Bonafide Certificate

MODEL ENGINEERING COLLEGE

THRIKKAKARA, KOCHI-21

DEPARTMENT OF ELECTRONICS AND COMMUNICATION

COCHIN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Bonafide Certificate
This is to Certify that the Project Report entitled

Real Time Sound Source Localisation System

Submitted by

Akshay A.J

Ganjimala Pavankumar

Navaneeth Paliath

Paul V. Sebastian

Ranjith R Nair

is a bonafide account of their work done under our supervision.

Project Co-ordinator	Project Guide	Head of Department
Sheeba P S	Laila D.	Mr. Pradeep M
Assistant Professor	Professor	Associate Professor

ACKNOWLEDGEMENT

First and foremost we would like to thank God Almighty for his blessings all through the design phase of this project.

We would like to express our sincere gratitude to our guide Dr. Laila D., Professor, Department of Electronics and Communication, Model Engineering College, Thrikkakara for helping/guiding us to utilize our skills and knowledge to the best of our abilities to fabricate and implement a project that will enable us to develop our learning to greater height and also promote among us the spirit of teamwork, good organisation and time management.

We are also obligated to our Principal Prof.(Dr.) V.P Devassia and Mr. Pradeep M, Head of the Department of Electronics and communication, Model Engineering College for providing us with platform to exhibit our skills.

Last but not the least, our parents have also played a pivotal role in the course of this project by catering to our every need and providing us with the best of everything to ensure the completion of this project.

ABSTRACT

Humans live in a complex audio environment. We have very good skills of following a specic sound source while ignoring or simply acknowledging the others. Machine listening systems are still far from the level of human performance in recognizing them. Hence the need for a sound source separation system. The primary goal of this project is to build a system implementing the basic mechanisms underlying localization of sound. Two physical cues dominate the perceived location of an incoming sound source: Time difference of arrival and Intensity difference of arrival. In this project we attempt to localize a single sound source by using array of microphones placed accordingly. In an ordinary situation, the source can be located anywhere in the three dimensional space. The use of more microphones would increase the accuracy of the localization of the source at the cost of computational time; hence an optimum no of microphones are to be used. One of the major challenges is judging the location of a sound source in a natural acoustic environment, such as an enclosed room, where sound is reflected from various surfaces. The solution is to analyse only the direct sound which is arriving first, for sound localization, but not the reflected sound, which arrives later. So sound localization remains possible even in an echoic environment. Applications include Multi-speaker Teleconferencing, Gun-shot detection, Robotic hearing, Surveillance etc.

Table of Contents

List of Figures

Chapter 1

Introduction

Sound localization is the process of determining the location of a sound source. The brain utilizes subtle differences in intensity, spectral, and timing cues to allow us to localize sound sources. Localization can be described in terms of three-dimensional position: the azimuth or horizontal angle, the elevation or vertical angle, and the distance (for static sounds) or velocity (for moving sounds). The azimuth of a sound is signalled by the difference in arrival times between the ears, by the relative amplitude of high-frequency sounds (the shadow effect), and by the asymmetrical spectral reflections from various parts of our bodies. The distance cues are the loss of amplitude, the loss of high frequencies, and the ratio of the direct signal to the reverberated signal. Depending on where the source is located, our head acts as a barrier to change the timbre, intensity and spectral qualities of the sound, helping the brain calculate where the sound emanated from. These minute differences between the two ears are known as interaural cues.

We can discern the sound source despite additional reflections at 10 decibels louder than the original wave front, using the earliest arriving wave front. This principle is known as the Haas effect, a specific version of the precedence effect. By making use of these effects and principles, an artificial sound source localization system can be implemented. There are multiple challenges present in the design and implementation phases. A proper algorithm that can be related and can act similar to the proposed idea needs to be found. The hardware implementation includes challenges on choosing apt processors and sensors. Another concern is the optimization of codes and creating test benches for simulation.

Chapter 2

Theory

2.1 Human Auditory system

2.1.1 Binaural cues

Lord Rayleigh's 'duplex theory' [1] was the first comprehensive analysis of the physics of auditory perception, and his theory remains basically valid to this day, with some extensions. As Rayleigh noted, two physical cues dominate the perceived location of an incoming sound source, as illustrated in Fig. 2.1.1. Unless a sound source is located directly in front of or behind the head, sound arrives slightly earlier in time at the ear that is physically closer to the source, and with somewhat greater intensity. This inter aural time difference (ITD) is produced because it takes longer for the sound to arrive at the ear that is farther from the source. The interaural intensity difference (IID) is produced because the 'shadowing' effect of the head prevents some of the incoming sound energy from reaching the ear that is turned away from the direction of the source, especially at higher frequencies.

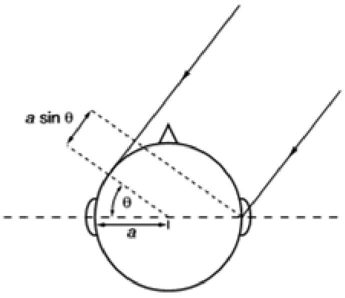

Figure 2.1.1: Inter aural differences of time and intensity impinging on an ideal spherical head from a distant source[1]

The ITD and IID cues operate in complementary ranges of frequencies at least for simple sources in a free field (such as a location outdoors or in an anechoic chamber). Specifically, IIDs are most pronounced at frequencies above approximately 1.5 kHz because it is at those frequencies that the head is large compared to the wavelength of the incoming sound, producing substantial reflection (rather than total diffraction) of the incoming sound wave. Inter aural timing cues, on the other hand, exist at all frequencies, but for periodic sounds they can be decoded unambiguously only for frequencies for which the maximum physically-possible ITD is less than half the period of the waveform at that frequency. Since the maximum possible ITD is about 660 s for a human head of typical size, ITDs are generally useful only for stimulus components below about 1.5 kHz. Note that for pure tones, the term interaural phase delay (IPD) is often used, since the ITD corresponds to a phase difference. If the head had a completely spherical and uniform surface, as in Fig. 1.1, the ITD produced by a sound source that arrives from an azimuth of radians can be approximately described using diffraction theory by the equation

$$\tau = (a/c)2sin\theta \tag{2.1}$$

for frequencies below approximately 500 Hz, and by the equation

$$\tau = (a/c)(\theta + sin\theta) \tag{2.2}$$

for frequencies above approximately 2 kHz. In the equations above, 'a' represents the radius of the head (approximately 87.5 mm) and 'c' represents the speed of sound. The actual values of IIDs are not as well predicted by wave diffraction theory, but they can be measured using probe microphones in the ear and other techniques. They have been found empirically to depend on the angle of arrival of the sound source, frequency, and distance from the sound sources (at least when the source is extremely close to the ear). IIDs produced by distant sound sources can become as large as 25 dB in magnitude at high frequencies, and the IID can become greater still when a sound source is very close to one of the two ears.

