LOCALIZATION OF MULTIPLE SOUND SOURCES IN THREE DIMENSIONAL ENVIRONMENTS

by

Murat Engin Ünal

B.S., Industrial Engineering, İstanbul Technical University, 2002

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Systems and Control Engineering

Boğaziçi University

ACKNOWLEDGEMENTS

I am grateful to my thesis advisor Prof. Dr. Fikret Gürgen who guided me throughout my thesis work, shows patience to my faults, and never looses his belief on my work. I also want to thank to "Mr. Piano" recording studio, for helping me to make my first set of experiments, and Can Karadoğan, research assistant at İTÜ MİAM who had never hesitated to help me on whatever makes my second set of experiments go in an easy and smooth way. My colleagues also were quite helpful by sharing my work load and helping me focusing on my thesis, especially to Nüfer Yasin Ateş, who spent a lot of time with me while we were working together, kept me concentrated and motivated. My special thanks go to my elder sister Elif Ünal, who was like a professional advisor, motivating, guiding, and checking my errors all the time, a truly magnificent psychological coach.

ABSTRACT

LOCALIZATION OF MULTIPLE SOUND SOURCES IN THREE DIMENSIONAL ENVIRONMENTS

Localization of sound sources has several applications like teleconferencing, speech recognition, speaker identification, speech acquisition in an automobile environment, sound capture in reverberant enclosures, large room recording-conferencing, and hearing aid devices. One way of finding the location of a sound source is to utilize the direction of arrival (DOA) values. This value indicates the angle between two lines, first of which connects the mid point of the microphone array and the sound source, and second of which carries the microphone array. DOA values are usually estimated by first estimating the time delay of arrival (TDOA) value of the signals received from two microphones and then converting TDOA estimates to DOA estimates. TDOA value is estimated using the cross-power spectrum phase (CSP) coefficients.

The existence of multiple sound sources in an environment brings two complications to this problem. First of all, the correlation between the sound sources distorts the TDOA estimation procedure. Second, multiple DOA estimates, which are calculated for multiple microphone arrays, have to be matched to the sound sources to find the correct intersection points among multiple ones. Another major complication is the generalization of the environment to the three dimensional space.

In this thesis, a new method is proposed, for localizing multiple sound sources in three dimensional environments. The synchronous addition of CSP coefficients method is utilized for finding the undistorted DOA estimates. Then these estimates are clustered using a new design-specific, an inconsistency measure based clustering algorithm. Having found DOA estimate triples for each sound source, the location of sound sources are determined by finding the intersection point of three cones formed by three DOA values. This intersection is found by first finding a closed formula, which consists of a single

variable, for a locus of the intersection, which is a three dimensional path and then finding the suitable point on this path.

Experiments are done on simulation and real acoustical environments. The results are promising considering the complexities that the algorithm faced, and the number of the microphones that are available for the experiments, and the diversity of them.

ÖZET

ÜÇ BOYUTLU ORTAMLARDA BULUNAN ÇOK SAYIDAKİ SES KAYNAĞININ YERLERİNİN TESPİTİ

Ses kaynağı yeri belirlenmesinin telekonferans, konuşma tanıma, konuşmacı belirleme, otomotiv ortamında konuşma sinyali alma, yankılı ortamlarda ses yakalama, büyük odalarda ses kaydı ve işitme cihazı tasarlama gibi birçok uygulaması bulunmaktadır. Ses kaynağının yerini bulmanın bir yolu geliş yönü değerlerini kullanmaktır. Bu değer, mikrofon dizisinin orta noktası ve ses kaynağı arasındaki doğru ile mikrofon dizisini taşıyan doğru arasındaki açıyı ifade etmektedir. Geliş yönü değerleri genellikle, ilk olarak iki farklı mikrofon tarafından algılanan sinyallerin geliş zaman gecikmesi değerlerinin tahmin edilmesi, sonra da bu değerin diğerine dönüştürülmesi ile belirlenir. Geliş zaman gecikmesi değeri ise çapraz-güç tayfı fazı katsayıları ile tahmin edilir.

Bu tezde, üç boyutlu ortamlarda bulunan çok sayıdaki ses kaynağının yerlerini tespit edebilecek yeni bir yöntem önerilmiştir. İlk olarak, CSP katsayılarının eşzamanlı bir biçimde toplanması yöntemi ile DOA tahminleri bulunmuştur. Daha sonra bu tahminler, mikrofon dizisi tasarımına bağlı olarak çalışan ve bir tutarsızlık ölçüsünü esas alan yeni bir yaklaşımla gruplandırılmıştır. Her bir ses kaynağı için DOA tahmini üçlüleri bulunduktan sonra bunların yerleri üç koninin kesişim noktalarını bulmak için geliştirilen, tek değişkene bağlı bir fonksiyon üzerinde çalışan bir arama metodu yardımıyla bulunur.

Ortamda birden fazla ses kaynağının bulunması bu probleme iki zorluk getirmektedir. Birincisi, ses kaynakları arasındaki ilgileşimin geliş zamanı değerlerinin tahmini sürecini saptırması, İkincisi ise, tüm kesişim noktaları arasından doğru kesişimleri bulmak için çoklu mikrofon dizileri tarafından hesaplanan çoklu geliş yönü tahminlerinin ses kaynakları ile eşleştirilmesi zorunluluğudur. Bir başka önemli zorluk ise ortamın üç boyuta genellenmiş olmasıdır.

Deneyler hem benzetim hem de gerçek akustik ortamda gerçekleştirilmiştir. Sonuçlar, algoritmanın karşılaştığı karmaşıklıklar, deneylerde kullanılabilen mikrofon sayısı ve bunları çeşitliliği göz önünde bulundurulduğunda oldukça umut vericidir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS / ABBREVIATIONS

α_i	Half of the vertex angle of the cone i
τ_{ij}	Time delay of arrival estimate
a	Steering vector
nD	Number of dimensions
R_{xy}	Cross correlation between signals x and y
tn_i	Vertex location of the cone i
C(.)	Combination operator
CSP	Cross-power Spectrum Phase
DFT(.)	Discrete Fourier Transform
DOA	Direction of Arrival
env	Environment
GMM	Gaussian Mixture Model
ICA	Independent Component Analysis
MUSIC	Multiple Signal Classification
nsc	Number of sound sources
TOF	Time of Flight
TDOA	Time Delay of Arrival

1. INTRODUCTION

Head-mounted and desk-stand microphones are widely used as the input transducer for the acquisition of speech data. A steerable microphone array has numerous advantages over these traditional sound acquiring systems such as improving the speech signal quality by electronically steering the array and focusing on the desired speaker, and removing the clutter around the speaker, leaving him/her a more comfortable space. A list of potential abilities of a steerable microphone array, that a single microphone is not capable of, includes automatic detecting, locating and tracking of active talkers in its environment. Researchers have been studying on other applications like teleconferencing, speech recognition, speaker identification, speech acquisition in an automobile environment, sound capture in reverberant enclosures, large room recording-conferencing, acoustic surveillance and hearing aid devices. Some of these are as follows:

Multiple microphones are widely used on humanoid robots for enhancing the sound and localizing the sound source (Nadakai *et al.*, 2003, 2004)

Acoustic beam forming is a fundamental technique of signal processing on microphone arrays. The aim of a beam former is to discriminate the desired signals from the undesired interferers (Morell *et al*., 1995).

Among various sound source directions, deciding which one to focus i.e. to steer the microphone array is another question. A speaker can be distinguished from other interfering sounds by its statistical properties. Nishiura *et al.* argue that these statistical properties can be modeled with Gaussian Mixture Model (GMM) (Nishiura *et al.*, 2002).

Human beings do quite well in localizing a sound source even in the presence of reverberation or interfering sound sources. It is believed that this ability comes from the reflections in pinna, and than detection of these reflections by cochlea. Based on this assumption, Ono and his friends tried to mimic the pinna and the cochlea to achieve a performance like humans do (Ono *et al.*, 2001).

Direction of arrival (DOA) can be obtained as a by product of independent component analysis. Ikram and Morgan, suggest finding the angles with respect to the microphone array by finding the null steering directions (Ikram and Morgan, 2002). But Sawada and his friends argued that it becomes harder to find null steering directions when the number of sources increases. Instead they proposed another approach which is both able to find DOAs of three or more sources, and computationally cheaper than the former method (Sawada *et al.*, 2002).

Mahajan and Walworth present a formulation for localization of sound source in three dimensions (Mahajan and Walworth, 2001). They utilize the differences between time of flight (TOF) values which is actually the value called "time delay of arrival (TDOA) values" in the literature. They form a system of equations variables of which are the coordinates of the sound source, distance of nearest microphone to the source, and optionally the speed of sound.

Nishiura and his friends suggest a new method to overcome one of the difficulties of the direction of arrival problem which is the fake peaks of the cross correlations due to the presence of multiple correlated sound sources (Nishiura *et al.*, 2000).

1.1. The Scope of This Work

There are detailed studies on estimation of the DOA and sound source localization, but they either lack handling multiple sound sources or work only on two dimensional spaces. Some others are computationally so expansive which makes them impractical to use. This paper presents a new way to find the localizations of multiple sound sources in three dimensional environments. Finding the location or localization of a sound source in three dimensions by using the DOAs is equivalent to find intersection point of three cones. Although this seems to necessitate expensive searches in three dimensional spaces, some partially closed forms can be found and the search can be reduced to one dimension. Localizing sound sources in three dimensions is important, since it can be used to improve the sound quality. It is also essential, because focusing on a certain sound source requires beam forming, and beam forming requires exact location information. These issues are

faced in teleconferencing, humanoid robot designs, meeting room designs without desk microphones etc.

This thesis is organized as follows: Section 2 “Background Information” gives a detailed explanation of some selected previous work on DOA and sound source location estimation. Section 3: the proposed method is explained in detail. Section 4 covers the details of the simulation method that is used to conduct the experiments. Section 5 covers description of the real experiment environments. Finally Section 6 contains the conclusion and future research directions.

2. BACKGROUND INFORMATION

2.1. Direction of Arrival

The direction of arrival estimation is the process of finding the angle between the line passing through a pair microphones and the line connecting the sound source to the one of the microphones. The sound source is assumed to be sufficiently far away from the microphone pair with respect to the distance between the microphones. Thus, the angles of the lines connecting sound source to the microphones are sufficiently close to each other; allowing us to ignore the difference. A hyperboloid can be very well approximated by a cone with an appropriate angle for the points far away from the foci. A simple analysis of this approximation is made, and the error and allowed regions plots are given in the figures.

