2007

EXPERIMENTAL EVALUATION OF MODIFIED PHASE TRANSFORM FOR SOUND SOURCE DETECTION

Anand Ramamurthy
University of Kentucky, ananram@gmail.com

ABSTRACT OF THESIS

EXPERIMENTAL EVALUATION OF MODIFIED PHASE TRANSFORM FOR SOUND SOURCE DETECTION

The detection of sound sources with microphone arrays can be enhanced through processing individual microphone signals prior to the delay and sum operation. One method in particular, the Phase Transform (PHAT) has demonstrated improvement in sound source location images, especially in reverberant and noisy environments. Recent work proposed a modification to the PHAT transform that allows varying degrees of spectral whitening through a single parameter, β, which has shown positive improvement in target detection in simulation results. This work focuses on experimental evaluation of the modified SRP-PHAT algorithm. Performance results are computed from actual experimental setup of an 8-element perimeter array with a receiver operating characteristic (ROC) analysis for detecting sound sources. The results verified simulation results of PHAT-β in improving target detection probabilities. The ROC analysis demonstrated the relationships between various target types (narrowband and broadband), room reverberation levels (high and low) and noise levels (different SNR) with respect to optimal β. Results from experiment strongly agree with those of simulations on the effect of PHAT in significantly improving detection performance for narrowband and broadband signals especially at low SNR and in the presence of high levels of reverberation.

KEYWORDS: Microphone array, Steered Response Power (SRP), Phase Transform (PHAT), Sound Source Location (SSL)

Anand Ramamurthy

November 19, 2007

EXPERIMENTAL EVALUATION OF MODIFIED PHASE TRANSFORM FOR SOUND SOURCE DETECTION

By

Anand Ramamurthy

Dr. Kevin D. Donohue
Director of Thesis

Dr. YuMing Zhang
Director of Graduate Studies

November 19, 2007

RULES FOR THE USE OF THESIS

Unpublished theses submitted for the Masters degree and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due regard to the rights of the authors. Bibliographical references may be noted, but quotations or summaries of parts may be published only with the permission of the author, and with the usual scholarly acknowledgments.

Extensive copying or publication of the thesis in whole or in part also requires the consent of the Dean of the graduate School of the University of Kentucky.

A library that borrows this dissertation for use by its patrons is expected to secure the signature of each user.

Name Date

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

THESIS

Anand Ramamurthy

The Graduate School

University of Kentucky

2007

EXPERIMENTAL EVALUATION OF MODIFIED PHASE TRANSFORM FOR SOUND SOURCE DETECTION

THESIS

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science in the
College of Engineering at the University of Kentucky

By

Anand Ramamurthy

Lexington, Kentucky

Director: Dr. Kevin D. Donohue,

Databeam Professor of Electrical and Computer Engineering

Lexington, Kentucky

2007

DEDICATION

To Appa, Amma, Arun

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. Kevin D. Donohue for his unwavering support and guidance in this project. I cherish the many discussions that I have had with him throughout this research effort which has improved my understanding in the critical aspects of the subject and spurred me to think independently. Thank you Sir, I have greatly enjoyed working with you.

I would also like to thank Dr. Bruce Walcott, Dr. Robert Heath and Dr. Daniel Lau for agreeing to take part in my committee and provide their valuable insight. I would like to extend my special thanks to Dr. Jens Hannemann for his help throughout this work, my lab mates Shantilal and Arul and all my friends for their help and patience in enduring me through these days.

TABLE OF CONTENTS

APPENDICES

List of Tables

List of Figures

List of Files

ETD_thesis.pdf

CHAPTER 1

Introduction and Literature Review

1.1 Sound Source Localization

Modern society craves better comfort, flexibility, quality of living. Technology has kept up to this growing demand with new generation of applications. Sound source location (SSL) with microphone arrays is one such development which finds importance in day-to-day applications like Bluetooth headsets, automobile speech enhancement, noise cancellation for audio communication, teleconferencing, speech recognition, talker characterization and voice capture in reverberant environments [1-3]. Other specialized applications involving this technology are: speech separation, robot navigation, security surveillance systems and as a key component of many new human-computer interface applications under development [4].