2.1.2 Sensitivity to differences in interaural time and intensity

Humans are remarkably sensitive to small differences in interaural time and intensity. For low-frequency pure tones, for example, the just-noticeable difference (JND) for ITDs is on the order of 10 μs, and the corresponding JND for IIDs is on the order of 1dB. The JND for ITD depends on the ITD, IID, and frequency with which a signal is presented. The binaural system is completely insensitive to ITD for narrowband stimuli above about 1.5 KHz, although it does respond to low-frequency envelopes of high-frequency stimuli, as will be noted below. JNDs for IID are a small number of decibels over a broad range of frequencies. Sensitivity to small differences in interaural correlation of broad-band noise sources is also quite acute, as a decrease in interaural correlation from 1 to 0.96 is readily discernable.

2.1.3 Cone of Confusion

A simplification used in one of the projects, which was not found in any recognition methods researched, is the use of a wrist band to remove several degrees of freedom. This enabled three new recognition methods to be devised. The recognition frame rate achieved is comparable to most of the systems in existence (after allowance for processor speed) but the number of different gestures recognised and the recognition accuracy are amongst the best found.

2.1.4 Dynamic binaural cues

If the head is rotated, the ITD and ILD change dynamically, and those changes are different for sounds at different elevations. For example, if an eye-level sound source is straight ahead and the head turns to the left, the sound becomes louder (and arrives sooner) at the right ear than at the left. But if the sound source is directly overhead, there will be no change in the ITD and ILD as

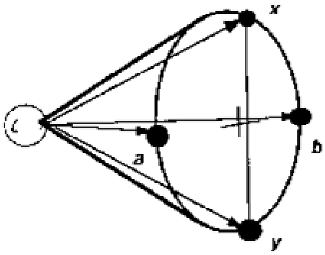

Figure 2.1.2: Cone of confusion

the head turns. Intermediate elevations will produce intermediate degrees of change, and if the presentation of binaural cues to the two ears during head movement is reversed, the sound will be heard behind the listener.

2.1.5 Monaural Cues

The structure of the outer ears (or pinnae) impose further spectral coloration on the signals that arrive at the eardrums. This information is especially useful in a number of aspects of the localization of natural sounds occurring in a free field , including localization in the vertical plane, and the resolution of front-back ambiguities in sound sources. These are known as monaural cues.

2.1.6 Head related Transfer function

Although measurement of and speculation about the spectral coloration imposed by the pinnae have taken place for decades, the 'modern era' of activity in this area began with the systematic and carefully controlled measurements of Wightman and Kistler [1] and others, who combined careful instrumentation with comprehensive psycho acoustical testing. Following procedures developed by Mehrgardt and Mellert [1] and others, Wightman and Kistler and others used probe microphones in the ear to measure and describe the transfer function from sound source to eardrum in anechoic environments. This transfer function is commonly referred to as the head-related transfer function (HRTF), and its time-domain analog is the head-related impulse response (HRIR). A head-related transfer function is a response that characterizes how an ear receives a sound from a point in space; a pair of HRTFs for two ears can be used to synthesize a binaural sound that seems to come from a particular point in space. It is a transfer function, describing how a sound from a specific point will arrive at the ear. To find the sound pressure that an arbitrary source $x(t)$ produces at the ear drum, all we need is the impulse response $h(t)$ from the source to the ear drum. This is called the Head-Related Impulse Response (HRIR), and its Fourier transform H(f) is called the Head Related Transfer Function (HRTF). The HRTF captures all of the physical cues to source localization. Once you know the HRTF for the left ear and the right ear, you can synthesize accurate binaural signals from a monaural source. The HRTF is a surprisingly complicated function of four variables: three space coordinates and frequency. In spherical coordinates, for distances greater than about one meter, the source is said to be in the far field, and the HRTF falls off inversely with range. Most HRTF measurements are made in the far field, which essentially reduces the HRTF to a function of azimuth, elevation and frequency.

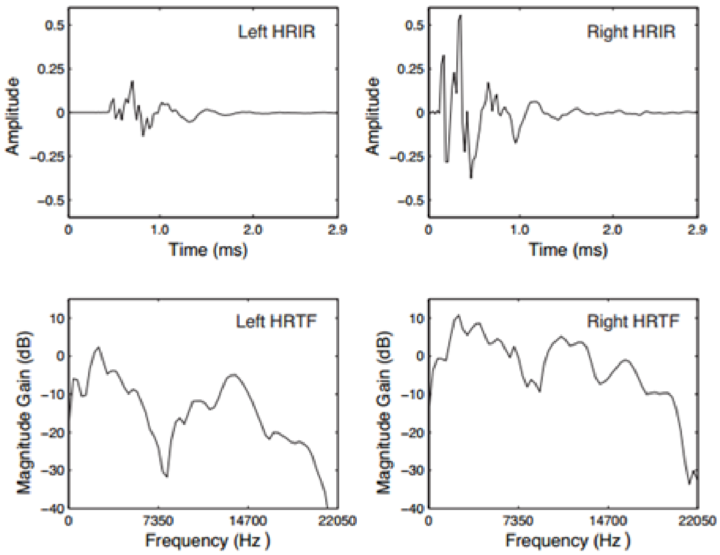

Figure 2.1.3: Head related impulse responses (HRIRs)and the corresponding head-related transfer functions (HRTFs)

