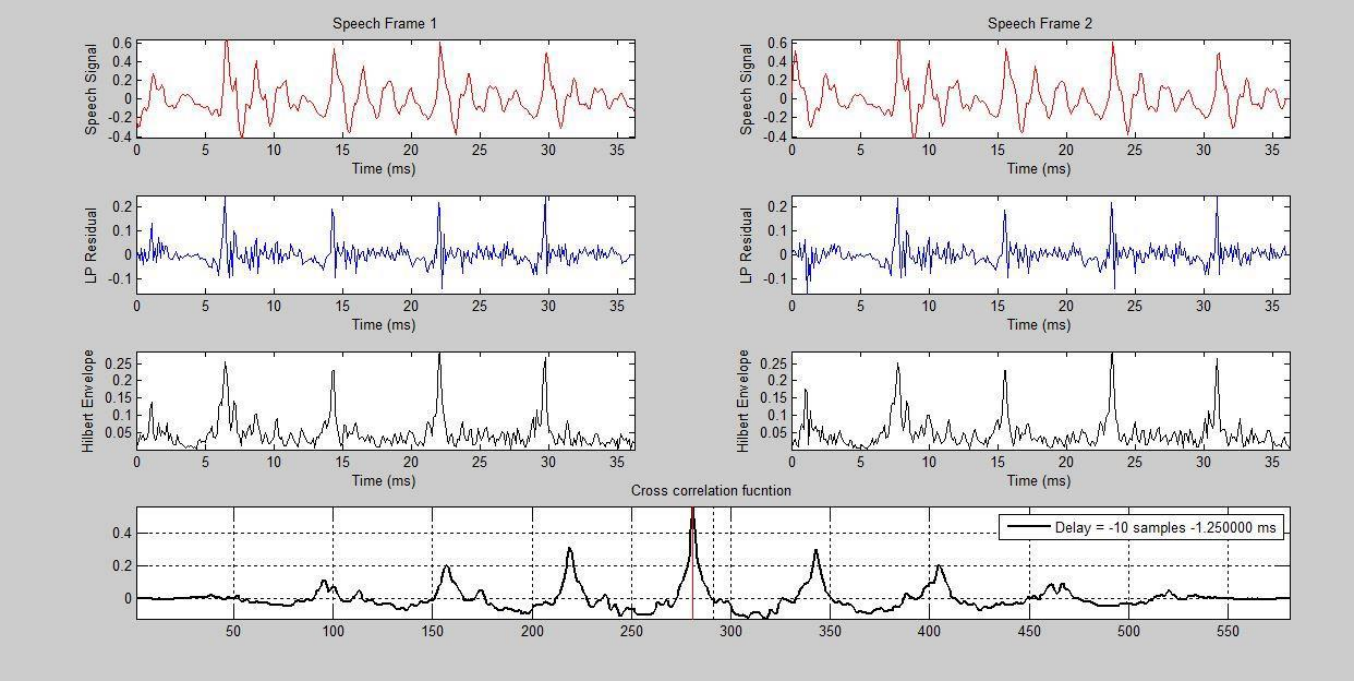


Fig 4.1 (a) Speech signals from different microphones (b) Their 10th order linear prediction residuals (c) Respective Hilbert envelopes (d) cross-correlation of the Hilbert envelopes