Distributed microphone systems have been considered for applications including advanced human computer/machine interfaces, talker tracking, and beamforming for signal-to-noise ratio (SNR) enhancements [1-3]. Many of these applications require detecting and locating a sound source. For example, application in a meeting or conference environment requires detecting and locating all voices and then beamforming on each voice to effectively create independent channels for each speaker. The failure to detect an active sound source or a false detection can significantly degrade the performance of such systems. As a major research topic, sound source location using microphone array has reached levels of performance where it is being integrated and deployed in real environments. E.g. voice-capture and automatic camera steering products using a 4-element microphone array (by Polycom Inc.) [5] and systems for high performance speech recognition in noisy environments [6, 7]. The primary goal of any SSL system is to ensure acceptable performance in different operational conditions [8].

When it comes to real-world applications, the source location estimates need to meet different reliability constraints. The primary reason for failure of such systems is the poor

performance in adverse environments, such as a room with ambient noise [9]. This problem can be addressed with a judicious decision on microphone array design and choice of a robust SSL algorithm [3, 10].

In general, SSL estimation performance is dependent on factors like:

1) quantity and quality of microphones used
2) microphone placement geometry
3) number of active sources in the FOV
4) ambient noise and reverberation levels

The above factors play a major role in the decision process for SSL. Increasing the number of microphones in the array is the simplest means to achieve marginal performance improvement in adverse environmental conditions. However, in most situations, a modest number of microphones can be used to achieve adequate performance provided the ambient conditions are favorable and microphones are positioned accordingly [10]. The optimal solution for number and geometry of an array is driven by factors like room layout, prevailing acoustic conditions, number and type of sources [11]. So, many practical SSL system designs take into consideration, factors like: the specific application conditions, the hardware availability, and other cost criteria.

1.2 Localization and Tracking

Obtaining the best accuracy forms the primary objective of localization and tracking systems. The sensor configuration and geometry have a strong bearing on performance. The room layout, speaking scenarios, acoustic conditions, and the prevailing environment have to be taken into consideration while designing the system. However, approaches differ depending on overall objective (e.g. detecting single/multiple sources), specific tracking framework, sensor configuration and use of different approaches such as audio, video, or their combinations.

1.3 Acoustic Localization Methods

Among the different localization and tracking techniques, acoustic source localization techniques have following advantages:

a) operational convenience independent of lighting conditions,

b) omni-directional sensing performance and

c) localization independence from visual occlusion.

1.3.1 Time Difference of Arrival: TDOA

Commonly used acoustic source localization algorithms are based on time delay estimation (TDE) or time-difference of arrival (TDOA) technique. The knowledge of microphone position-geometry along with time difference of arrival of the source signal at different microphones pairs is used to estimate the source location. The reliability of a time delay estimate depends on the spatial coherence of the acoustic signal reaching the sensors, and is influenced by the distance between the microphones, the level of background noise and the extent of the room reverberation.

Most of the TDOA schemes are based on estimating the maximum Generalized Cross-Correlation (GCC) between the delayed microphone-pair signals [12]. The GCC is a popular method for estimating time-delays. Its popularity is due to its low computational complexity which is achieved by Fast Fourier Transform (FFT) implementations. Let $x_i(t)$ denote the signal at i^{th} microphone and $X_i(\omega)$ be its Fourier transform over a finite interval $0 \leq t \leq T$. The cross correlation between 2 microphone channels is:

$$\hat{R}_{GCC}(\tau) \triangleq \int_{-\infty}^{\infty} |U(\omega)|^2 \ \hat{P}_{12}(\omega)\, e^{j\omega\tau} \tag{1}$$

where, $|U(\omega)|$ is the weighting function and the cross power spectrum $\hat{P}_{12}(\omega)$ is:

$$\hat{P}_{12}(\omega) \triangleq X_2(\omega)X_1^*(\omega) \tag{2}$$

The superscript $(\cdot)^*$ denotes complex conjugate.