2.1.7 Localization of single sources

As noted above, Wightman and Kistler developed a systematic and practical methodology for measuring the HRTFs that describe the transformation of sounds in the free field to the ears. They used the measured HRTFs both to analyze the physical attributes of the sound pressure impinging on the eardrums, and to synthesize 'virtual stimuli' that could be used to present through headphones a simulation of a particular free-field stimulus that was reasonably accurate (at least for the listener used to develop the HRTFs) . These procedures have been adopted by many other researchers. Wightman and Kistler and others have noted that listeners are able to describe the azimuth and elevation of free-field stimuli consistently and accurately. Localization judgments obtained using 'virtual' headphone simulations of the free-field stimuli are generally consistent with the corresponding judgments for the actual free-field signals, although the effect of elevation change is less pronounced and a greater number of front-to-back confusions in location is observed. On the basis of various manipulations of the virtual stimuli, they also conclude that under normal circumstances the localization of free-field stimuli is dominated by ITD information, especially at the lower frequencies, that ITD in formation must be consistent over frequency for it to play a role in sound localization, and that IID information appears to play a role in diminishing the ambiguities that give rise to front-back confusions of position . While the interaural cues for lateralization are relatively robust across subjects, fewer front-to-back and other confusions are experienced in the localization of simulated free-field stimuli if the HRTFs used for the virtual sound synthesis are based on a subject's own pinnae . Useful physical measurements of ITD, IID, and other stimulus attributes can also be obtained using an anatomically realistic manikin such as the popular Knowles Electronics Manikin for Acoustic Research (KEMAR). Examples of HRTFs (and the corresponding HRIRs) recorded from a KEMAR manikin are shown in Fig. 2.1.3. Experimental measurements indicate that localization judgements are more accurate when the virtual source is spatialized using HRTFs obtained by direct measurement in human ears, rather than from an articial head such as the KEMAR. However, there is a strong learning effect, in which listeners adapt to unfamiliar HRTFs over the course of several experimental sessions.

2.1.8 The precedence effect

A further complication associated with judging the location of a sound source in a natural acoustic environment, such as an enclosed room, is that sound is reected from various surfaces before reaching the ears of the listener. However, despite the fact that reections arise from many directions, listeners are able to determine the location of the direct sound quite accurately. Apparently, directional cues that are due to the direct sound (the 'first wave front') are given a higher perceptual weighting than those due to reected sound. The term precedence effect is used to describe this phenomenon .

2.2 Estimation of Time Difference of Arrival (TDOA)

Among the cue available for localization,Time difference of arrival(TDOA) cues are the most commonly used. Most algorithms are therefore developed based on the estimation of TDOA. Some of the TDOA Estimation algorithms are Cross correlation, Generalized Cross correlation Phase transform (GCC-PHAT), Maximum Likelihood (ML) method, Average Square Difference Function(ASDF) method and Least Mean Square (LMS) Adaptive filter method. Of all of these cross correlation is the most basic one and GCC-PHAT is the most widely used method to estimate time delay in the context of source localization and hence these two are mentioned in little more detail in later chapters.

2.2.1 Cross Correlation Algorithm

The first approach to estimate the time delay is to compute the cross correlation between the received signals at two microphones. If $s(t)$ is the source sound signal and $n1(t)$ and $n2(t)$ are the uncorrelated noise signals. Received signals $x1(t)$ and $x2(t)at$ the each of the microphones are given by:

$$x1(t) = s(t) + n1(t) \tag{2.3}$$

$$x2(t) = s(t - T_d) + n2(t) \tag{2.4}$$

Cross correlation of $x1(t)$ and $x2(t)$ is given by

$$R_{x1.x2}(\tau) = \int_{-\infty}^{\infty} x1(t).x1(t+\tau)dt \tag{2.5}$$

$$R_{x1.x2}(\tau) = \int_{-\infty}^{\infty} s(t).s(t - T_d + \tau)dt \tag{2.6}$$

Peaks of the cross correlation plot gives the Time difference of arrival (T_d).The location of the maximum peak cross correlation result represents the estimated time delay(T_d). The sound source direction is then given by $\theta = cos^{-1}(\frac{c\tau}{F_s d})$ in reference to the figure 2.2.1

A 'basic' frequency domain cross-correlator does an FFT on each signal, conjugates one of them, multiplies one versus the other, then inverse transforms. A peak in the result indicates a time delay. There are a lot of problems with this simple approach. For instance: a linear cross-correlation is needed instead of a circular one due to strong sinusoidal interference which will generates multiple peaks, or very low SNR, the noise may not be as expected etc.

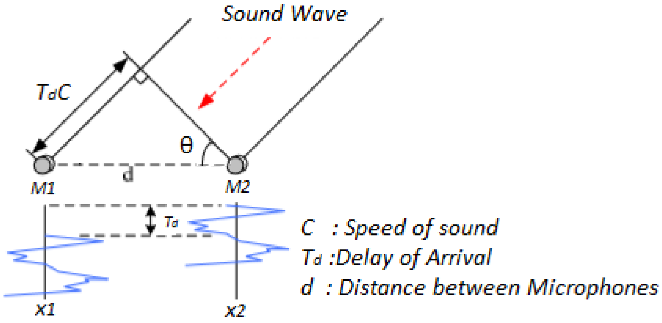

Figure 2.2.1: TDOA estimation

2.2.2 Generalized Cross Correlation(GCC)

A way to sharpen the cross correlation peak is to whiten the input signals by using weighting function, which leads to the so-called generalized cross-correlation technique (GCC). The block diagram of a generalized cross-correlation processor is shown in Figure 2.2.2. The procedure of GCC has received considerable attention due to its ability to avoid spreading of the peak of the correlation function.

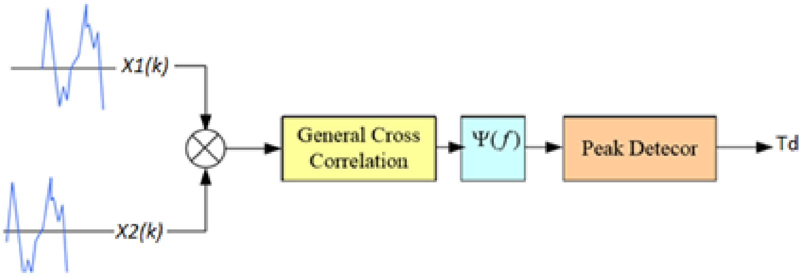

Figure 2.2.2: Estimation of TDOA by GCC

2.2.3 Stream processing

Real-time sound signal processing is done using Stream Processing since data is continuosly being received. Stream processing is a memory-efficient technique used for handling large amounts of data. Stream processing divides incoming data into frames and fully processes each frame before the next one arrives as in Figure 2.2.2.