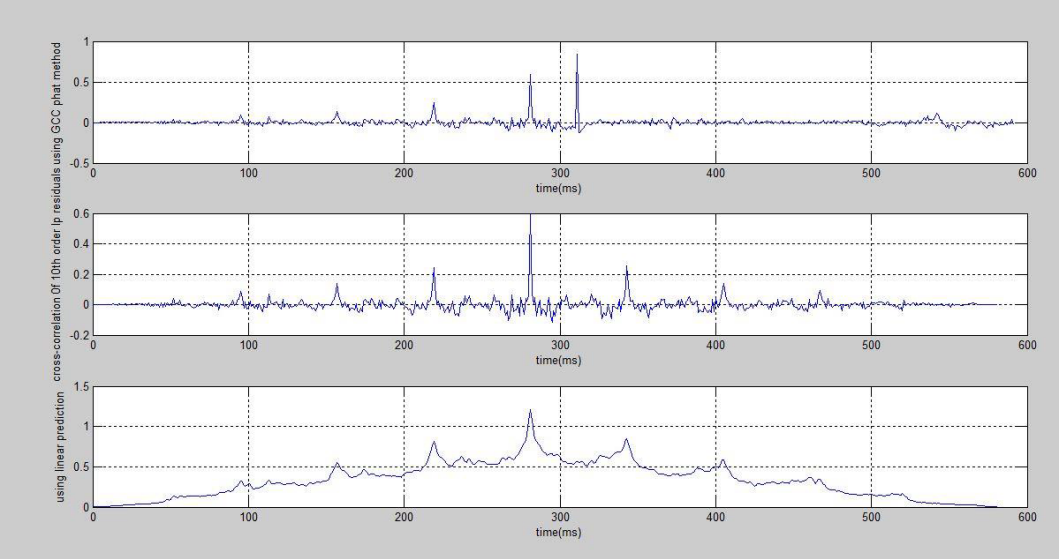
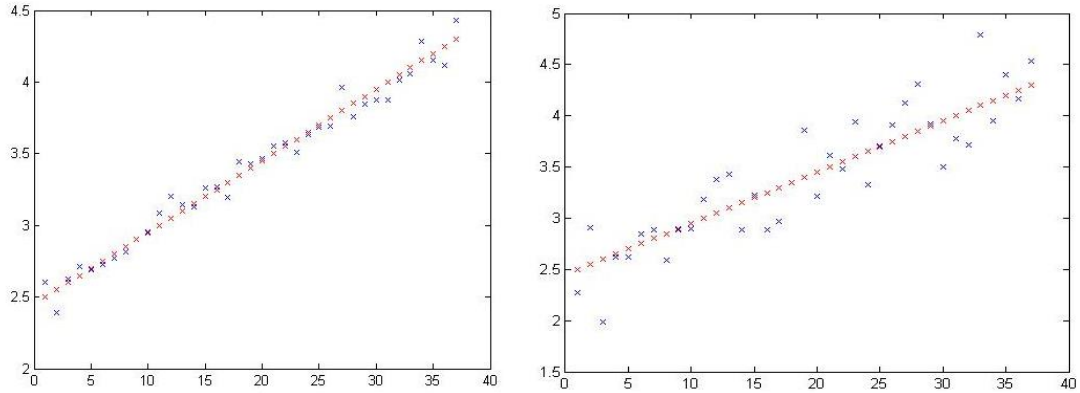
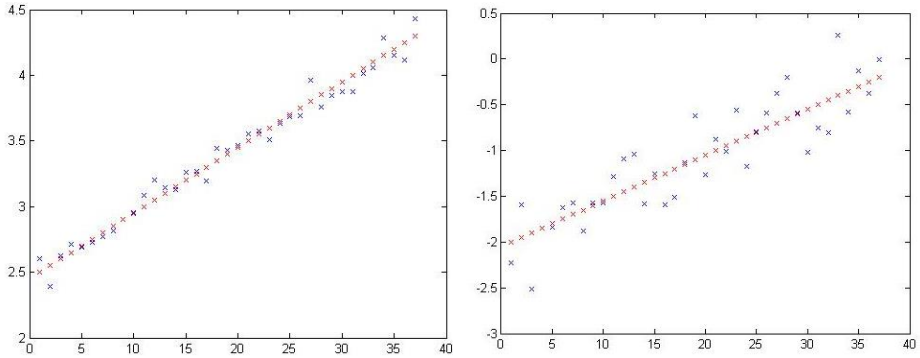