In the GCC method, the weighting function $|U(\omega)|$ is set to '1' in equation 1, and the estimated time-delay $\hat{\tau}$ is given by:

$$\hat{\tau} = arg\max_{\tau}(\hat{R}_{GCC}(\tau)) \quad (3)$$

The performance of GCC suffers in conditions of multi-source presence and even worse for moderate to high levels of background noise and reverberation. In such cases, the GCC with Phase Transform (GCC-PHAT) method is found to have significantly better performance over conventional SSL approaches for TDOA based SSL systems [13]. The weighting function for GCC-PHAT is defined for the equation1 above, as:

$$|U(\omega)|^2 = \frac{1}{|P_{12}(\omega)|} \quad (4)$$

1.3.2 Enhancements to TDOA:

In effort to enhance the accuracy of TDOA estimates and handle multi-speaker cases, Kalman filter smoothing [14] and a combination of TDOA with particle filter approach [15] has been investigated.. The basic Kalman filter is limited to a linear assumption. Kalman filter assumes dynamics to be linear and Gaussian However, most non-trivial systems are non-linear. For example, when the sound source is human, the linearity assumption is not true for sudden changes in source position. Furthermore, in spontaneous speech, short utterances (typically less than a second) that makeup considerable portion of the speech poses further challenges when trying to implement the Kalman filter approach.

In such situations, the Extended Kalman Filter (EKF) where the state transition and observation models need not be linear functions but may instead be differentiable functions. Unlike its linear counterpart, the EKF is *not* an optimal estimator. In addition, if the initial estimate of the state is wrong, or if the process is modeled incorrectly, the filter may quickly diverge [16, 17]. However, the above approaches still encounter difficulties in delivering consistent performance when dealing with spontaneous speech, that is variable in both space (source movement) and is sporadic over time (short intervals of signal energy). Also, the increased computational requirement of complex algorithms prohibits their use in real-time applications.

Single acoustic source localization and tracking applications are found in [18, 19]. However, fast-changing source movements as encountered in spontaneous multi-party speech requires either specific multi-source models [20] or adapting the single-source model to switch between speakers [21] . Some attempts have been made to combine the TDOA and SRP based approaches to alleviate the disadvantages of TDOA based approach [22].

Measures to improve the performance of TDOA based SSL systems designed assuming presence of ideal conditions could still hurt the performance in normal application environments. The following section describes research on a more robust approach (beamformer based).

1.3.3 Steered Response Power: SRP

Most state-of-the-art speech processing systems rely on close-talking microphones for speech acquisition to achieve good performance. But, in the case of multiparty conversational setting like meetings, the setup is often not suitable. For such scenarios, microphone arrays present a potential solution by offering distant, hands-free and reliable audio signal acquisition by making use of beamforming techniques. Beamforming consists of filtering and discriminating active speech sources from noise sources based on their spatial location [23]. The simplest technique is *delay-sum beamforming*, in which a delay filter is applied to each microphone channel before summing them to give a single enhanced output. A more sophisticated filter-sum beamformer that has shown good performance in speech processing applications is super-directive beamforming, in which filters are calculated to maximize the array gain for the look direction [24] . The post filtering of the beamformer output significantly improves desired signal enhancement by reducing background noise.

The localization and tracking of multiple active sources is crucial for optimal performance of microphone-array based systems. Many computer vision systems have been studied to detect and track people [25], but are affected by occlusion and illumination effects. Acoustic source localization algorithms can be implemented to work efficiently in such environments independent of lighting conditions.

1.3.4 Evolution of SRP-PHAT-β

Several weighting functions (filters) have been studied for improving the performance of the conventional SRP, such as: maximum likelihood (ML), smoothed coherence transforms (SCOT), the phase transform (PHAT) and the Roth processor. [12, 26-29]. The difference between the above mentioned approaches to SRP is in the weighting function used in each case which is summarized in the table below, where $P_{x_i x_j}(\omega)$ is the cross power spectrum described in equation 2.

Table 1: Weighting functions used for SRP

Weighting function	**PHAT**	**SCOT**	**Roth processor**
Equation	$\frac{1}{\lvert P_{x_1 x_2}(\omega) \rvert}$	$\frac{1}{\sqrt[2]{P_{x_1 x_1}(\omega) P_{x_2 x_2}(\omega)}}$	$\frac{1}{P_{x_1 x_1}(\omega)}$

The weighting function that is found to be robust to reverberant conditions is the PHAT function [5, 12].