The just-in-time and memory-sensitive nature of stream processing presents special challenges. Streaming algorithms must be efficient and keep up with the rate of data updates. To handle large data sets, the algorithms must also manage memory and state information, store previous data buffers only as needed, and update each buffer and state frame-by-frame. Algorithm components called System objects simplify stream processing in MATLAB. System objects provide a workflow for developing streaming algorithms and test benches for a range of streaming applications, which involve just a few lines of MATLAB code

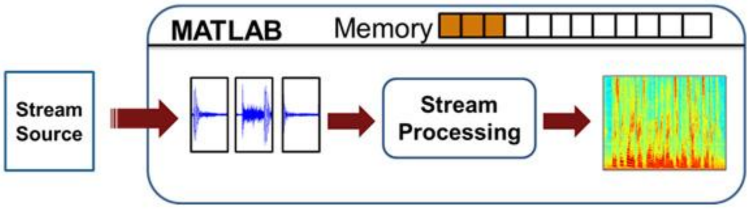

Figure 2.2.3: Stream processing in MATLAB

2.2.4 Accelerating algorithm execution

(i) **Optimize MATLAB code**: This can be done by making use of techniques like pre allocation and vectorization. Preallocation solves a basic problem in simple program loops, where an array is iteratively enlarged with new data (dynamic array growth). Unlike other programming languages (such as C, C++, C or Java) that use static typing, Matlab uses dynamic typing. This means that it is natural and easy to modify array size dynamically during program execution.The basic idea of preallocation is to create a data array in the final expected size before actually starting the processing loop. This saves any reallocations within the loop, since all the data array elements are already available and can be accessed. This solution is useful when the final size is known in advance

(ii) **Using system objects**: System Toolboxes include System objects and most System Toolboxes also have MATLAB functions and Simulink blocks. System objects process frames and then overwrite past frames with incoming data, initialize parameters only once as they are created, manage buffer updates, state updates, and indexing automatically, support MATLAB code generation and parallel computing workflows all which accelerate algorithm development.

(iii) **MATLAB to C**: Generating MEX files (MatlabEXecutable files) automatically with MATLAB coder or converting MATLAB code to C manually(depends on algorithm complexity).

2.3 Microphones

A microphone is a transducer that converts sound into an electrical signal. Microphones are used in many applications such as telephones, hearing aids, public address systems for concert halls and public events, motion picture production, live and recorded audio engineering, two-way radios, megaphones, radio and television broadcasting, and in computers for recording voice, speech recognition, VoIP, and for non-acoustic purposes such as ultrasonic checking or knock sensors. Most microphones today use electromagnetic induction (dynamic microphones), capacitance change (condenser microphones) or piezo-electricity (piezoelectric microphones) to produce an electrical signal from air pressure variations. Microphones typically need to be connected to a preamplifier before the signal can be recorded or reproduced. The sensitive transducer element of a microphone is called its element or capsule. Sound is first converted to mechanical motion by means of a diaphragm, the motion of which is then converted to an electrical signal. A complete microphone also includes a housing, some means of bringing the signal from the element to other equipment, and often an electronic circuit to adapt the output of the capsule to the equipment being driven. A wireless microphone contains a radio transmitter. Microphones categorized by their transducer principle, such as condenser, dynamic, etc., and by their directional characteristics. Sometimes other characteristics such as diaphragm size, intended use or orientation of the principal sound input to the principal axis (end- or side-address) of the microphone are used to describe the microphone.

2.3.1 Condenser Microphone

The condenser microphone, invented at Bell Labs in 1916 by E.C.Wente, is also called a capacitor microphone or electrostatic microphone, capacitors were historically called condensers, hence the name condenser microphone. Here, the diaphragm acts as one plate of a capacitor, and the vibrations produce changes in the distance between the plates. There are two types, depending on the method of extracting the audio signal from the transducer: DC-biased microphones, and radio frequency (RF) or high frequency (HF) condenser microphones. With a DC-biased microphone, the plates are biased with a fixed charge (Q). The voltage maintained across the capacitor plates changes with the vibrations in the air, according to the capacitance equation ($C = QV$), where Q = charge in coulombs, C = capacitance in farads and V = potential difference in volts. The capacitance of the plates is inversely proportional to the distance between them for a parallel-plate capacitor. The assembly of fixed and movable plates is called an "element" or "capsule". A nearly constant charge is maintained on the capacitor. As the capacitance changes, the charge across the capacitor does change very slightly, but at audible frequencies it is sensibly constant.

2.3.2 Dynamic Microphone

Dynamic microphones (also known as moving-coil microphones) work via electromagnetic induction. They are robust, relatively inexpensive and resistant to moisture. This, coupled with their potentially high gain before feedback, makes them ideal for on-stage use. Dynamic microphones use the same dynamic principle as in a loudspeaker, only reversed. A small movable induction coil, positioned in the magnetic field of a permanent magnet, is attached to the diaphragm. When sound enters through the windscreen of the microphone, the sound wave moves the diaphragm. When the diaphragm vibrates, the coil moves in the magnetic field, producing a varying current in the coil through electromagnetic induction.

2.3.3 MEMS Microphone

The MEMS (Micro Electrical-Mechanical System) microphone is also called a microphone chip or silicon microphone. A pressure-sensitive diaphragm is etched directly into a silicon wafer by MEMS processing techniques, and is usually accompanied with integrated preamplifier. Most MEMS microphones are variants of the condenser microphone design. Digital MEMS microphones have built in analog-to-digital converter (ADC) circuits on the same CMOS chip making the chip a digital microphone and so more readily integrated with modern digital products. Major manufacturers producing MEMS silicon microphones are Wolfson Microelectronics (WM7xxx) now Cirrus Logic, InvenSense (product line sold by Analog Devices), Akustica (AKU200x) etc. Recently, there has been increased interest and research into making piezoelectric MEMS microphones which are a significant architectural and material change from existing condenser style MEMS designs.

2.3.4 Factors affecting selection of Microphone

There are a number of models and choices when deciding to purchase a microphone and preamplifier system to measure sound or unwanted sound, called noise. In some cases, multiple microphone and preamplifier models could be used for the same application for improved system accuracy. The basic factors that govern selection of microphones for an application are sensor properties (acoustic, accels, pressure), power supplies, cables and the environment for testing which are

- Indoors or outdoors

- Duration of test
- Background noise, and frequencies of background noise
- Humidity of the test area
- Temperature in test area
- Test object location and sensor positioning
- Location of the test object
- Location and positioning of the sensor relative to test object
- Characteristics of surrounding objects
- Minimum and maximum frequency and amplitudes required
- Preference for pre-polarized or externally polarized microphones
- Budget for microphones and the cost per channel of the system

2.4 Estimation of Location

2.4.1 Microphone Array Geometry

A three-element-two-dimensional microphone is taken for preliminary analysis. The array consists of three microphones arranged in an 'L' fashion in a 2-dimensional plane. As shown in the fig 2.4.1 the microphones M3-M1-M2 form the array with M1 being the center microphone. M1 is at the origin of the coordinate axis, M1-M2 form the x-axis, M1-M3 the y-axis. The distances between adjacent microphones is d. The angle of arrival θ_1 is measured in clockwise direction w.r.t the line perpendicular to M1-M2 axis and passing through the mid point of the axis. θ_2 is measured in counter clockwise direction w.r.t the line perpendicular to M1-M3 axis and passing through the mid point of the axis. This convention is chosen for experimental convenience. It is assumed that microphones are omnidirectional and also that sound source is present in the first quadrant.