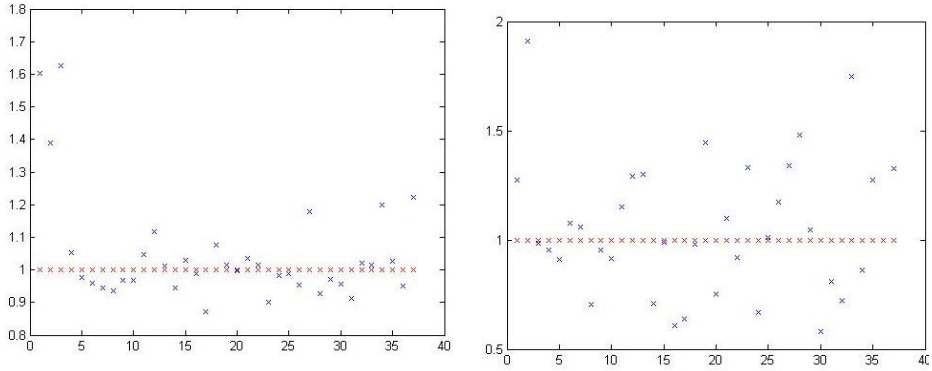
Fig 4.2 (a) PHAT weighting of GCC, (b) Cross-correlation of the 10*th* order LP residuals and (c) Cross-correlation of the Hilbert envelopes of the corresponding LP residuals of to 37.5ms speech segments



X coordinate for LP and GCC PHAT



Y coordinate for LP and GCC PHAT



Z coordinate for LP and GCC PHAT

Fig 4.3(a)(b)(c)Plot of the actual and the estimated *x*, *y* and *z* coordinates of the speaker using the used approach and using the GCC approach

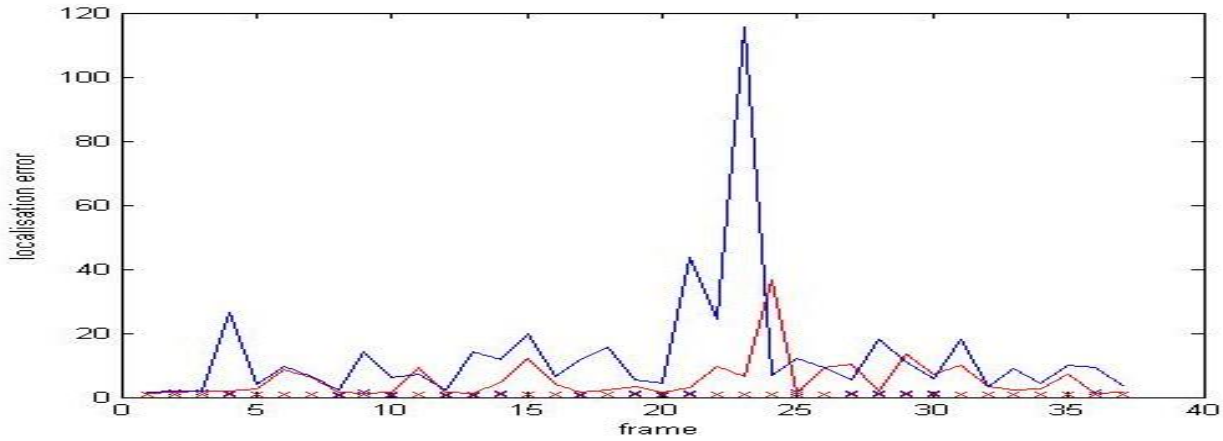


Fig 4.4 Localisation error (in cm)

Inferences:

- Results obtained from the linear prediction algorithm which is based on information regarding source of excitation are compared with that of the existing GCC-PHAT approach which is based on content based on spectrum. The values that are estimated using linear prediction are more close to the actual tracking path. Significant information regarding moving speaker is given by features of source and features of vocal tract system. Epochs present in information of excitation source are used in present method to calculate the delay of time. The delay of time is given by the displacement of peak from center. Hence by combining the source of excitation information and features of spectrum, an efficient way of calculating the parameters for tracking of a speaker is found