The GCC-PHAT method [30] used for TDOA (refer equations1 to 4), is based on estimating the maximum GCC between the delayed signals and is robust to reverberations due to the influence of the PHAT. The steered response power (SRP) method [31] delays signals from different microphone channels to estimate the power output and is robust to background noise. The advantages of both the methods i.e., robustness to reverberation and background noise are combined in the SRP-PHAT method [5].

Donohue et al. (2007) introduced a modification to the PHAT, referred to as the PHAT-β transform [32], that investigates the effect of changing the degree of spectral magnitude information used by the transform using a single parameter (β) . In this work, performance results of the 'β' parameter were computed using a Monte Carlo simulation of an 8 element perimeter array and analyzed using receiver operating characteristic (ROC) analysis. Results in [32] have shown that standard PHAT significantly improves detection performance for broadband signals. Proper choice of β can result in performance improvements for both narrowband and broadband signals.

1.3.5 Motivation:

Research work on sound source location has focused on algorithms for enhancing detection and localization of targets. SRP along with the Phase Transform (PHAT) weighting has shown promising results as a robust algorithm for detecting sound sources [33, 34]. A detailed analysis focused on target detection performance has shown that a variant of the PHAT, referred to as modified PHAT or PHAT-β [32, 35], actually outperforms the conventional PHAT for SRP for a variety of signal source types and operating conditions (low SNR, high reverberation).

The performance results for PHAT-β demonstrated through simulation results in [32] presented a means to parametrically influence performance of PHAT with respect to signal type and bandwidth of interest. The work described in [32] and subsequently this thesis attempts to evaluate the effect of 'β' for SRP-PHAT based approach in terms of detection performance. Detection performance is assessed using the area under the Receiver Operating Characteristics (ROC) curve [36-38] .

1.3.6 Hypothesis

The objective of this thesis is to verify the results presented in [32] and develop experiments to validate and test the influence of 'β' parameter on target detection performance. Separate tests were designed to study performance with respect to sound source detection in reverberant and noisy rooms and present an effective methodology for its solution.

For an efficient evaluation of the acoustic degradations on SSL performance, this thesis will focus on the implementation SRP-PHAT-β algorithm as a function of source type, reverberation levels, and ambient noise (in terms of SNR), rather than focusing on influence of changes in specific environmental scenario and microphone geometry. Prior knowledge about the time frames where the sources was active is assumed for analysis. This is because a received signal could contain not only segment of interest but also of noise source and periods of silence.

While the focus of the experiments and analysis will be the single-source scenario, the techniques described are applicable to situations involving multiple sources with little modification.

1.4 Organization of the Thesis

Chapter 2 gives an introduction to concepts of beamforming used with respect to the delay and sum beamformer implementation for steered response power computation. The later sections of this chapter discuss the SRP algorithm implementation using the PHAT weighting approach and finally the PHAT-β is introduced for SRP implementation.

Chapter 3 presents the specifications of the experimental setup where the data used for all analysis in this thesis were collected. This chapter also discusses the decision choices made, and other implementation criterion used for computing and analyzing the SRP-PHAT β.

Chapter 4 focuses on the results obtained from the analysis of the data gathered from the experimental setup described in chapter 3. It also presents a case-by-case discussion of the performance results obtained with respect to the simulation results published by Donohue et.al in [32] indicating the agreement of results with those in [32] and also the disagreements.

Chapter 5 summarizes the conclusion and future research directions.

Appendices A at the end of this thesis gives an introduction to the basics of acoustic signal modeling and the parameters involved.

Appendix B is a review of commonly used SSL approaches.

CHAPTER 2

Steered Response Power with modified PHAT (PHAT-β)

This chapter discusses the concepts of beamforming and Steered Response Power algorithms used for SSL. The implementation of PHAT for SRP is discussed in section 2.4 and the final section 2.5 introduces the PHAT- β for SRP implementation and the expected performance improvement for the new algorithm.

An important application of SSL based beamforming has been its use in speech-array applications for voice capture [1, 6, 23, 41-43]. When applied to source localization, the beamformer output is maximized when the array is focused on the target location. The SRP algorithm exploits the multitude of microphones in order to overcome the limitation in estimation accuracy of TDOA based approaches in the presence of noise and reverberation. SRP exploits the spatial filtering ability of a microphone array which further increases its applicability for the SSL problem. SRP also enables the selective enhancement of signal from the source of interest while suppressing other unwanted signals [12, 39]. This property of SRP algorithm makes it a more robust choice for SSL applications [32].