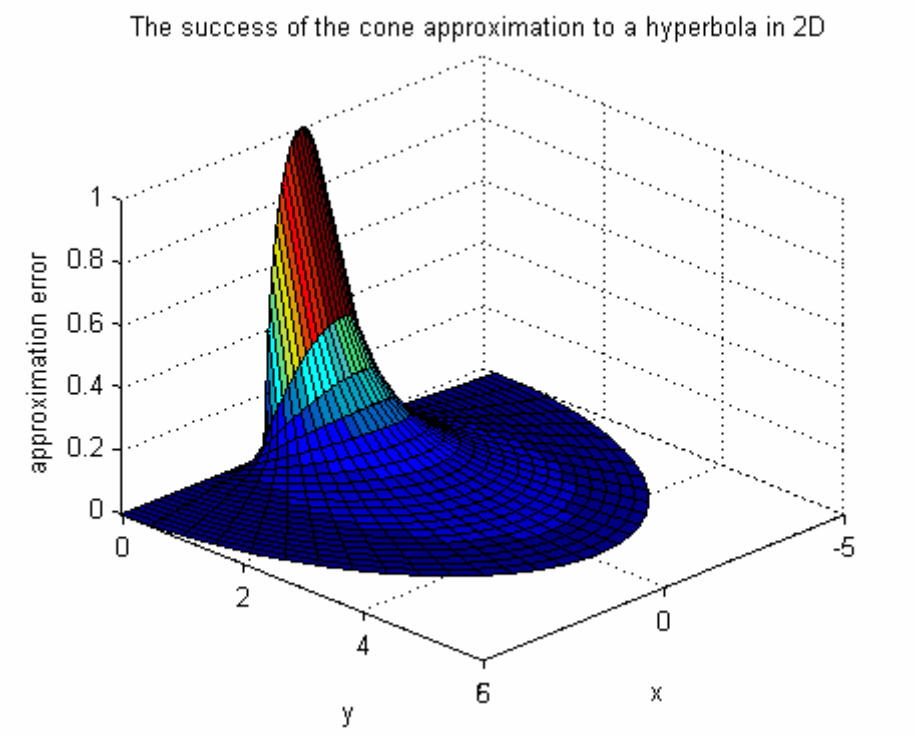

Figure 2.1. The error plot of the cone approximation to a hyperbola

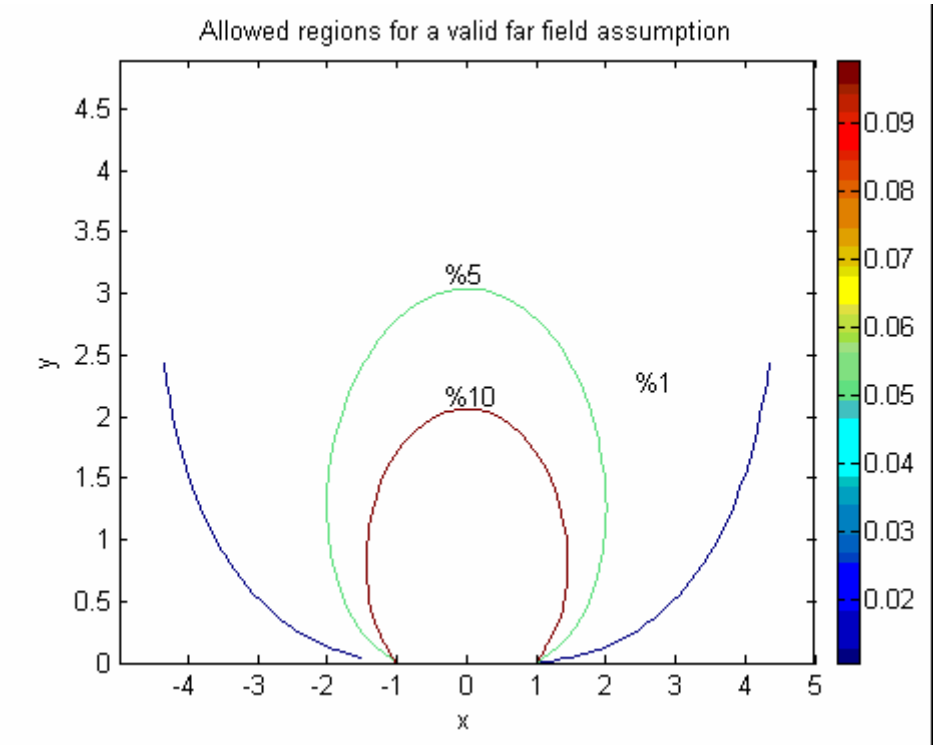

Figure 2.2. The sound source locations for a valid approximation

There are different ways of estimating the direction of arrival. The most common one is estimating the time delay of arrivals to different microphones and then calculating the angle. Some other methods utilize Multiple Signal Classification (MUSIC) or Independent Component Analysis (ICA). These methods are explained in the following sections.

2.1.1. Time Delay of Arrival Estimation (TDOA)

A powerful way of estimating the DOA is using the "Time Delay of Arrivals". Since speed of sound is constant, this time delay is due to the differences of distances from microphones to the sound source. The TDOA estimate can be easily converted to the DOA estimate by a simple calculation.

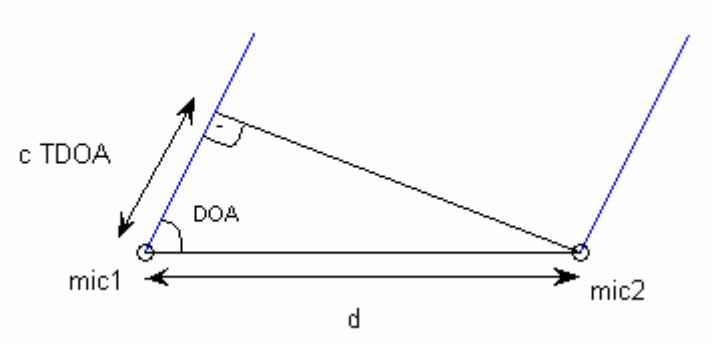

Figure 2.3. Geometric relation between TDOA and DOA estimates

$$DOA = \cos^{-1}\left(\frac{c \cdot TDOA}{d}\right) \tag{2.1}$$

where d is the distance between the microphones, c is the propagation speed of sound in air, and $\cos^{-1}()$ denotes the inverse of cosine function.

So finding a DOA estimate for a pair of microphones means that the sound source is on one side of a hyperboloid in three dimensional space.

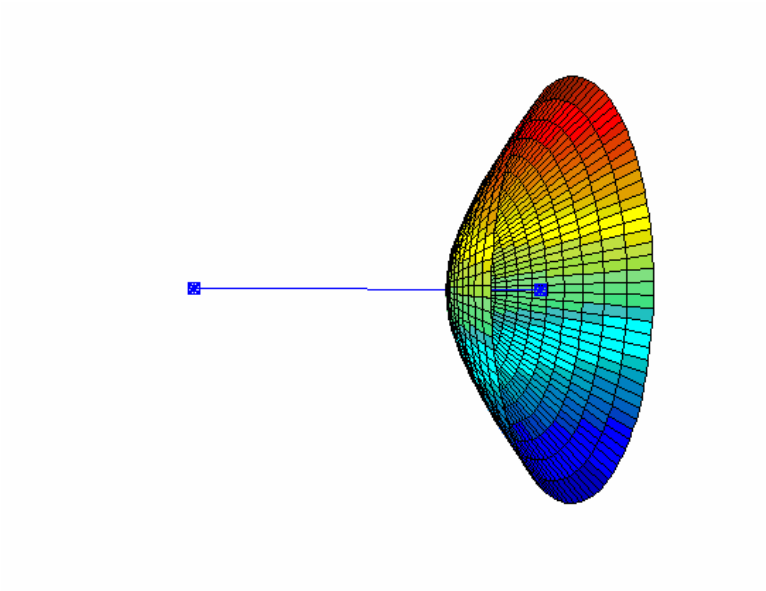

Figure 2.4. The hyperboloid formed by a DOA estimate

One way of estimating TDOA is calculating the cross correlation values between the signals (the sound record of the microphones) for different delays, and then finding the one with the largest correlation value.

The mathematical representation of cross correlation between two signals *x* and *y* for a specific time delay m is given below.

$$R_{xy}(m) = E\{x_{n+m} y_n\} = E\{x_n y_{n-m}\} \tag{2.2}$$

where x_n and y_n are jointly stationary random processes $-\infty<n<\infty$, and $E\{\}$ is the expected value operator.

For finite digital signals an unbiased estimate can be calculated by the following equation:

$$\hat{R}_{xy}(m) = \begin{cases} \dfrac{\sum_{i=0}^{N-m-1} x_{i+m} y_i}{N-m} & m \geq 0 \\ \hat{R}_{xy}(-m) & m < 0 \end{cases} \tag{2.3}$$

Using this formula, 2*N*-1 cross-correlation values can be found between two signals, but this would be computationally expensive. Instead, using the properties of discrete Fourier transform, calculation can be accelerated.

$$\hat{R}_{xy} = DFT^{-1}\left(\frac{DFT(x_n) DFT(y_n)^*}{\left| DFT(x_n) \right\| DFT(y_n) \right|} \right) \tag{2.4}$$

where *DFT ()* and DFT^{-1} *()* denote discrete Fourier transform and inverse discrete Fourier transform, and *()** denotes complex conjugate. This estimate is also called Cross-power Spectrum Phase (CSP) coefficients. The maximum of these values gives the desired TDOA estimate in terms of samples.

$$\tau_{ij} = \arg\max_k (\hat{R}_{ij}(k)) \tag{2.5}$$

Another way of finding the TDOA is finding the slope of the line of the phase of R. This has advantage over the correlation form that inter sample resolution exists. On the other hand one has to solve the phase unwrapping problem first, since the phase is always between $-\pi$ and $+\pi$.

2.1.1.1. Synchronous addition of CSP Coefficients. In the case that there are multiple sound sources, or there exists reverberation, the cross-correlation function may have additional undesired peaks along with the expected ones which are the sound source location pointing ones. Nishiura and his friends propose a method to overcome this issue (Nishiura, 2000). They suggest adding CSP coefficients of multiple microphone pairs, which share a common mid-point, in a meaningful manner. This way the desired cross correlation values are amplified while undesired ones remain same. Thus, the probability of detecting the correct cross correlation value increases. Their results show that increasing number of simultaneous addition of CSP coefficients improves the localization quality.

2.1.2. Multiple Signal Classification (MUSIC)

This method can be classified as a subspace based method since it uses the subspace properties like orthogonality.

Primary aim of the MUSIC algorithm is to detect frequencies in signals, but it also can be used to find the DOA estimates to different sensors.