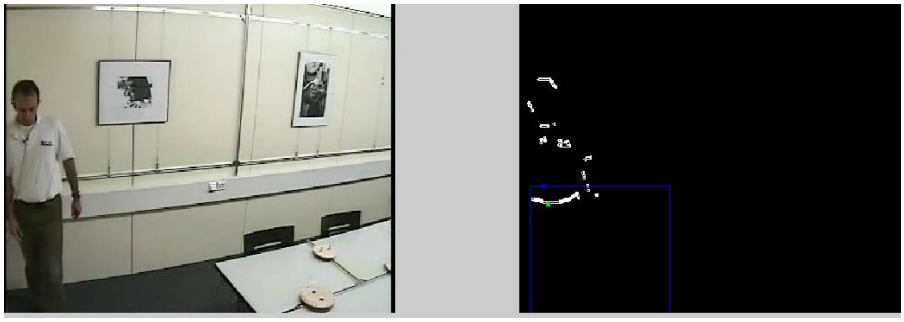
4.2 Video Tracking Results:

Frame 120

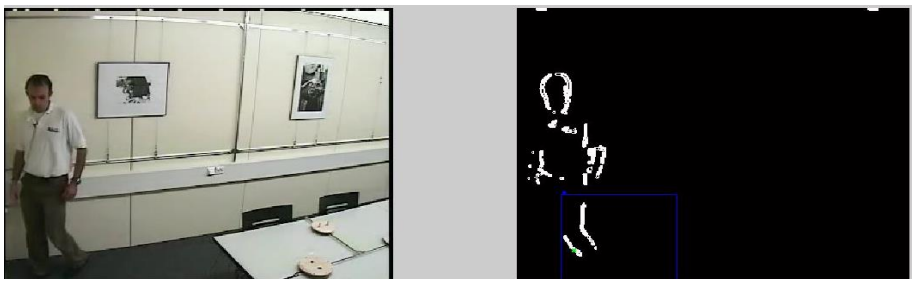
Frame 130

Frame 140

Frame 120



Frame 128



Frame 135

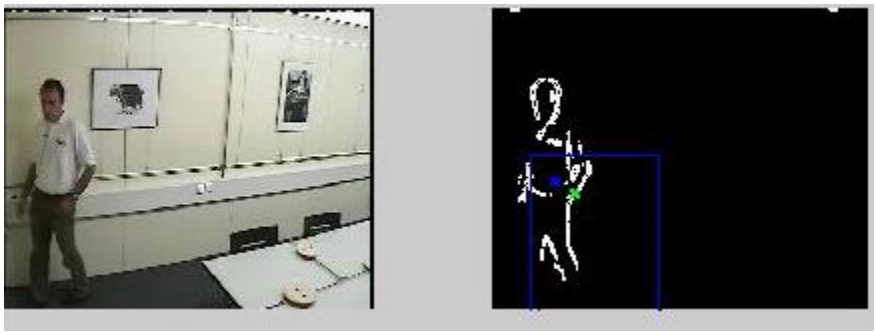


Fig 4.5 Plot of tracking results by the Linear Prediction algorithm

Red-linear prediction

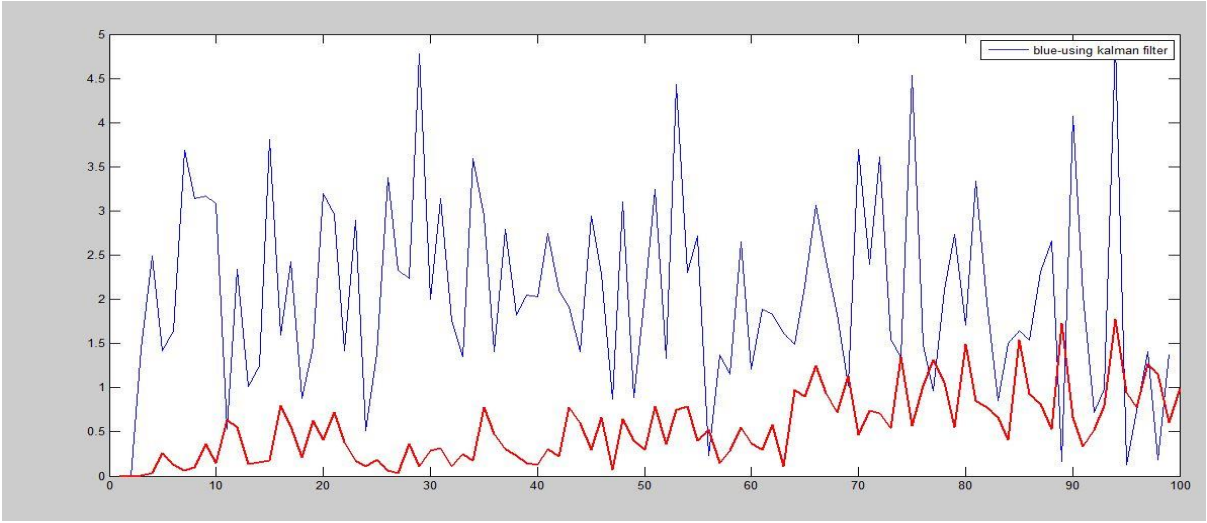


Fig 4.6 Average mean error Vs each frame

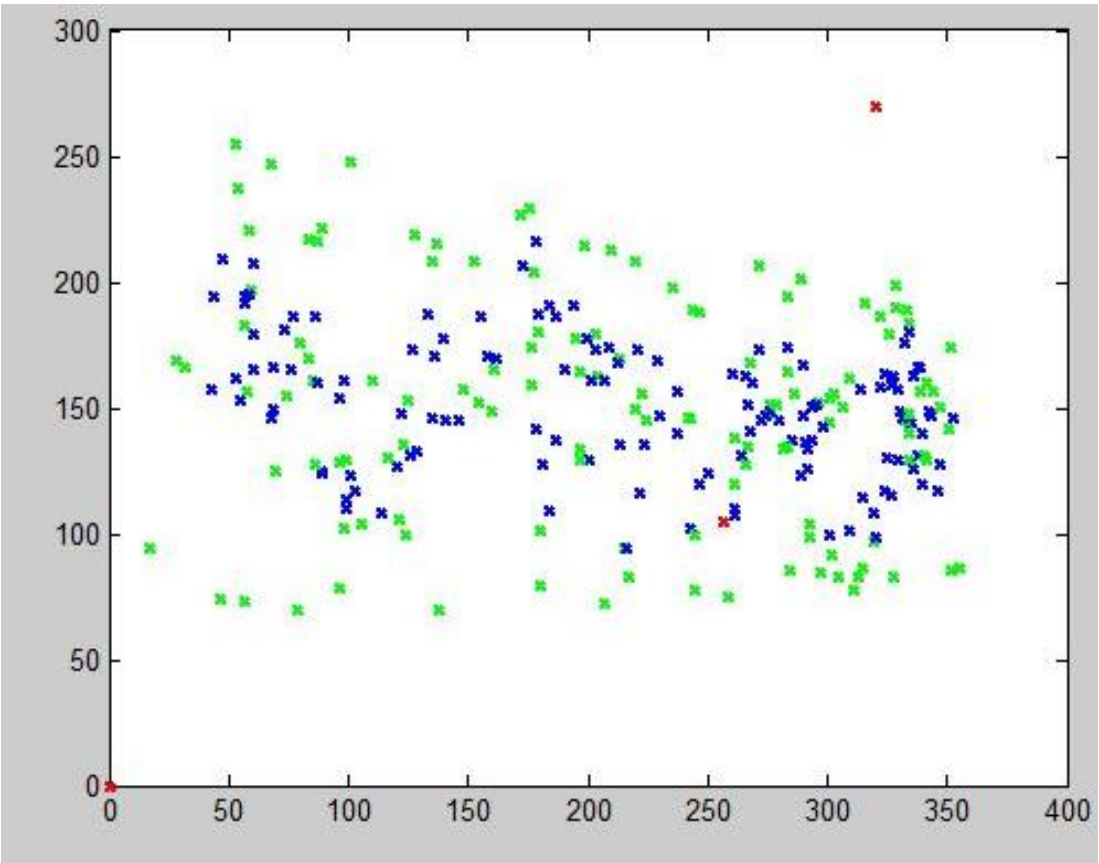


Fig 4.7 Accuracy comparison for the centroid locations of the target

Blue-linear prediction Green--kalman filter

Inferences:

- Hexagonal edge detected image is more robust to noise compared to other edge detection algorithms because of its uniformity.
- An alternative tracking algorithm which is better compared to the ones that are existing by using Linear Prediction technique has been presented. In this algorithm, the moving object's centroid is detected using Projection Histograms method. Before applying Projection Histograms technique, detection of motion is done according to the Hexagonal Edge Detection method
- For tracking, it is found that linear prediction of second order is enough for getting tracking results that are good . The system used for tracking in each sequence and hence succeeds to track the detected moving object accurately and real-time speed is achieved.

- Though accuracy of tracking using Kalman Filter is efficient for linear estimates but it is not fast in real time. Tracking using linear prediction is faster in real time. For each frame of image, the result obtained from the Linear prediction method and the Kalman Filter is compared . It is clear from the Linear Prediction algorithm that it exhibits better accuracy. The results show the tracking results of different frame numbers.