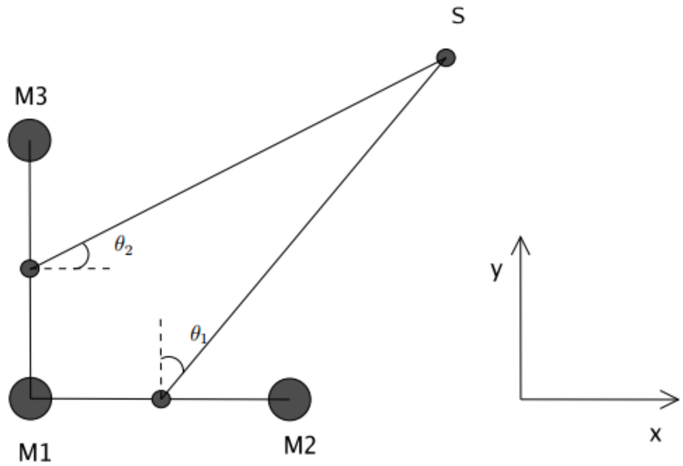

Figure 2.4.1: Microphone Array Geometry

Now we derive a relationship between the frequency content of the incident signal and the maximum allowed separation between each pair of microphones in the array. The maximum phase difference is restricted to π. Any phase difference out of the range of $-\pi$ and π is wrapped around to within this range. This places an important restriction on the array geometry when performing DOA estimation.

Consider a signal incident on the pair of microphones as shown in Fig. 2.4.1 at an angle θ. Let the broad band signal have a maximum frequency of f_{max}. At f_{max}, if we restrict the phase difference between signals of pair of microphones to be less than or equal to π, then we require

$$2\pi f_{max}\tau \leq \pi \tag{2.7}$$

and

$$\tau = \frac{d\sin\theta}{v} \tag{2.8}$$

where
τ = signal time delay between the two microphones,
d = distance between the pair of microphones,
θ = incident angle,
v = velocity of sound.

Rearranging these terms, we have

$$d \leq \frac{1}{2}\left(\frac{v}{f_{max}}\right)\frac{1}{\sin\theta} \tag{2.9}$$

Since $(\sin\theta)_{max} = 1$ and $\frac{v}{f_{max}}$ is same as λ_{min}, the minimum wave length present in the signal

$$d \leq \frac{\lambda_{min}}{2} = \frac{v}{2f_{max}} \tag{2.10}$$

which means that the distance between any pair of microphones in the array should not exceed half the minimum wavelength present in the signal. This condition becomes very important when we perform TDE from phase difference estimates of the signals.

2.4.2 Source Localisation in 2D Space

The object localization requires three microphones in a 2D plane since two coupled microphones can give only one information i.e., direction of source. The acoustic waves generated by the source reaches the first microphone earlier than the second. The difference in the propagation delay and that the acoustic velocity in air is known, we calculate the path difference of the acoustic waves. By definition, a hyperbola is the set of all points in the plane whose location is characterized by the fact that the difference of their distance to two fixed points is a constant. The two fixed points are called the foci. In our case the foci are the microphones. Each hyperbola consists of two branches. The emitter is located on one of the branches. The line segment which connects the two foci intersects the hyperbola in two points, called the vertices. The line segment which ends at these vertices is called the transverse axis and the midpoint of this line is called the center of the hyperbola.

The time-delay of the sound arrival gives us the path difference that defines a hyperbola on one branch of which the emitter must be located. At this point, we have infinity of solutions since we have single information for a problem that has two degrees of freedom. We need to have a third microphone, when coupled with one of the previously installed microphones, it gives a second hyperbola. The intersection of one branch of each hyperbola gives one or two solutions with atmost of four solutions being possible. Since we know the sign of the angle of arrivals, we can remove the ambiguity.

Hyperbolic Position Location

Hyperbolic position location (PL) estimation is accomplished in two stages. The first stage involves estimation of the time difference of arrival (TDOA) between the sensors (microphones) through the use of time-delay estimation techniques. The estimated TDOAs are then utilized to make range difference measurements. This would result in a set of nonlinear hyperbolic range difference equations. The second stage is to implement an efficient algorithm to produce an unambiguous solution to these nonlinear hyperbolic equations.The solution produced by these algorithms result in the estimated position location of the source. Accurate position location estimation of a source requires an efficient hyperbolic position location estimation algorithm. Once the TDOA information has been measured, the hyperbolic position location algorithm will be responsible for producing an accurate and unambiguous solution to the position location problem. Algorithms with different complexity and restrictions have been proposed for position location estimation based on TDOA estimates. When the microphones are arranged in non-collinear fashion, the position location of a sound source is determined from the intersection of hyperbolic curves produced from the TDOA estimates. The set of equations that describe these hyperbolic curves are nonlinear and are not easily solvable. If the number of nonlinear hyperbolic equations equals the number of unknown coordinates of the source, then the system is consistent and a unique solution can be determined from iterative techniques. For an inconsistent system, the problem of solving for the position location of the sound source becomes more difficult due to non-existence of a unique solution.
Let (x, y) be the source location and (X_i, Y_i) be the known location of the i^{th} microphone. The squared range difference between the source and the i^{th} microphone is given as

$$r_i = \sqrt{(X_i - x)^2 + (Y_i - y)^2} \tag{2.11}$$

$$r_i = \sqrt{X_i^2 + Y_i^2 - 2X_i x - 2Y_i y + x^2 + y^2} \tag{2.12}$$

Also, TDOA between the i^{th} and M1 using range difference equations can be wriiten as

$$r_{i,1} = v\tau_{i,1} = r_i - r_1 \tag{2.13}$$

For a three microphone system, producing two TDOA's, the unknown location coordinates x and y can be solved in terms of r_1 for $i = 1$. The soluution is in the form of

$$\begin{bmatrix} x \\ y \end{bmatrix} = -\begin{bmatrix} x_{2,1} & y_{2,1} \\ x_{3,1} & y_{3,1} \end{bmatrix}^{-1} \times \begin{bmatrix} r_{2,1} \\ r_{3,1} \end{bmatrix} r_1 + \frac{1}{2}\begin{bmatrix} r_{2,1}^2 - K_2 + K_1 \\ r_{3,1}^2 - K_3 + K_1 \end{bmatrix} \tag{2.14}$$

where

$$K_1 = X_1^2 + Y_1^2$$
$$K_2 = X_2^2 + Y_2^2$$
$$K_3 = X_3^2 + Y_3^2$$

When the equation (2.14) is inserted into equation above, with $i = 1$, a quadratic equation in terms of r_1 is produced. Substituting the positive root back into the equation (2.14) results in the final solution. Two positive roots may exist from the quadratic equation that can produce two different solutions, resulting in an ambiguity. This problem has to be resolved by using a priori information.