First the correlation matrix R of sensor observations is calculated. Then R is decomposed to its eigenvectors, such that

$$R = V\Lambda V \tag{2.6}$$

where V is composed of eigenvectors v_k and Λ has zeros everywhere except the corresponding eigenvalues at its diagonal. Eigenvectors and eigenvalues are sorted in descending order. If a function is defined like

$$U(\theta) = \sum_{k=P+1}^{Q} \left| v_k^T * a(f,\theta) \right|^2 \tag{2.7}$$

the θ values that make this function 0 will be the desired directions of arrival.

Here is the explanation of MUSIC algorithm. All signals are represented in the subspace that is spanned by first P eigenvectors. The remaining $Q - P + 1$ eigenvectors represents the noise subspace. The signal subspace and the noise subspace are orthogonal to each other since they are spanned by different eigenvectors. When θ coincides with one of the actual directions, the noise subspace and the steering vector are orthogonal to each other, so the function $U(\theta)$ approaches zero. Apparently Q must be greater than P since $U(\theta)$ is calculated for $Q - P$ values.

2.1.3. Independent Component Analysis (ICA)

2.1.3.1. ICA Approach One. Solving the frequency domain ICA means, obtaining the separating matrix W(f) of the system

$$Y(f, m) = W(f)\, X(f, m) \tag{2.8}$$

where Y is the vector of separated signals, and W is the PxQ separation matrix. Each row of W, w_r, extracts one source signal coming from a specific direction. So it forms spatial nulls in the direction of thee other sources to suppress them. The directivity patterns are formed by $w_r(f)$ is calculated using the steering vector $a(f, \theta)$ as

$$Br(f, \theta) = wr(f)\, a(f, \theta) \tag{2.9}$$

It is observed that for some θ all of the B_r values approaches to zero except one signal that w_r is extracting. Although this method works well for two sources, it becomes difficult to identify the nulls in the directivity patterns for more sources.

2.1.3.2. ICA Approach Two. If the ICA method successfully obtains the separation matrix W, then the frequency response of the mixing system can be estimated as W-1. Since frequency domain ICA has both permutation and scaling ambiguities, H matrix columns may arbitrary scales and permutations compared to the real mixing system.

Therefore, the approximation to H_{qp} has to be revised. Revision is done like this:

$$H_{qp}(f) = A_{qp} e^{j\varphi_p} e^{j2\pi f \frac{d_q \cos\theta_p}{c}} \tag{2.10}$$

where A_{qp} is a real valued attenuation, and phase modulation $e^{j\varphi}{}_p$ at origin. The scaling ambiguity can be canceled out by calculating the ratio between two elements H_{qp} and $H_{q'p}$ corresponding to the same source p. then the angle is calculated with the formula:

$$\theta_p = \cos^{-1}\left(\frac{angle(H_{qp} / H_{q'p})}{2\pi f \frac{d_q - d_{q'}}{c}} \right) \tag{2.11}$$

If the absolute value of the argument of $\cos^{-1}$ appears to be larger than one, then no angles are obtained. In this case another pair of sensors *q*, *q'* should be tried.

Since there is a permutation ambiguity, the calculated θ_p values may not correspond to s_p but to another signal. However all directions can be calculated.

2.2. Localization of the Sound Source

2.2.1. Sound Source Localization in the Existence of Multiple Sound Sources

In two dimensional cases, one has to find two intersecting lines for finding the location of the sound source. So, there has to be at least two pairs (or arrays) of microphones. Although the number of lines increase as the number of sound sources increase, two pairs of microphones are still enough. However, as the number of sound sources increases, a line matching problem arises. The explanation is below.

There is the same number of DOA estimates and lines as the number of sound sources for each pair of microphones, but we do not know which line passes from which sound source. Assuming that the two lines originated from the same pair of microphones are not coinciding, there are *n* factorial possible *n*-tupples for *n* sound sources.

Nishiura and his friends try to solve this problem by clustering the DOA estimates, relying on the specific microphone array formation of their own (Nishiura *et al.*, 2000). In their paper, the all of the 16 microphones are located on a straight line and have equal spaces between two consecutive microphones. Since two lines are necessary for estimating the sound source location, two arrays should be formed. The first 14 microphones form the first array and last 14 form the second one. The 12 microphones in the middle are shared by both of the arrays. So the mid-points of two arrays are close to each other. The natural consequence of this configuration is that two DOA estimates for a sound source would be also close to each other. Excluding some particular cases like existence of close sound sources, the assumption states that a DOA estimate of an array matches to the nearest unmatched DOA estimate of the other array would be valid.

2.2.2. A Closed Form Location Estimator in Three Dimensional Environment

Brandstein and his friends proposed a method which estimates the sound source location with the help of a closed form formula (Brandstein *et al.*, 1996). They place four microphones to each wall of the room and the method first forms a bearing line using two DOA estimates generated from four microphones, and then finds a point that is likely to be closest one to each bearing line especially to the most reliable one. Details of the procedure are explained below.

The microphone quadruples are located such that each of them would lie at the mid-point of the sides of a rectangle. These rectangles positioned on the surface of the side-walls of the room. The TDOA estimates should be given as an input for the microphone couples that are facing at each other. The DOA estimate corresponding to the TDOA estimate is calculated with the equation (2.1).

The lines which connect the microphone pairs, from which the DOA estimates are calculated, are vertical to each other. So, the axes of the cones, which are formed by the DOA estimates, are also vertical to each other. Only the points that lie inside of the room are eligible as sound source locations. The intersection of these cones forms a single beam which is called the bearing line.

Given that all of the DOA estimates are exactly correct, all bearing lines should intersect at a single point namely the sound source location. The imperfections of the DOA calculations may result in disjoint lines. In order to cope with this situation, the method forms every possible line couples from the set of all bearing lines. The closest points on each line to its couple are found -called the candidate points- and weighted according to probability that the specific bearing line passes through the actual sound source. This probability is calculated based on the variance accompanied with the TDOA estimates, which are supposed to be supplied as an input to the method. The ultimate sound source location estimate is then calculated by the weighted average of these candidate points.

2.3. Simulation

2.3.1. Mirror Method

In this method every possible path from the source to the microphones is aimed to be calculated. The walls act like mirrors and reflect sound beams. So in order to find a path that includes a reflection, one first has to find the location of the image of the sound source or the microphone. A transfer function is then found using transmitting properties of the source, the propagation paths, reflection behavior of the walls, absorption condition of the air, and finally sensory specifications of the microphone. Although this algorithm is more realistic and straight forward, because the number of the paths of propagation in a closed room increases exponentially, as the number of reflections increase, it quickly becomes unfeasible to have a sufficiently realistic simulation.

2.3.2. Beam Method

A computationally more efficient solution than the mirror method to the acoustic room simulation is beam method. In this method, only a several number of sound beams are assumed to be spreaded from the source. These beams are propagated in different directions and are assumed to reach the microphone eventually. The implementations of this algorithm have to be aware of the situations in which this assumption will not hold at all or will hold in an undesirably long time. First of all the sensor should occupy some space like a sphere, instead of a point which has no volume at all. Otherwise only very precise beams would reach, which is very unlikely if the directions are selected randomly.

Second cyclic paths should be avoided i.e. if a beam reaches a point that it already have passed from and then reflects to the same direction or a very close one, than this means that it will never reach to the microphone or will reach after a very long time.

3. THE PROPOSED APPROACH

The approach presented here offers a way to find a three dimensional solution to the cases that multiple sound sources are present. Such an approach is selected not only because it is not investigated before, but also it is an important problem for the reasons that are explained in the introduction section.

Nishiura and his friends' approach forms a bases for this new method (Nishiura *et al*., 2000). They suggest synchronous addition of CSP coefficients for improving the localization success. In this thesis, that strategy is tried to be expanded to a general three dimensional space solution.

Four complexities aroused during this expansion job. All of them are common problems with 2-D case, but they were not mentioned at all or handled in a different way in Nishiura's paper. These are as follows: 1.The inability to add different DOA estimates directly due to a precision issue. 2. The microphone array positioning problem. 3. DOA matching problem in the case of existence of multiple sound sources in an environment. 4. The calculation of the location using matched DOA estimates in three dimensional which is the core problem. In this study each of these complexities is explained in detail and the solutions to these complexities are given in the following sub sections.

3.1. Precision Correction: Resampling for Synchronous Addition of CSP Coefficients

Since the new method relies on cross correlation, the calculated DOA estimates have a specific precision determined by the sound acquisition system design. The effective design parameters about precision include, sampling frequency, and inter-microphone distance. The effectiveness of both of these design parameters can be assigned to the time delay capturing ability of the system. The sound waves propagates with a constant and high speed in the air, so the actual TDOA estimate for a microphone pair have to be in a small time interval. The time delay capturing ability of the system depends on how much samples the system can get in this small interval. The more sampling frequency we have,

the more precise solutions come up, and the more inter-microphone distance there is, the more precise solutions we have. Below a demonstration of the precision is given.

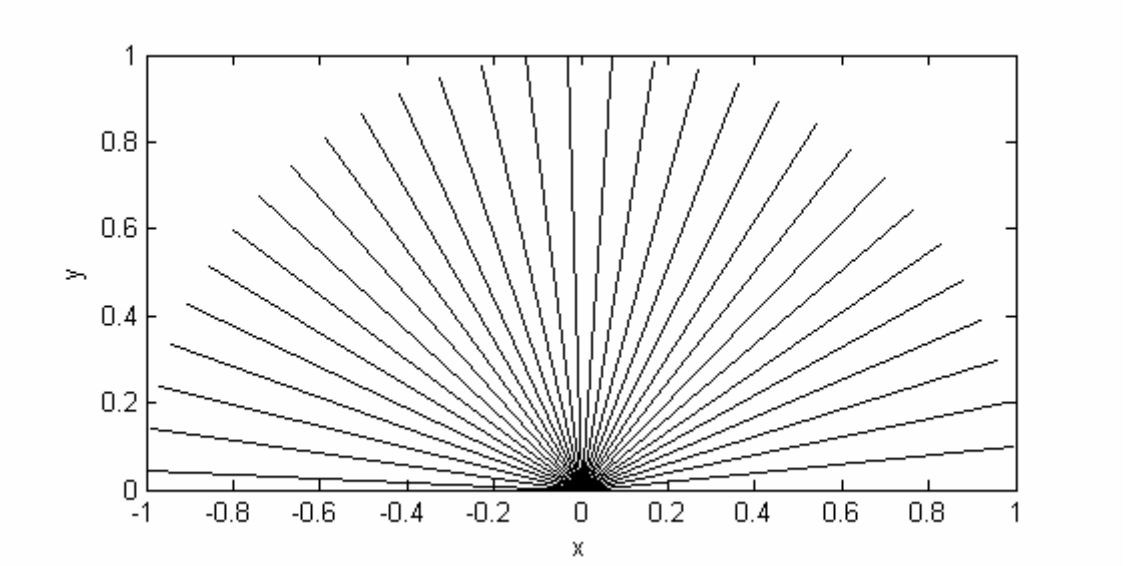

Figure 3.1. A demonstration of the DOA precision

In synchronous addition of CSP coefficients method, the coefficients extracted from different microphone pairs are supposed to be added. But since these microphone pairs have to be aligned and have to share a common mid-point, the inter-microphone distances have to be different than each other. Since the sampling frequencies are the same, the DOA precisions for those pairs are different than each other. The one situated inside the other has a lower precision since the inter-microphone distance is smaller than the outer pair. In order to level the precisions, the product of the sampling frequency and the inter-microphone distance should be leveled too. There are two ways to achieve this: up sampling the received signals of the inner pair, or down sampling the outer pair. Although these products can meet in a mid-point, there are no particular reasons for doing that. So doubling up the computation (both up sample and down sample) makes that option obsolete.