The features of SRP which make it a better approach than TDOA in terms of robustness to reverberation for the SSL problem is discussed in this chapter and a new filter is introduced. This filter is derived from the phase transform (PHAT) [32], which applies a magnitude-normalizing weighting function to the cross-spectrum of two microphone signals.

1.1 Beamforming for SRP

Consider a set of microphones and sound sources at different spatial locations. Let $s_i(t; \vec{r_i})$ be the pressure wave resulting from the i^{th} source. The waveform received by the m^{th} microphone is given by [27]:

$$x_{m,i}(t;\ \overrightarrow{r_m},\vec{r_i}) = s_i(t;\ \vec{r_i}) * h_{p,i}(t;\ \overrightarrow{r_m},\vec{r_i}) + n_m(t) \qquad (5)$$

where, $h_{p,i}(t;\ \overrightarrow{r_m},\vec{r_i})$ is the impulse response of the propagation path from $\vec{r_i}$ to $\overrightarrow{r_m}$ and $n_m(t)$ represents all the noise sources.

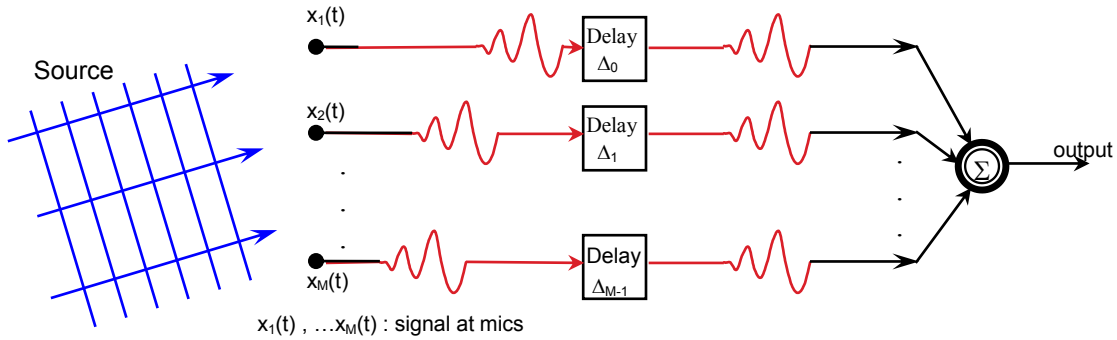


Figure 1: The SRP algorithm using delay-sum beamforming

Figure 1 above shows that for an array of *M* microphones, a delayed and filtered version of the source signal $x_i(t)$ exists in each microphone channel. By time-aligning the delayed versions of $x_i(t)$, the resulting signals can be summed together so that all copies add constructively while the uncorrelated noise signals present in $n_m(t)$ cancel out.

The copies of $s_i(t)$ at each of the individual microphones can be time-aligned by setting the steering delays equal to the negative values of the propagation delays plus some constant delay, τ_0:

$$\Delta_m = \tau_0 - \tau_m; \qquad (6)$$

where, m takes values from 1,2,…..M, τ_0 defines the *phase center* of the array, and is set to the largest propagation delay among all microphones in the array, making all the steering delays greater than or equal to zero. This implies all shifting operations are causal, which satisfies the requirement for practical implementation in a system. This also makes the steering delay values relative to one microphone. Hence, the output equation for *delay-and-sum beamformer* shown in Figure1:

$$y_i(t;\ \Delta_1 \dots \dots \Delta_m) \equiv \sum_{m=1}^{M} x_m(t - \Delta_m) \tag{7}$$

where, $\Delta_1 \dots \dots \Delta_m$ are the M steering delays, which focus or *steer* the array to the source's spatial location or direction and $x_m(\cdot)$ is the signal received at the m^{th} microphone.