Chapter 3

Hardware Details

Hardware details includes the descriptions about the block diagrams, circuit diagrams and the components that are used in this project.

3.1 Block Diagram

The block diagram of the prototype microphone array system is depicted in the figure. The system is divided into three modules by function: microphone array (Module 1),pre-processing and sampling module (Module 2: PS) ,real-time processing or data acquisition module (Module 3: P/A). The microphone array is a uniform circular array with four condenser microphones. The microphone signals are filtered After preprocessing of synchronized filters and amplifiers, simultaneous sampling ADCs are used to capture signals from the microphones. The synchronized filters and amplifiers mean that a strict demand on the consistency of the four channels is requested.

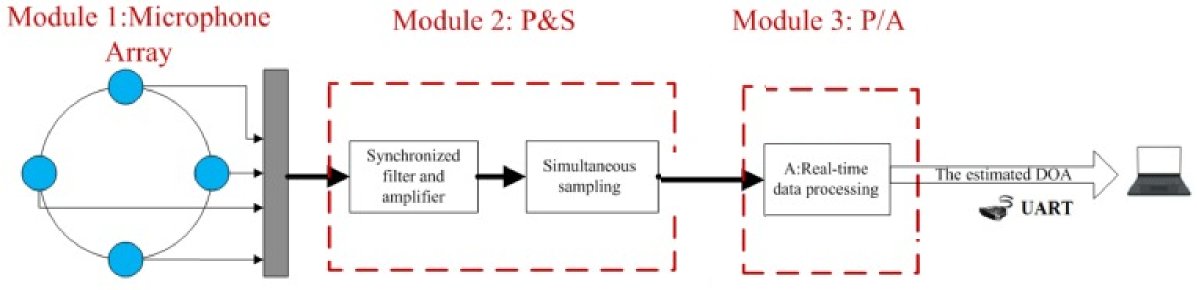
Module 1:Microphone Array
Module 2: P&S
Module 3: P/A
Synchronized filter and amplifier
Simultaneous sampling
A:Real-time data processing
The estimated DOA
UART

Figure 3.1.1: Block Diagram of the microphone array system

3.2 Processing Hardware

3.2.1 TMS320C6713 DSP Starter Kit

The TMS320C6713 DSP Starter Kit (DSK) is the core hardware unit of this project. The TMS320C6713 DSP Starter Kit (DSK) developed jointly with Spectrum Digital is a low-cost development platform designed to speed the development of high precision applications based on TIs TMS320C6000 floating point DSP generation. Both experienced and novice designers can get started immediately

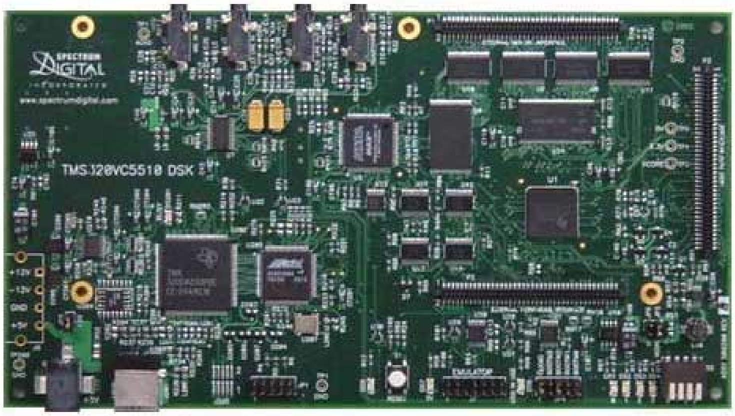

Figure 3.2.1: TMS320C6713 DSK

with innovative product designs with the DSKs full featured Code Composer Studio^{TM} IDE and eXpressDSP^{TM} Software which include DSP/BIOS and Reference Frameworks. The C6713 DSK tools includes the latest fast simulators from TI and access to the Analysis Toolkit via Update Advisor which features the Cache Analysis tool and Multi-Event Profiler. Using Cache Analysis developers improve the performance of their application by optimizing cache usage. By providing a graphicalview of the on-chip cache activity over time the user can quickly determine if their code is using the on-chip cache to get peak performance.The C6713 DSK allows you to download and step through code quickly and uses Real Time Data Exchange (RTDX^{TM}) for improved Host and Target communications. The DSK includes the Fast Run Time Support libraries and utilities such as Flash burn to program flash, Update Advisor to download tools, utilities and software and a power on self test and diagnostic utility to ensure the DSK is operating correctly.

Some key features of the C6713 DSK are

- A Texas Instruments TMS320C6713 DSP operating at 225 MHz
- An AIC23 stereo codec
    - 8-96 Khz sampling rates
    - 16-32 bit quantisation
    - Two audio IN and two audio OUT channels
- 16 Mbytes of synchronous DRAM
- 512 Kbytes of non-volatile Flash memory (256 Kbytes usable in default configuration)
- 4 user accessible LEDs and DIP switches
- Software board configuration through registers implemented in CPLD
- Configurable boot options
- Standard expansion connectors for daughter card use
- JTAG emulation through on-board JTAG emulator with USB host interface or external emulator

- Single voltage power supply (+5V)