Here the designers have to make a choice between high precision, and fast computation. If she/he decides on up sampling, the computation time of the consequent operations will increase; on the other hand, if she/he decides on down sampling, the precision of the localization calculations decrease. In this thesis, former way of leveling is used, predicating on the fact that the order of the computational complexity of the

remaining operations is not so worrisome; however, sacrificing the precision has a direct inhibitory effect on localization performance.

3.2. Microphone Array Positioning Design

A design of microphone array positioning in a room, can be evaluated using some criteria. These criteria may include robustness against DOA calculation errors, coverage area, and number of microphones used.

The error robustness is a function of the distance between the mid-points of the microphone arrays. If this distance is small, then DOA estimates that correspond to a sound source location for two arrays will be close to each other. In this case, small shifts in DOA estimates may lead to big differences in intersection points of the lines. This means that, a design which locates mid-points of the arrays close to each other would not be so robust against small DOA errors, while the visa-versa case is more robust.

Due to the discontinuous behavior of DOA, in most of the cases, the system may not cover the full range of $[0, \pi]$ interval. Instead they may cover only $[0+x, \pi -x]$ interval where x depends on the inter-microphone distance and sampling frequency. So, the designer should consider this information while positioning the microphone array. The intersection of the region where the people are more likely to talk and the coverage region of the microphone array should be maximized. This can be easily achieved by choosing a line that is parallel to the principle component of the active speaking region and overlapping the mid-point of the array with the mid-point of the region.

Number of microphones that are used in the design is an important factor, because it is hard to find a big number of identical appropriate microphones for a real application. Second, it takes more time to make simulation, as the number of microphones increase. The only way of keeping the number of microphones at a lower level while preserving the number of additions of CSP coefficients is to share a number of microphones between different arrays. This criterion conflicts with the first one, since the more number of shared microphones used, the closer the mid-points of the microphone arrays become. In this

thesis choice is made in the favor of the first criterion, again for preferring localization performance to computational performance.

3.3. Clustering DOA estimates

When there are a number of sound sources in the environment, each microphone array come up with the same number of DOA estimates without any information on the DOA-source matching. Since two or three DOA estimates are necessary to find the location of a sound source in two dimensions or three dimensions, the matching has to be known.

Nishiura and his friends, find an easy solution to this problem, owing to their way of constructing the microphone arrays. The details of their solution are explained in section 2.2.1. In this thesis; however, this problem is not so easy to solve, especially in three dimensional cases, because another way of forming the microphone arrays is used. The steps of the algorithm are given below.

"*nsc*" denotes the number of sound sources.

"*nD*" denotes the number of dimensions.

"C(*x*,*y*)" denotes the combination operator.

1. Beamform each array to each nsc directions.

2. Find TDOA estimate for each microphone array pair and DOA estimates.

3. Find all possible sound source locations (points). It is found by finding the intersection point of two lines in two dimensions or intersection point of three cones in three dimensions. There are nsc^2 possible points in two dimensions, while nsc^3 in three dimensional.

4. Find the distances of all points to the all mid-points of microphone arrays.

5. Find another TDOA estimate by differentiating the distances and then dividing to the speed of sound for each point.

6. Find a measure of "inconsistency" for each point and each DOA estimate by comparing the TDOA estimates that are found in step 2 and 5.

7. Find all possible *nsc*-tuples. It should be noted that if a DOA estimate is used to calculate an intersection; then that DOA estimate may not be used to calculate any other point in a *nsc*-tuple. While there are *nsc* factorial number of *nsc*-tuples in two dimensions, this number squares itself when the environment is three dimensional.

Table 3.1. All possible matchings for two source case

Match No	Source No	DOA x	DOA y	DOA z
1	S1	1	A	a
	S2	2	B	b
2	S1	1	A	b
	S2	2	B	a
3	S1	1	B	a
	S2	2	A	b
4	S1	1	B	b
	S2	2	A	a

Table 3.2. The first six matching of all possible matchings on three source case

Match No	Source No	DOA x	DOA y	DOA z
1	S1	1	A	a
	S2	2	B	b
	S3	3	C	c
2	S1	1	A	a
	S2	2	B	c
	S3	3	C	b
3	S1	1	A	b
	S2	2	B	a
	S3	3	C	c
4	S1	1	A	b
	S2	2	B	c
	S3	3	C	a
5	S1	1	A	c
	S2	2	B	a
	S3	3	C	b
6	S1	1	A	c
	S2	2	B	b
	S3	3	C	a

8. Eliminate *nsc*-tuples which contains a point that is outside of the room.

9. Find total inconsistency of all *nsc*-tuples.

10. Choose the *nsc*-tupple with the minimum inconsistency measure.

3.4. Finding Sound Source Location in Three Dimensions Using DOA Estimates

A DOA estimate estimated from a pair of microphone forms a cone. The intersection of two cones form a curve and the intersection of three cones form a single point. So, three different pairs of microphone are required. If two microphones are sufficient to find the actual DOA estimates, then four microphones would be enough to find the position of a sound source in three dimensional environments if all the microphones are not on the same plane.

In order to find the intersection point of three cones, the cones are expressed in parameterized form. Parametric equations for cones axes of which are the *x, y* and *z* axes respectively, are given below.

$$\begin{aligned} x &= t_1 + tn_1 \\ y &= t_1 \tan(\alpha_1)\cos(s_1) \\ z &= t_1 \tan(\alpha_1)\sin(s_1) \end{aligned} \tag{3.1}$$

$$\begin{aligned} x &= -t_2 \tan(\alpha_2)\cos(s_2) \\ y &= t_2 + tn_2 \\ z &= t_2 \tan(\alpha_2)\sin(s_2) \end{aligned} \tag{3.2}$$

$$\begin{aligned} x &= -t_3 \tan(\alpha_3)\sin(s_3) \\ y &= t_3 \tan(\alpha_3)\cos(s_3) \\ z &= t_3 + tn_3 \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} t_1, t_2, t_3 &\in (-\infty, +\infty) \\ s_1, s_2, s_3 &\in [0, 2\pi] \end{aligned} \tag{3.4}$$

are parameters of the equations, α_i are the vertex angles of the cones, and finally tn_i are the distances of vertices to the origin.

At first step the parametric expressions of the *x, y,* and *z* coordinates for the two cones are equated to each other. Since each cone has two parameters, a total of four parameters exist at this first step. With the help of these three equations, three of them can be represented as a function of the fourth one. As a result all of the parametric equations for the intersection curve of the first two cones can be found as a function of a single variable. These equations form a curve in three dimensional. Special attention should be paid on which side of the cone is selected, since the equations form a full two sided cone, but we are interested in only one of them. The next and last step is to find the point that is both on this curve and the third cone. To check whether a point is on a cone or not,

comparing the angle between the line connecting the point and the vertex of the cone with the half of the vertex angle of the cone. So this last job is finding the root of the function showing the difference of the two angles. The figure below demonstrates the intersection of three cones.

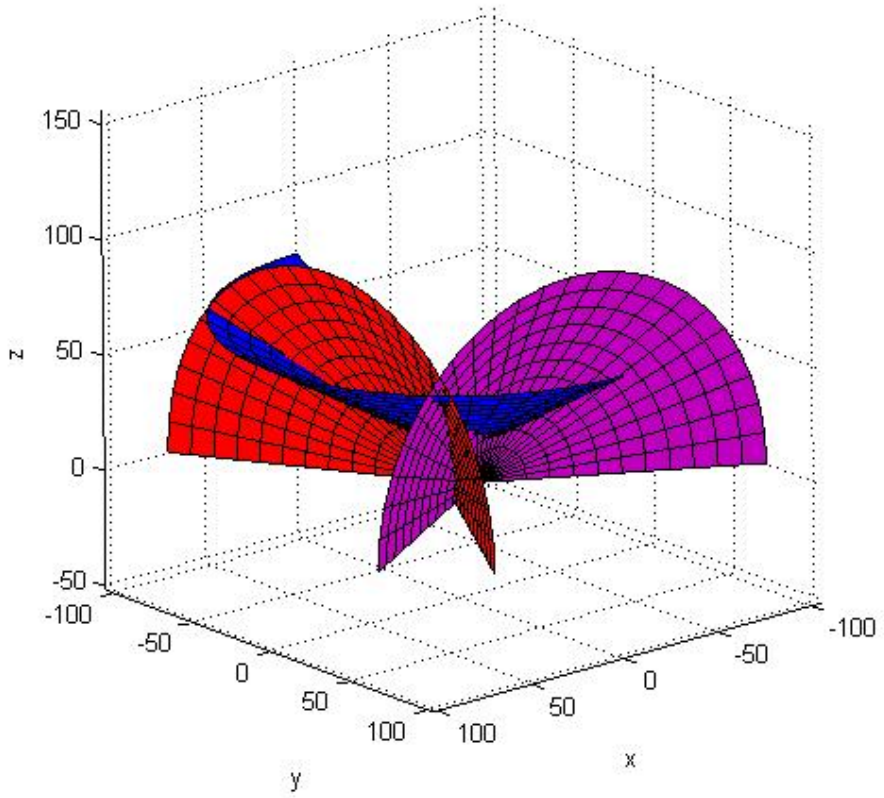

Figure 3.2. A plot showing the intersection of three cones

4. RESULTS OF SIMULATIONS

Mirror method was used for simulation. A computer program was designed and coded the author in Matlab. The codes of the program are given in the compact disc attached to this thesis. The program has the following assumptions and functions (abilities):

Assumptions:

1. Sound propagates in the air with 340 meters per second.

2. As the sound travels through the air, it attenuates in a frequency independent manner.

3. When the sound reflects from a wall, it attenuates in a frequency independent manner.

4. All of the sources and microphones are omni directional.

Abilities and constraints:

1. Any convex space with planar boundary surfaces can be modeled.

2. Any number of microphones and sound sources can be placed anywhere in the space.

3. Any number of successive reflections can be represented. However, there is a practical upper bound on the number of the successive reflections can be studied because the computational cost grows exponentially while the number of reflections increases.