4.3 Conclusion and Discussion:

In this project a novel audio and video tracking systems for single moving speaker using linear prediction method is used and compared with GCC PHAT and Kalman filter methods. The results prove that linear prediction method is better suitable for tracking compared to other methods.

In source localization ,a normal room of size 3m ×4m×5m whose reverberation time was approximately 200ms.that contains 4 set of 4 pairs i.e., 16 sets of microphones is assumed. The speech signal is sampled with 16 bit resolution and 8 KHz. For tracking, at his normal level of speaking the speaker was made to move in the room while speaking. The speaker was asked to move in a path that is predetermined whose coordinates were known in order to validate the results. In the results shown the speaker is made to move from one end of the room towards the array of microphones along the trajectory of a straight line. It is clear from the delays that are estimated by the Linear prediction method are more uniform compared to those estimated using the GCC-PHAT method. Also if the frame length is increased, the delays obtained will become more consistent.

Figure 4.4 shows the corresponding error of localization as a function of the frame number. The Euclidean distance between the actual position and the position that is estimated gives the localisation error. Similar results were observed for all the simulations done. We will show the results for only one case because of the space constraints. Hence, linear prediction method is compared to GCC-PHAT method by considering signals from a pair of microphones and it is proved that used algorithm is better compared to other approaches.

The performance (in accuracy) of the used tracking method is shown in Figure 4.5, which depicts the centroid locations (in pixel) of the tracked object obtained between using the

2nd order LP and Kalman Filter. The object for that sequence is moving from upper right to the bottom left portion. It can be seen that the accuracy of Kalman Filter degrades when the object manoeuvres itself slightly near the bottom left portion of Figure 4.5, whereas Linear Prediction is able to track the object steadily. Comparison of the tracking accuracy (in terms of the average mean error) for each frame of the image sequences obtained between the used method and the Kalman Filter is depicted in Figure 4.7. It is clear that the used tracking algorithm exhibits better accuracy.

For audio-visual fusion, both of results are fed to kalman filter, fusion could be implemented to demonstrate how audio modality can help video to track trough occlusions when high measurement ambiguities occur in a non-meeting scenario. In particular, by symmetrically fusing synchronised audio and video signals at the likelihood level the system is able to track object with more certainity. However , the last part of work could not be possible to complete in limited time and hence planned for future work.