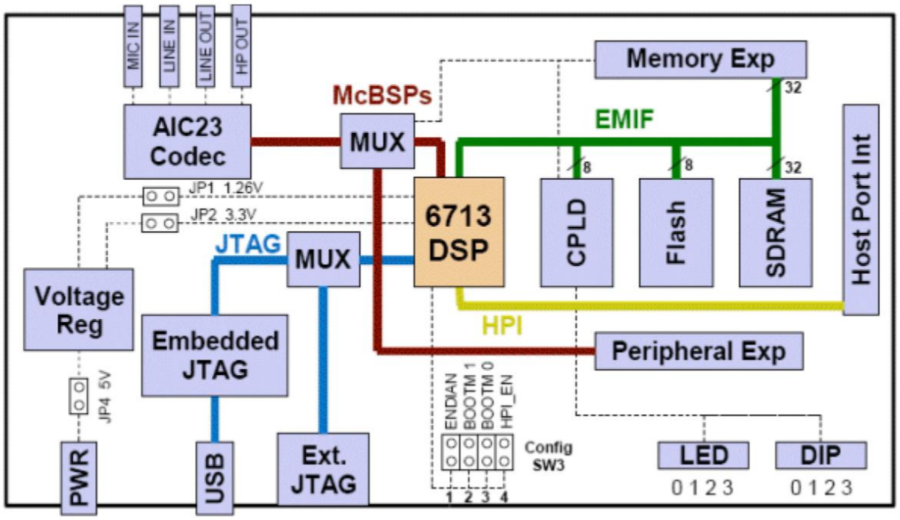

Figure 3.2.2: C6713 DSK's block diagram

3.2.2 Functional Overview of DSK

The DSP on the 6713 DSK interfaces to on-board peripherals through a 32-bit wide EMIF (External Memory InterFace). The SDRAM, Flash and CPLD are all connected to the bus. EMIF signals are also connected daughter card expansion connectors which are used for third party add-in boards.
The DSP interfaces to analog audio signals through an on-board AIC23 codec and four 3.5 mm audio jacks (microphone input, line input, line output, and headphone output). The codec can select the microphone or the line input as the active input. The analog output is driven to both the line out (fixed gain) and headphone (adjustable gain) connectors. McBSP0 is used to send commands to the codec control interface while McBSP1 is used for digital audio data. McBSP0 and McBSP1 can be re-routed to the expansion connectors in software.
A programmable logic device called a CPLD is used to implement glue logic that ties the board components together. The CPLD has a register based user interface that lets the user configure the board by reading and writing to its registers. The DSK includes 4 LEDs and a 4 position DIP switch as a simple way to provide the user with interactive feedback. Both are accessed by reading and writing to the CPLD registers.
An included 5V external power supply is used to power the board. On-board switching voltage regulators provide the +1.26V DSP core voltage and +3.3V I/O supplies. The board is held in reset until these supplies are within operating specifications. Code Composer communicates with the DSK through an embedded JTAG emulator with a USB host interface. The DSK can also be used with an external emulator through the external JTAG connector

3.2.3 TMS320C6713 DSP Features

The DSK features the TMS320C6713 DSP, a 225 MHz device. The C6713 device is based on the high-performance, advanced very-long-instruction-word (VLIW) architecture developed by Texas Instruments (TI). Operating at 225 MHz, the C6713 delivers up to 1350 million floating-point operations per second (MFLOPS), 1800 million instructions per second (MIPS), and with dual fixed/floating-point multipliers up to 450 million multiply-accumulate operations per second (MMACS).

Some key features of the processor are

- 32 bit registers
- Harvard architecture
- 8 functional units
  - Two fixed point ALU's
  - Four floating point ALU's
  - Two multipliers
- Load store architecture
- 256 bit instruction fetch in one cycle
- 4.4-, 5-, 6- instruction cycle times

3.3 Microphone

Condenser microphones are the most common types of microphones found. They have a much greater frequency response and transient response - which is the ability to reproduce the ”speed” of an instrument or voice. They also generally have a louder output, but are much more sensitive to loud sounds.Condenser microphones are generally much more expensive than dynamic microphones, but many cheap condensers also exist for smaller applications. Also structural advantage exists for condenser microphone since they do not use heavy coils as in dynamic microphones.

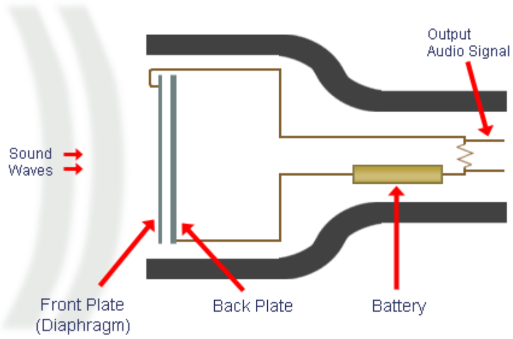

Figure 3.3.1: Condenser microphone

Chapter 4

Software Details

4.1 Flowchart

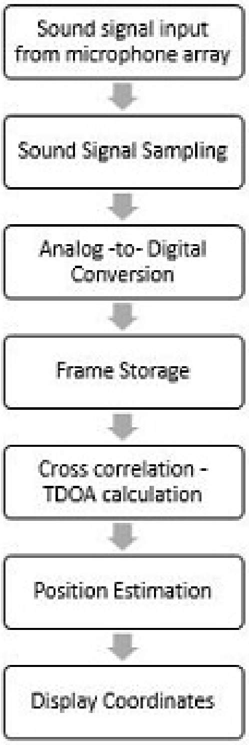
Sound signal input from microphone array
Sound Signal Sampling
Analog -to- Digital Conversion
Frame Storage
Cross correlation - TDOA calculation
Position Estimation
Display Coordinates

Figure 4.1.1: Flow Chart

The four parallel sound signals obtained from the microphone array is divided and stored into respective frames for real-time computation. Cross correlation is done on the frame pairs to obtain six different values of TDOA. Coordinates of the incoming sound source is estimated using TDOA values and microphone array dimensions. The coordinates obtained are then displayed on a user interface.

4.2 Code Composer Studio

Code Composer Studio is an integrated development environment (IDE) that supports TI's Microcontroller and Embedded Processors portfolio. Code Composer Studio comprises a suite of tools used to develop and debug embedded applications. It includes an optimizing C/C++ compiler, source code editor, project build environment, debugger, profiler, and many other features. The intuitive IDE provides a single user interface taking you through each step of the application development flow. Familiar tools and interfaces allow users to get started faster than ever before. Code Composer Studio combines the advantages of the Eclipse software framework with advanced embedded debug capabilities from TI resulting in a compelling feature-rich development environment for embedded developers.

Code Composer Studio features for the TMS320C6713 DSK include:

- A complete Integrated Development Environment (IDE), an efficient optimizing C/C++ compiler assembler, linker, debugger, an advanced editor with Code Maestro^{TM} technology for faster code creation, data visualization, a profiler and a flexible project manager.
- DSP/BIOS^{TM} real-time kernel
- Target error recovery software
- DSK diagnostic tool
- Ability for third-party software for additional functionality

4.3 Matlab

MATLAB (matrix laboratory) is a multi-paradigm numerical computing environment and fourth-generation programming language. A proprietary programming language developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, C, Java, Fortran and Python.
It provides capabilities for the numerical solution of linear and nonlinear problems, and for performing other numerical experiments. It also provides extensive graphics capabilities for data visualization and manipulation.