4.1. Validation of the Results of the Simulation Experiments

A validation of the results can be made by testing the system. Tests are made by giving some specific inputs and observing the outputs of the system, and then comparing these outputs with the obvious results of those specific inputs.

The mean localization error plots for different inputs are given. Each data point in these plots are the overall averages of experiments conducted for 120 different location couples, 40 different frames of a signal, two sound source locations, a total of 9600 values

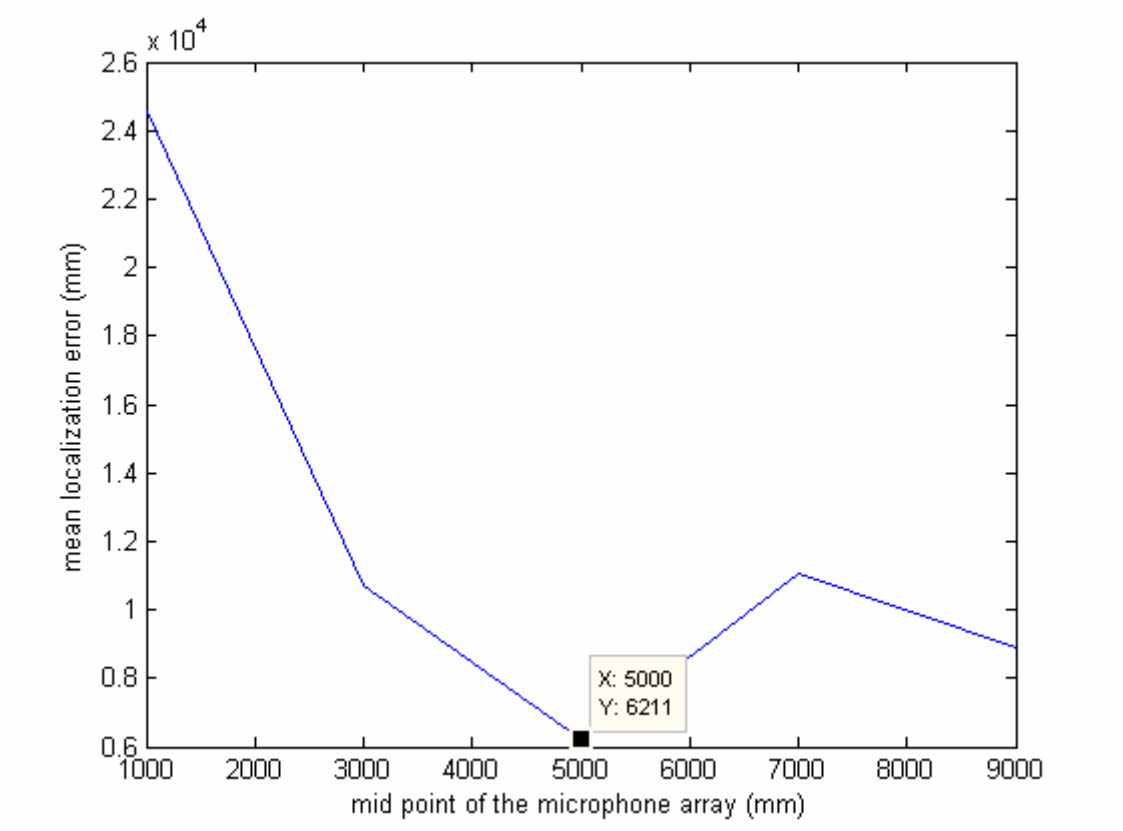

Figure 4.1. The effect of microphone array location

This plot verifies that the coinciding the mid points of the microphone arrays and the apses of the center of the active speaking region.

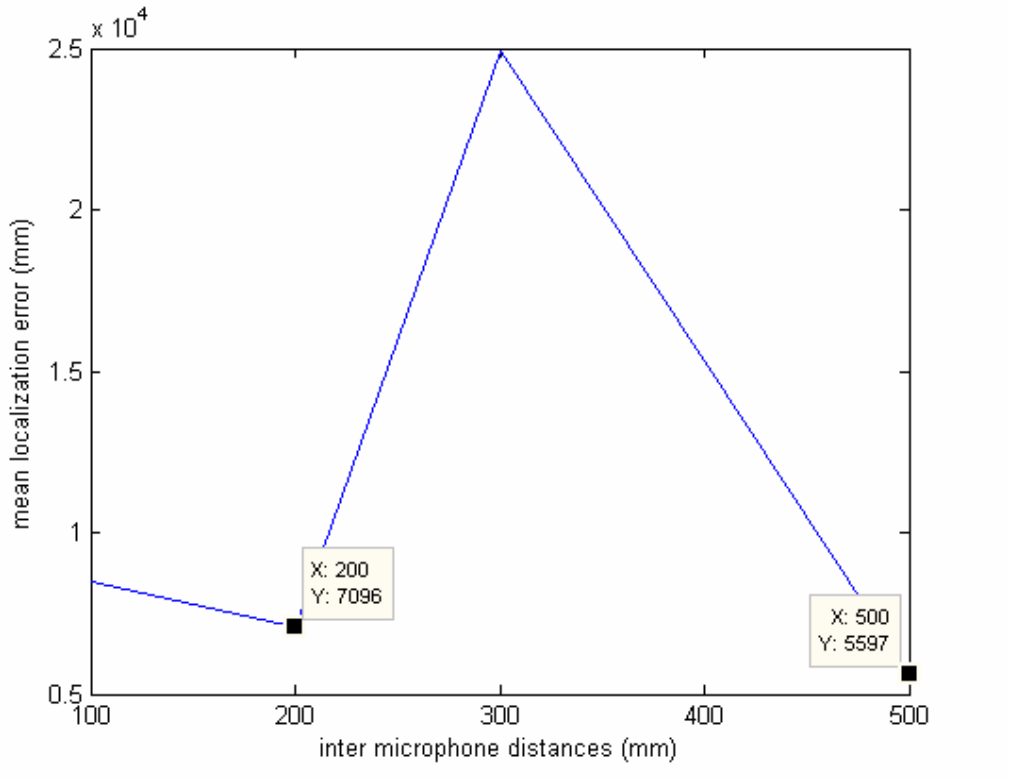

Figure 4.2. The effect of inter microphone distances on error

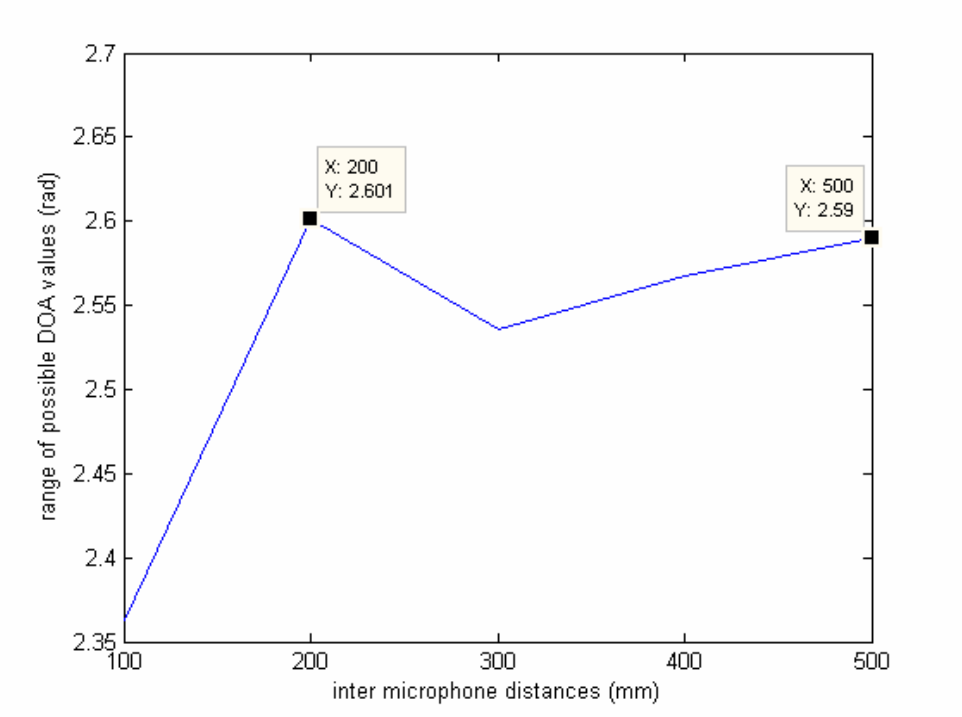

Figure 4.3. The effect of inter microphone distances on DOA range

These two graphics show, how together inter-microphone distances and the range of the DOA values in other words coverage area affects the performance of the localization.

4.2. Experimental Study

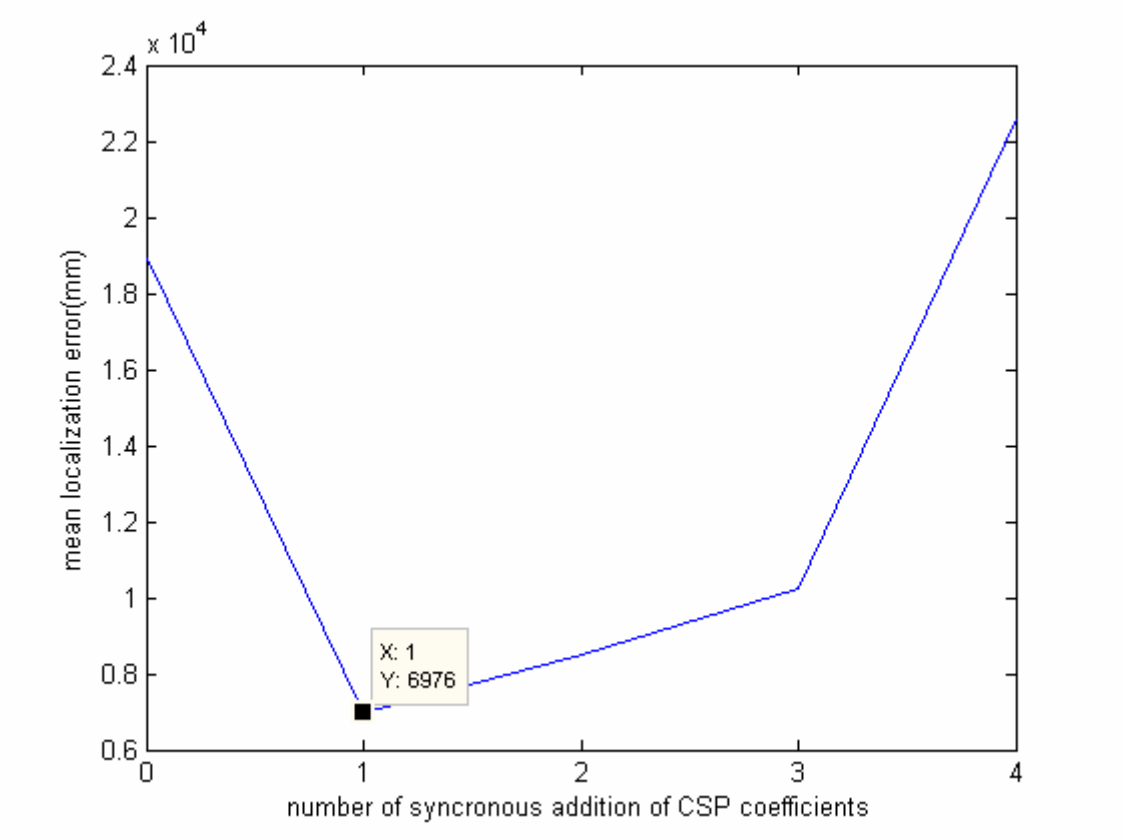

Figure 4.4. Effect of synchronous addition

This plot shows the effect of increasing the number of synchronous addition of CSP coefficients. We see that increasing the number of additions does not help after some point. This situation can be assigned to the weakness of the far field assumption. For more addition, more microphone pairs are necessitated so the inter microphone distance grows up and far field assumption becomes obsolete.