4.4 Future work:

In the future we want to be able to track single moving speaker in a larger area when audio and video ambiguities coexist for longer by using Linear Prediction method by audio-visual fusion. Also we want to be able to track multiple moving speakers by using audio-visual fusion. More sophisticated tracking algorithms and several additional audio and video features including gestures and speaker identification other than using vocal tract information.

Bibliography

[1] B. Y. D. Vikas C. Raykar, "Speaker Localization using Excitation Source Information in Speech," IEEE transactions on audio and speech processing, vol. 13, no. 5, pp. 751-761, Sept. 2005.

[2] A.-B. S. Yeoh P.Y, "Accurate Real Time Object Tracking using Linear Prediction," in International Conference on Image Processing, 2003.

[3] N. M. R. Eleonora D'Arca, "Person Tracking Using Audio-Visual fusion," in Data Fusion & Target Tracking Conference (DF&TT 2012): Algorithms & Applications, 9th IET, may 2012.

[4] C. H. K. a. G. C. Carter, "The generalized correlation," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 24, no. 7, pp. 320-327, 2003.

[5] "Linear prediction: A tutorial review," IEEE Proceedings, vol. 63, pp. 561-580, 1975.

[6] S. R. M. P. a. K. S. R. B. Yegnanarayana, "Speech enhancementusing excitation source information," IEEE Int. Conf.Acoust., Speech, Signal Processing, vol. 1, pp. 541-544, 2002.

[7] S. G. P. H. a. M. B. C. Wang, "Real-time automated video and audio capture with multiple camera and microphones," Journal of VLSI Signal Processing Systems, vol. 29, pp. 81-100, 2001.

[8] T. V. A. a. B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," IEEE Trans. Acoust., Speech, Signal Processing, vol. 27, no. 4, pp. 309-319, 1979.

[9] C. a. H. R. Mehmet, "Moving Object Tracking Using Local Windows," Proceeding of the IEEE, pp. 180-185, 1998.

[10] J. a. V. L. Frau, "Predictive Tracking of Targets Using Image Sequences," Proceeding of IEEE RSJ International Workshop on Intelligent Robots and Systems (IROS '91)., 1991.

[11] A. B. a. R. J. G. Syed, "Detection of Edges Based on Hexagonal Pixel Formats," Proceeding of the 3rd. International Conference on Signal Processing (ICSP, vol. 2, p. 1114 –1117., 1996.

[12] P. Y. a. S. A.-B. Yeoh, "Performance Study on Linear Prediction For Tracking of A Moving Object," Proceeding of the Malaysian Science and Technology Congress, 2003.

[13] H. a. Y. S. B. Kyu-Bum, "Visual Servo Tracking Strategy Using Time-Varying Kalman Filter Estimation," Proceeding of The Fourth Conference on Motion and Vibration Control, 1998.

[14] B. S. Cruceru, “Kalman based video tracking In a video Surveillance system,” 10th International Conference on development and application system, 2010.

[15] D. d. V. a. P. V. A. J. Berkhout1, “Acoustic Control by wavefrontsynthesis,” Delft University of Technology ,laboratory of Seismics and Acoustics, 1993.

[16] Karasulu, “ Review And Evaluation Of Well-known Methods For Moving object detection and Tracking Videos,” 2001.

[17] R. D. B. Yegnanarayana, “Processing of Reverberant Speech for Time-Delay Estimation,” IEEE Trans. Acoust., Speech, Signal Processing" vol. 13, no. 6, 2005.

[18] W. T. F. T. D. a. P. V. John W. Fisher, “Learning joint statistical models for audio-visual fusion and segregation,” 2001.

[19] K. N. H. E. U. K. a. J. M. T. Gehrig, “Kalman filters for audio-video source localization,” 2005.

[20] J.-M. O. a. D. G.-P. Guillaume Lathoud, “ Audio-Visual Corpus for Speaker Localization and Tracking,” Proceedings of the MLMI'04 Workshop, 2004.