4.4 Software implementation

Signal processing algorithms and mathematical equations are easily and efficiently implemented in Matlab from a user's perspective. From the processing point of view C provides a much better performance. Code Composer Studio only has a C compiler to convert high level language to the target C6713 processor's instructions. Therefore Matlab code cannot be directly executed on C67 processors. Matlab Coder is used to convert Matlab code to equivalent C code. C compiler then converts C code to a C6713 executable code. The program is burnt via the JTAG/USB interface

Chapter 5

Applications

Sound Source localization systems are being utilized in many different areas from consumer electronics to military systems. Noise reduction, speech recognition, hands-free communication, automatic video camera steering and multiparty teleconferencing are some of the applications being studied.

5.1 Sound source separation

Sound source separation systems make use of the spatial location of the sound source. One such system which is widely used is known as Beamforming. The term Beamforming refers to the design of a spatio-temporal filter which operates on the outputs of the microphone array. This spatial filter can be viewed in terms of dependence upon angle and frequency. Beamforming is achieved by filtering the microphone signals and combining the outputs to extract (by constructive combining) the desired signal and reject (by destructive combining) interfering signals according to their spatial location. Beamforming for broadband signals like speech can, in general, be performed in the time domain or frequency domain. The simplest deterministic Beam-forming technique is delay-and-sum Beamforming, where the signals at the microphones are delayed and then summed in order to combine the signal arriving from the direction of the desired source coherently, expecting that the interference components arriving from off the desired direction cancel to a certain extent by destructive combining. The delay-and-sum Beamformer as shown in Figure 3.1 is simple in its implementation and provides for easy steering of the beam towards the desired source. Assuming that the broadband signal can be decomposed into narrowband frequency bins, the delays can be approximated by phase shifts in each frequency band.
The performance of the delay-and-sum Beam-former in reverberant environments is often insufficient. A more general processing model is the filter-and-sum Beam-former as shown in Figure 3.2.

Where, before summation, each microphone signal is filtered with FIR filter of order M. This structure, designed for multipath environments, namely reverberant enclosures, replaces the simpler delay compensator with a matched filter.

5.2 Surveillance Applications

Video analytics used in surveillance applications performs well in normal conditions. But it may not work as accurately under adverse circumstances. Taking advantage of the complementary aspects of video and audio can lead to a more effective analytics framework resulting in increased system robustness. For example, sound scene analysis may indicate potential security risks outside field-of-view, pointing the camera in that direction. Many video analytics solutions used in surveillance applications to identify security-risk events perform relatively well in normal conditions. They might not, however, detect emergency events accurately in cases of view obstruction, low or rapidly changing lighting, out-of view activities, or other adverse conditions (e.g. rain, fog, smoke). In such cases, audio analytics can be used to provide additional information about the environment under surveillance. Audio analytics can analyze the sound scene of a surveyed environment and provide additional data about activities not readily discerned by a camera. Sound identification may alert to potential security risks and sound localization may be used to point the camera in the direction of interest. Taking advantage of the complementary aspects of video and audio can provide a powerful framework that should lead to increased system robustness and positive alarm detection rate. A potential application of sound localization lies in Gunshot detection systems like 'ShotSpotter'. ShotSpotter systems consist of highly sensitive microphones installed all around a large area to track the instance of gunshot. ShotSpotter gunfire data enables intelligent analysis. With that, law enforcement can move from the reactive to the proactive. ShotSpotter has been called "a force multiplier" because it provides critical information for better, more timely resource allocation.

5.3 Robot Audition System

Robot audition is an active research area which realizes natural human-robot interaction in a daily environment. The main claim in robot audition is listening to several things simultaneously using a robot's own ears. However, various types of sound sources coexist with a target speech source. Thus, one of the hottest research topics in robot audition is sound source separation and speech enhancement.

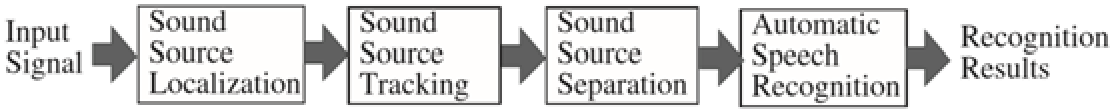

Figure 5.3.1: Robot audition system flow

The basic processing flow of a typical robot audition system is shown in Figure 5.3.1. There are mainly three processing blocks such as Sound Source Localization (SSL), Sound Source Tracking (SST) and Sound Source Separation (SSS) before an ASR block. These preprocessing blocks should consider moving sound source situations. For SSL and SST, some studies mentioned mobile functions of robot audition, that is, low-level active audition.

5.4 Multi Speaker Teleconferencing

Attending to multiple speakers in video conferencing setting is a complex task. From a visual point of view, multiple speakers at different locations present radically different appearances. The camera used to capture the visuals of the speaker can utilize the localization cues to direct the field of view of the camera to the active speaker. From an audio point of view, multiple speakers may be speaking at the same time, and background noise may make it difficult to localize sound source without some a priori estimate of sound source locations. A directional audio system based on beam-forming is used to confirm potential speakers and attend to them.

Chapter 6

Conclusion and future scope

The proposed system can only deal with single sound source accurately.But in practical situations there can be multiple sound sources.This system could be later developed to localize all the individual sound sources.This system can be improved by integrating features like sound source isolation and enhancement to achieve directional filtering.

Bibliography

[1] D. Wang and G. Brown,*Computational auditory scene analysis*, 1st ed. Hoboken, N. J.: Wiley-Interscience, 2006, pp. 147-185.

[2] J. Stachurski, L. Netsch, R. Cole, “Sound Source Localization for Video Surveillance Camera”, *10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, Krakow, 2013, pp. 93-98.

[3] F. Keyrouz, ”Advanced Binaural Sound Localization in 3-D for Humanoid Robots”, *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 9, pp. 2098-2107, 2014.

[4] H. Adel et al.,”Beamforming Techniques for Multichannel audio Signal Separation”, *International Journal of Digital Content Technology and its Applications*, vol. 6, no. 20, pp. 659-667, 2012.

[5] Petr Dostálek, Jan Dolinay ,Vladimír Vašek ,”Embedded system for audio source localization based on beamforming “, *International Journal of Circuits, systems and signal processing*, vol. 6, no. 6, pp. 367-375, January 2012.

[6] A. K. Tellakula, ”Acoustic Source Localization Using Time Delay Estimation”, M.S. thesis, IISc, Bangalore, August 2007.