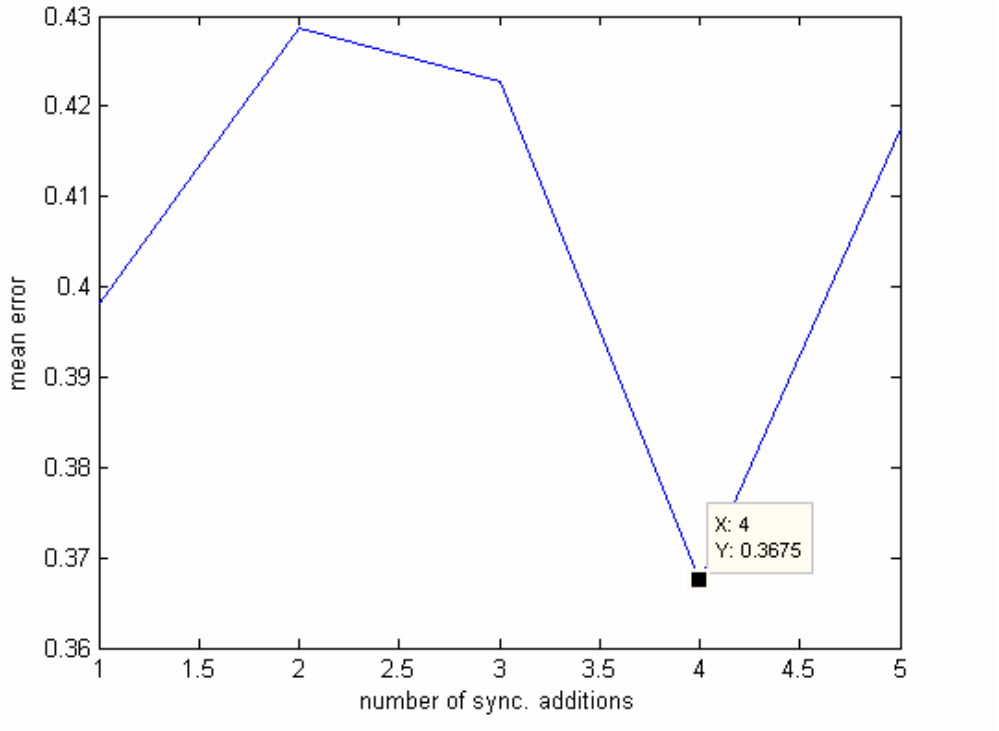

Figure 4.5. Effect of synchronous addition on three dimensional environment.

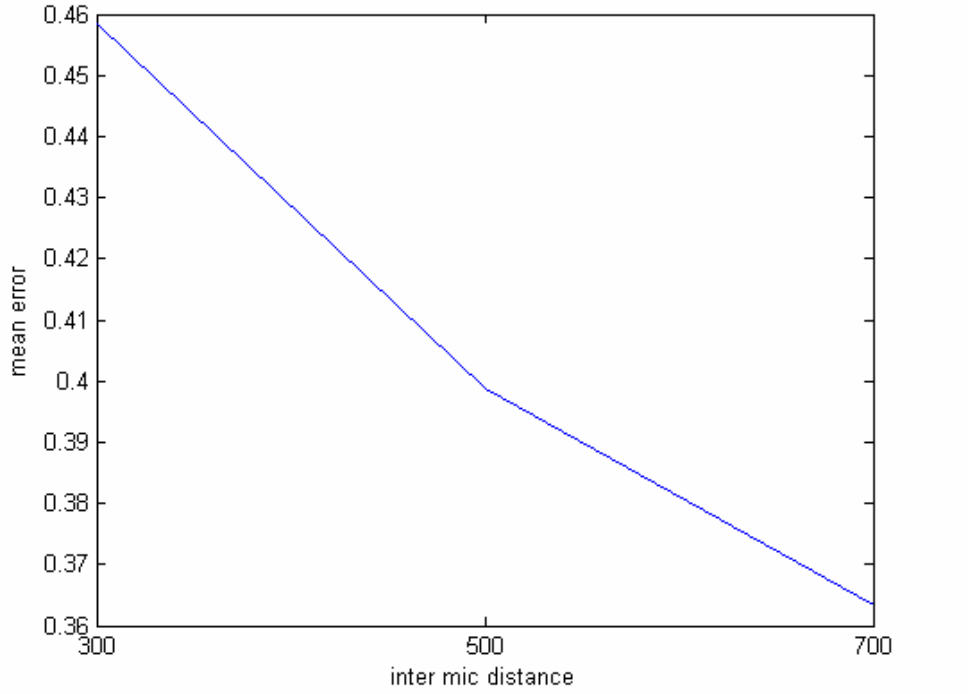

Figure 4.6. Effect of inter microphone distances on three dimensional environment

Following tables shows the mean localization errors, mean DOA errors and percentage of the outliers for different sound source locations. The sound source locations are numbered in the same manner with the real experiment environment two for convenience.

Table 4.1. Simulation results for different sound source locations.

Location of the second source	**Mean Localization error (mm)**	**DOA Errors (radians)**			**Ratio of Outliers**
		X	**Y**	**Z**	
5	198.32	0.04	0.07	0.04	0
6	198.32	0.04	0.07	0.04	0
7	204.08	0.05	0.04	0.02	0
8	205.89	0.05	0.04	0.02	0
9	365.20	0.03	0.06	0.05	0
10	368.75	0.03	0.06	0.05	0
11	1300.67	0.14	0.02	0.06	0
12	1300.67	0.14	0.02	0.06	0
13	165.43	0.06	0.03	0.02	0
14	166.19	0.06	0.03	0.02	0
15	243.90	0.01	0.06	0.02	0
16	243.44	0.01	0.06	0.02	0
17	330.59	0.04	0.04	0.03	0
18	330.59	0.04	0.04	0.03	0
19	651.21	0.09	0.01	0.05	0
20	651.21	0.09	0.01	0.05	0
Mean	**432.78**	**0.06**	**0.04**	**0.04**	**0**

The reverberation conditions of these experiments were 0.1 reflection coefficient for walls and at most two successive reflections occur. The following table contains the results for a more reverberant environment i.e. there could be three successive reverberations. The mean error values for all of the sound source locations indicates as the reverberations increase, localization errors also increse.

Table 4.2. Simulation results

Location of the second source	Mean Localization error (mm)	DOA Errors (radians)			Ratio of Outliers
		X	Y	Z	
5	529.39	0.06	0.08	0.05	0
6	522.20	0.06	0.08	0.05	0
7	240.94	0.02	0.04	0.03	0
8	240.94	0.02	0.04	0.03	0
9	360.81	0.03	0.04	0.02	0
10	360.81	0.03	0.04	0.02	0
11	661.51	0.06	0.02	0.04	0
12	661.51	0.06	0.02	0.04	0
13	862.22	0.02	0.10	0.06	0
14	861.67	0.02	0.10	0.06	0
15	1119.15	0.04	0.10	0.05	0
16	1119.15	0.04	0.10	0.05	0
17	864.94	0.04	0.05	0.04	0
18	864.94	0.04	0.05	0.04	0
19	1760.52	0.05	0.08	0.03	0
20	1732.15	0.05	0.08	0.03	0
Mean	**797.68**	**0.04**	**0.06**	**0.04**	**0**

5. RESULTS OF REAL EXPERIMENTS

5.1. Environment One

First Experiments were conducted in a record studio environment. Walls were insulated to prevent echoes to a certain extent. Seven microphones were used. Microphones were positioned as if they construct a three axis of the coordinate system as shown below.

Figure 5.1. Experiment environment one

Each microphone was 20cm away from the next one. The origin of the coordinate system was located at 160cm high. Because the microphones are not designed to sense distant speech (dynamic microphones) and humans can not speak loud enough, speakers are used as the sound sources. Two speakers were used. Prerecorded sound files were played during the experiment. The sound files were created by recording female and male

speakers reading text. Three sound files are recorded: file one, a 15 minute female speaker, file two, a five minute female speaker, and file three, another five minute male speaker.

These files were recorded on the computer at 44.1 kHz. Sound file one was played from each speaker. Seven minutes of the fifteen minute sound file was played from the left speaker and the rest was played from the right one. Additional data was collected while sound file two was playing from the right and sound file three was playing from the left speaker simultaneously. All the sound files were recorded from all the microphones at 96 kHz and down sampled to 48 kHz afterwards.

5.2. Environment Two

Another series of experiment are conducted to be able to record real human sound instead of reproduction of it from the speakers. This time condenser and omni-directional microphones were used as sensors. Although not all of the microphones were of the same kind, this weakness became less important by matching same kind of microphones as pairs. A much larger space exists to make enough experiments on. 12 microphones are used, four for each axis. Each microphone was 20 cm away from the next one; the first microphones among the four are zero, 110 and 140 cm away from the origin. The microphones, which were positioned on x-y plane, were at 160 cm high. Twenty two points were marked on the floor as sample speaker locations. Each of them was one meter away from the others in any direction. All recordings were done at 88 kHz sampling rate and 24 bits per sample. Experiments consist of one speaker, two and three stable speakers spontaneously, a single sound source that was generated by clapping two metal bars to each other, a single sound source and a source of noise, and finally one speaker on the move.

Single source experiments were made in every marked position by each of three speakers. Two of the speakers were male, and one was female. Each speaker spoke a sentence notifying the position (as numbered on the floor) of him/her at that moment, followed by counting from one to five in Turkish.

Double sound source experiments were conducted as follows. One speaker remained fixed at a point and the other speaker moved around all other points while reading a

Turkish test spontaneously. Four sets of experiments were made. In each set the fixed speaker stands at one of the predetermined points.

Triple sound source situations also examined, but only three different configuration sets were studied. In these sets two speakers remained fixed at their locations and the third one moved around the other points. Metal bar clapping is also studied, because they give an explosive sound and should be easy to determine its location. Finally moving source experiment was run. A speaker continuously spoke words about his location while he walked through the marked points.

Figure 5.2. The microphone array setup for experiment two.

Figure 5.3. A wide view of experiment environment two

The following table summarizes the results for the experiments made for single sound source. The detailed tables are given at the appendix.

Table 5.1. Summary of results of experiments with single sound source

Speaker Location	Mean Localization error (mm)	Mean DOA Errors (radians)			Ratio of outliers (per cent)
		X	Y	Z	
5	374.13	0.24	0.14	0.12	7
6	523.91	0.22	0.14	0.09	6
7	1878.96	0.28	0.22	0.12	10
8	2730.66	0.59	0.30	0.31	29
9	694.65	0.25	0.20	0.21	13
10	927.22	0.16	0.19	0.19	10
11	1633.43	0.18	0.25	0.35	30
12	2551.86	0.36	0.29	0.40	28
13	1505.77	0.19	0.33	0.21	6
14	1213.54	0.19	0.32	0.23	14
15	2013.68	0.18	0.29	0.34	21
16	2825.18	0.33	0.31	0.35	25
17	1784.32	0.17	0.39	0.19	13
18	2074.39	0.20	0.35	0.22	17
19	1718.12	0.23	0.32	0.35	31
20	3001.70	0.28	0.47	0.32	26
Mean	**1715.72**	**0.25**	**0.28**	**0.25**	**18**

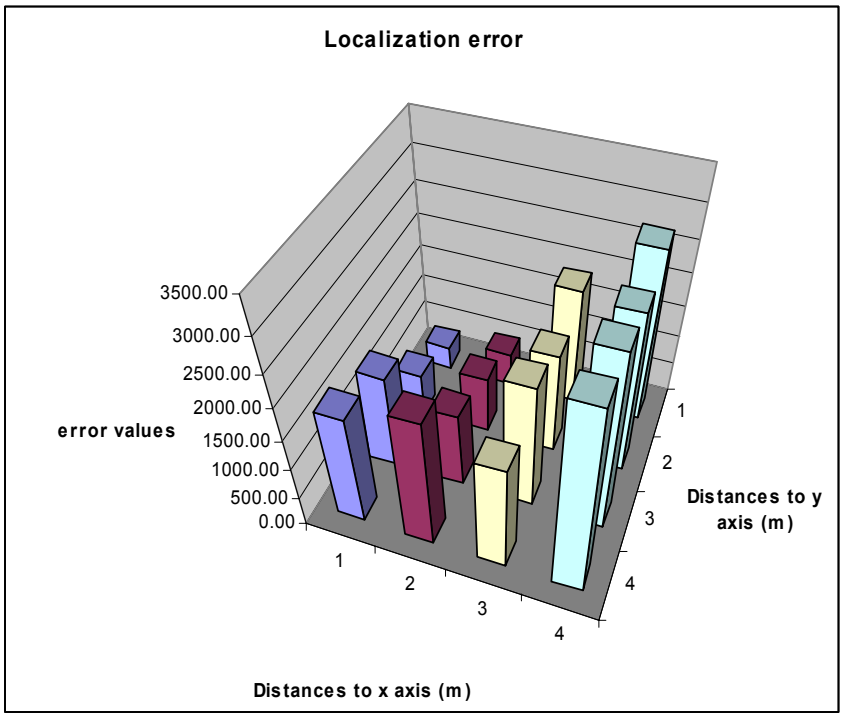

Figure 5.4. A plot of localization error according to sound source's location on the environment

These results show that the error increases as the sound source gets away from the origin. This is related to the diminishing signal power which is an important aspect when detecting the cross correlation values.

The following table shows the mean errors when there are two sound sources, one of which is standing at point five throughout this set of experiments, and the other one speaks at different locations. These locations are shown in the table.

Table 5.2. Summary of results of experiments with two sound sources

Location of the second source	Mean Localization error (mm)	DOA Errors (radians)			Ratio of Outliers (per cent)
		X	Y	Z	
6	1415.00	0.37	0.18	0.21	8
7	1579.28	0.40	0.18	0.30	9
8	1333.93	0.39	0.18	0.24	10
9	1118.23	0.37	0.48	0.16	10
10	1219.39	0.31	0.28	0.19	9
11	1790.64	0.38	0.24	0.24	12
12	1287.78	0.43	0.25	0.22	18
13	1269.14	0.32	0.33	0.36	14
14	1474.46	0.33	0.32	0.24	12
15	1934.51	0.32	0.31	0.28	9
16	1711.31	0.33	0.27	0.23	5
17	1373.94	0.36	0.35	0.26	19
18	2050.96	0.32	0.40	0.18	8
19	1549.95	0.37	0.27	0.24	11
20	2981.25	0.32	0.28	0.25	7
Mean	**1605.98**	**0.35**	**0.29**	**0.24**	**10.73**

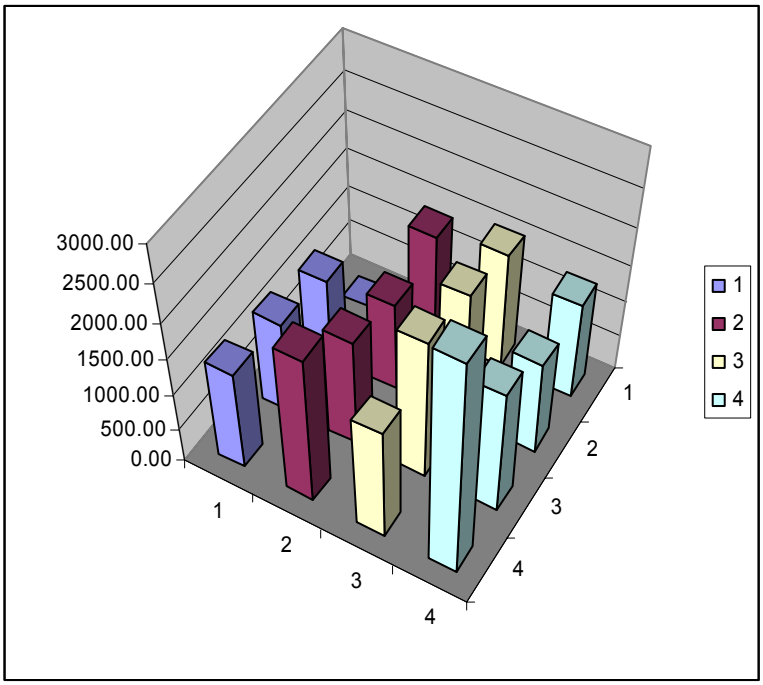

Figure 5.5. A plot of mean localization errors according to sound source's location on the environment

5.3. Comparison of Results with Simulations

A comparison between the real environment results and the simulation environment result is meaningful, because then it will be possible to track the performance degrading factors by comparing the simulation and the real acoustic environments.

The following table contains the average performance of experiments at real acoustic environment two and experiments on the simulation model of that environment.

Table 5.3. Comparison of simulated and real environments

Environment	Number of Reflections	Mean Localization error (mm)	DOA Errors			Ratio of Outliers
			X	Y	Z	
Simulation env. 1	2	432.78	0.06	0.04	0.04	0.00
Simulation env. 2	3	797.68	0.04	0.06	0.04	0.00
Simulation env. 3	4	991.84	0.06	0.06	0.05	0.00
Real env.	-	1715.72	0.25	0.28	0.25	18

These results show that the number of reflections allowed for the simulation environment has a direct inhibitive effect on the localization performance. More over the results of real acoustic environment are not as accurate as the simulations. This situation should be assigned to the assumptions and simplifications that are done during the simulation modeling. One draw back is directivity patterns of the sound sources are ignored during the simulations. If a speaker has the microphones at his back, then the reflections might have more power than the direct sound. This kind of situation would degrade the performance for sure. The bodies of the sound sources are also ignored. In the case of multiple sound sources, each sound source may prevent others to reach the microphones and distracts the calculations. Lots of objects in the room, frequency dependent reflection structures of the walls, different conditions affecting the propagation of the sound in the air are other differences between the simulation and our real acoustic environment.

6. CONCLUSIONS

A new approach to the multiple sound source localization problem in three dimensional environments is proposed in this thesis.

The first complication due to the existence of multiple sound sources, namely the undesired correlation between two signals, is overcome using the synchronous addition of the CSP coefficients method of Nishiura and his friends (Nishiura *et al.,* 2000). This solution comes with its own problem, which is called precision leveling. The synchronous addition of CSP coefficients necessitates a leveled TDOA estimate precision. The microphone pairs with different inter microphone distances have different precisions. In this thesis, this situation is leveled by resampling of the signals that are recorded from one of the microphone pairs. The details of precision leveling are given at Section 4.1.

The second complication, namely the matching of the DOA values, is solved by a new matching method, based on a measure called inconsistency measure. This measure is an indicator of the validity of a candidate point. It is calculated by comparing two TDOA values that are estimated in two different ways. The first method is the CSP analysis approach but this time the signals are the beamformed signals generated from the microphone arrays that are electronically directed or steered to the candidate sound source. The second is directly calculating the difference of distances from candidate sound source location and the mid points of the microphone arrays. After finding all of the possible solutions, the one with the minimum total inconsistency is selected as the actual solution. A solution includes locations of all of the sound sources. The details of DOA matching are given at section 4.3.

Finally the formation of candidate solutions is solved. They are formed by finding the intersection points of the cones representing the DOA values for three axes. The intersection of two cones is first found in a closed function of a single variable. This constructs a curve in three dimensional space. Then the intersection of this curve and the last cone is found by a search method. Sample points are generated on the curve and are

checked whether they are on the cone or not. The one on the cone is the intersection point of the three cones.

Coming up with solutions to all of the complexities and problems, the algorithm is tested on simulation and real acoustic experimental environments. The capabilities and limitations of the acoustic simulator are given in detail at Section 4. The real acoustic environment is explained in detail at Section 5.2. A comparison of simulation and real environment results are given at section 5.3.

Results show that the proposed method of localization of the multiple sound sources in three dimensional environments works reasonably well. It can be used various applications that needs sound source applications like speech acquisition in acoustical environments without desktop microphones or headsets.

7. FUTURE RESEARCH DIRECTIONS

In this thesis a simple search which checks every single point one by one on the intersection curve of two cones is utilized, but this degrades the computation speed. A cleverer search algorithm might be developed to find the intersection of three cones. Some clever ways of selecting a start point for the search may also be developed.

An extensive analysis of the relation between several design and environment parameters and the localization performance can be studied. The effects of microphone types, acoustic condition of the room, and presence of noise in the environment are some of the worth-to-discover relations.

The effect of correct beam formation on speech recognition systems can be studied. This study might make desktop microphones unnecessary and intelligent systems might become easier to communicate like clever homes.

APPENDIX A: DETAILED RESULTS OF REAL EXPERIMENTS

Table A.1. Detailed results of experiments

		Sound source locations							
		5	6	7	8	9	10	11	12
Frame Sequence	1	2575.50	268.27	2099.09	1765.47	NaN	983.44	NaN	5411.88
	2	159.54	268.27	4539.51	914.98	1998.59	12198.59	513.05	NaN
	3	159.54	274.27	2071.81	855.08	1687.62	1145.57	967.90	949.53
	4	159.54	268.27	478.50	1459.59	NaN	416.88	703.03	1109.30
	5	159.54	268.27	452.56	1794.17	294.71	377.09	509.05	1443.14
	6	252.43	749.27	452.56	62571.60	294.71	377.09	684.86	31538.14
	7	174.69	950.17	538.03	NaN	817.07	377.09	1472.21	1596.70
	8	159.54	646.81	NaN	899.43	845.10	377.09	NaN	1605.73
	9	191.20	242.65	977.53	558.42	2582.43	247.00	NaN	949.53
	10	191.20	261.65	358.71	1288.78	294.71	1620.47	870.73	949.53
	11	191.20	261.65	996.14	855.08	337.41	484.25	503.42	1240.25
	12	191.20	541.85	538.03	914.98	313.93	384.44	513.05	949.53
	13	NaN	287.98	358.71	2161.71	313.93	377.09	8078.28	NaN
	14	208.18	274.27	800.87	NaN	313.93	NaN	NaN	3541.08
	15	191.20	274.27	553.81	1732.90	313.93	949.32	503.42	NaN
	16	191.20	261.65	538.03	1373.12	430.93	5150.95	NaN	3385.48
	17	191.20	268.27	452.56	4587.38	313.93	450.51	1858.05	NaN
	18	208.18	268.27	965.25	5012.26	313.93	377.09	NaN	1109.30
	19	541.37	541.85	1001.59	NaN	318.87	377.09	605.33	NaN
	20	174.69	1310.70	538.03	NaN	318.87	377.09	NaN	2667.25
	21	540.38	4919.65	452.56	695.40	302.48	377.09	617.63	3402.28
	22	174.69	1163.47	682.63	NaN	407.86	646.82	5672.71	NaN
	23	1172.39	541.85	4382.76	2598.89	464.74	483.10	892.56	949.53
	24	191.20	541.85	1477.84	7099.19	NaN	377.09	309.13	1352.12
	25	159.54	305.72	682.63	2857.42	2018.53	377.09	503.42	NaN
	26	191.20	268.27	16167.87	3186.76	2418.30	377.09	NaN	NaN
	27	159.54	NaN	358.71	3629.19	7616.30	377.09	503.42	2664.27
	28	159.54	923.52	358.71	14832.86	NaN	377.09	3328.72	1084.45
	29	893.20	823.51	370.81	2686.63	318.87	377.09	605.33	2255.83
	30	847.43	305.72	358.71	3168.32	363.70	377.09	605.33	NaN
	31	NaN	1013.75	19242.92	3844.01	430.93	377.09	NaN	2862.80
	32	191.20	268.27	NaN	2890.98	258.81	450.51	1290.55	5938.73
	33	180.19	328.15	358.71	NaN	313.93	7072.79	739.42	1605.73
	34	NaN	1225.29	358.71	NaN	430.93	485.62	NaN	949.53
	35	191.20	2815.22	1436.62	NaN	318.87	729.63	NaN	1968.16

Table A.1. (continued) Detailed results of experiments

		Sound source locations							
		13	14	15	16	17	18	19	20
Frame Sequence	1	700.66	3398.94	3372.04	NaN	2560.49	1603.76	NaN	NaN
	2	700.66	1618.41	2224.61	876.62	NaN	918.57	866.16	3389.14
	3	700.66	1790.43	828.35	1199.62	11629.48	918.57	1172.39	4452.19
	4	2887.39	1394.37	826.99	NaN	779.36	918.57	1122.63	3852.98
	5	1108.62	1035.11	877.06	NaN	779.36	1074.86	3390.17	1522.24
	6	700.66	3640.72	1440.63	1141.64	1203.93	918.57	2977.90	1122.53
	7	700.66	1477.44	NaN	1141.64	1203.93	1229.17	3056.77	7241.37
	8	700.66	611.93	4818.31	1440.78	1203.93	1716.56	1172.39	2610.88
	9	700.66	611.93	828.35	11087.16	1276.30	1229.17	1172.39	6564.26
	10	700.66	989.51	1172.43	1141.64	1203.93	2393.66	NaN	1522.24
	11	700.66	611.93	NaN	2317.73	779.36	2183.05	1172.39	577.97
	12	700.66	611.93	NaN	6087.51	1203.93	1229.17	1172.39	2475.13
	13	700.66	NaN	828.35	NaN	779.36	918.57	NaN	1522.24
	14	837.44	NaN	1227.15	15929.69	1770.72	918.57	5014.51	NaN
	15	700.66	NaN	NaN	1588.17	1203.93	1229.17	NaN	2325.54
	16	700.66	611.93	828.35	566.33	779.36	1229.17	3825.83	NaN
	17	736.58	611.93	10765.24	1466.42	2322.34	4865.05	7656.69	1420.99
	18	837.44	611.93	516.39	1141.64	901.11	9104.55	1295.60	3403.48
	19	700.66	611.93	1632.56	1141.64	1425.44	2546.37	NaN	1122.53
	20	700.66	1076.41	828.35	742.15	1822.38	NaN	NaN	1522.24
	21	8521.08	6735.83	NaN	905.65	2370.35	5300.98	NaN	1122.53
	22	700.66	1050.30	NaN	NaN	1203.93	NaN	NaN	NaN
	23	697.42	611.93	3102.44	5934.98	2696.24	8237.62	5052.52	202.09
	24	700.66	611.93	NaN	4610.36	605.73	403.96	1500.75	4410.29
	25	3695.17	611.93	7179.54	2239.46	NaN	510.73	1172.39	2605.90
	26	4265.53	611.93	6678.75	2533.99	1203.93	2655.82	1172.39	NaN
	27	16508.86	733.51	877.06	3403.58	2553.49	1229.17	1172.39	2414.37
	28	1681.90	750.60	994.91	2889.42	1203.93	1229.17	1996.72	4143.54
	29	1413.58	611.93	828.35	3990.08	1203.93	1229.17	1172.39	NaN
	30	700.66	611.93	2609.85	3849.20	NaN	6952.94	2057.36	3861.44
	31	700.66	611.93	NaN	1452.11	1203.93	7435.92	NaN	3164.50
	32	700.66	2514.77	3040.28	1141.64	1203.93	NaN	NaN	1555.13
	33	700.66	3214.37	1397.16	3904.48	779.36	NaN	NaN	1122.53
	34	700.66	611.93	828.35	3852.51	6340.76	NaN	871.63	NaN
	35	5124.87	611.93	2171.75	4047.43	2941.19	2417.14	NaN	NaN

The entries of Table A.1 are localization errors in mm. The NaN values represents an outlier i.e. either the solution is outside of the room boundaries, or no solution is available for that frame in other words the cones do not intersect at all.

REFERENCES

Brandstein, M. S., J. E. Adcock, and H. F. Silverman, 1997, "A closed form location estimator for use with room environment microphone arrays", *IEEE Transactions on Speech and Audio Processing*, v.5, No.1, p. 45 – 50

Gröhn M., 2002, "Localization of a Moving Virtual Sound Source in a Virtual Room, The Effect of a Distracting Auditory Stimulus", *Proceedings of the Intl. Conference on Auditory Display,* Kyoto, Japan, July 2-5

Ikram M. Z., and D. R. Morgan, 2002, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation", Proceedings. (ICASSP '02), *IEEE International Conference on Acoustics, Speech, and Signal Processing*, v.12, No.5, p.530-538

Mahajan, A and M. Walworth, 2001, "3-D Position Sensing Using the Differences in the Time-of-Flights from a Wave Source to Various Receivers", *IEEE Transactions on Robotics and Automation*, v.17, No.1, p. 91-94

Morell A., A. Pascual-Iserte, and A. Perez-Neira, 2005, "Fuzzy Inference based robust beamforming", *Signal Processing*, v.85 Elsevier, p. 2014-2029

Nakadai K., D. Matsuura, H.G. Okuno, and H. Kitano, 2003, "Applying Scattering Theory to Robot Audition System: Robust Sound Source Localization and Extraction", Proceedings of the 2003 IEEE/RSJ, *Intl. Conference on Intelligent Robots and Systems*, Las Vegas, October, p. 1147-1152

Nakadai K., H.G. Okuno., and H. Kitano, 2004, "Robot Recognizes Three Simultaneous Speech by Active Audition", *Proceedings of the 2003 IEEE Intl. Conference on Robotics & Automation,* Taipei, September, p. 398-405

Nishiura T., M. Nakamura, A. Lee, H. Saruwatari, and K. Shikano, 2002, "Talker Tracking Display on Autonomous Mobile Robot with a Moving Microphone Array", *Proceedings of the Intl. Conference on Auditory Display,* Kyoto, Japan, July 2-5

Nishiura T., T. Yamada., S. Nakamura, K. Shikano, 2000, "Localization of Multiple Sound Sources based on a CSP Analysis with a Microphone Array", *ICASSP IEEE Intl. Conf. Acoustic Speech Signal Process Proc,* p. 1053- 1056

Ono, N., Y. Zaitsu, T. Nomiyama, A. Kimachi, S. Ando, 2001 "Biomimicry Sound Source Localization with .Fishbone", *Trans. IEE of Japan*, Vol. 121-E, No. 6, June, pp. 313-319.

Sawada H., R. Muaki, and S. Makino, 2003 "Direction of Arrival Estimation for Multiple Source Signals Using Independent Component Analysis", *IEEE,* p.411-414

Xiaojun L., Z. Xianda, and B. Zheng, 2002 "Estimating DOA using Independent Component Analysis", *ICSP Proceedings,* p.1389-1391

Yamamoto S., K. Nakadai, H. Tsujino, T. Yokoyama and H.G. Okuno, 2004, "Improvement of Robot Audition by Interfacing sound Source Separation and Automatic Speech Recognition with Missing Feature Theory", *Proceedings of the 2004 IEEE Intl. Conference on Robotics & Automation*, New Orleans, April, p.1517-1523