The delay-and-sum beamformer output $y_i(t;\ \Delta_1 \dots \dots \Delta_m)$ in equation7, can now be expressed in terms of the microphone signal model $x_{m,i}(t;\ \overrightarrow{r_m}, \vec{r_i})$ of equation5 and the steering delays Δ_m from equation6, giving:

$$y_i(t;\ \Delta_1 \dots \dots \Delta_m) \equiv\ s_i(t - \tau_0;\ \vec{r_i})\ * \sum_{m=1}^{M} h_{m,i}(t - \tau_0 + \tau_m;\ \overrightarrow{r_m}, \vec{r_i})\ + \\ + \sum_{m=1}^{M} n_m(t - \tau_0 + \tau_m) \tag{8}$$

Considering the impulse responses of individual microphone channels $h_{m,i}(t)$ to approximate a band pass filter, the output of the beamformer, as given by equation8, will be a band-limited version of $s_i(t)$ with amplitude M times larger than the signal from any single microphone. The degree, to which the noise signals are suppressed, depends on the nature of the noise. Separating the noise term from equation8:

$$y_i(t;\ \Delta_1 \dots \dots \Delta_m) \equiv\ s_i(t - \tau_0;\ \vec{r_i})\ * \sum_{m=1}^{M} h_{m,i}(t - \tau_0 + \tau_0;\ \overrightarrow{r_m}, \vec{r_i}) \tag{9}$$

Equation9 gives the output of an M-element, delay-and-sum beamformer in time domain. The frequency domain representation of equation9 is:

$$Y_i(\omega) \equiv \sum_{m=1}^{M} H_{m,i}(\omega)\ S_i(\omega)\ \ e^{-j\omega\,\Delta_m} \tag{10}$$

1.2 The Steered Response Power

The *steered response* is generally a function of *M* steering delays, $\Delta_1 \ldots\ldots \Delta_m$. The steering delays are used to aim a beamformer (acoustically focus the array) at a particular position or direction in space. The steered response is obtained by sweeping the focus of the beamformer. When the focus of the beamformer corresponds to the source location, the time-aligned signals in the microphone channels add up and the power of the steered response reaches maxima due to constructive interference. The equation8 can be re-written as:

$$y_{m,i}(t;\ \overrightarrow{r_m}, \overrightarrow{r_i}) = \int_{-\infty}^{\infty} h_{m,i}(t - \tau_0 + \tau_m;\ \overrightarrow{r_m}, \overrightarrow{r_i})\ s_i(t - \tau_0;\ \overrightarrow{r_i})\ d\lambda + \sum_{k=1}^{K} \int_{-\infty}^{\infty} h_{m,k}(t - \tau_0 + \tau_m;\ \overrightarrow{r_m}, \overrightarrow{r_k})\ n_k(t - \tau_0 + \tau_m;\ \overrightarrow{r_k})\, d\lambda + n_m(t) \tag{11}$$

where, $h_{m,i}(\cdot)$ represents the impulse response of the microphone and propagation path from $\vec{r_i}$ to $\overrightarrow{r_m}$, $n_k(\cdot)$ represents correlated noise sources resulting from sources and $n_m(t)$ is the uncorrelated electronic noise from the sensor, amplifier, and digitizer on the m^{th} microphone channel.

For reverberant rooms, the impulse response in equation11 can be separated into a signal component (direct path only) and noise component (includes multi path signals also). If the primary operations on the sound source are the effective delays from multiple reflections and attenuation from the propagation paths, the transfer function can be represented as:

$$h_{m,i}(t;\ \overrightarrow{r_m}, \overrightarrow{r_i}) = \ h_{m,i}(t) = \sum_{n=0}^{N} a_{m,i,n}\left(t - \tau_{m,i,n}\right) \tag{12}$$

where, $a_{m,i,n}(t)$ denotes the n^{th} path of the effective impulse response for the source at $\vec{r_i}$ and microphone at $\overrightarrow{r_m}$, and $\tau_{m,i,n}$ is the corresponding path delay. The direct path corresponds to $n = 0$. As the algorithms for SSL operate on small time segments, only target and noise scatterer delays falling in that segment contribute to the SRP estimate within the frame. For a single SRP frame, equation7 can be expressed in the frequency domain with the substitution of equation8 to give:

$$\hat{Y}_{m,i}(\omega) = \sum_{i=1}^{N_T} \hat{S}_{i,l}(\omega) \sum_{p|\tau_{m,i,n}} \hat{A}_{m,i,n}(\omega)\, e^{j\omega\,\tau_{m,i,n}} + \sum_{k=1}^{K} \widehat{N}_k(\omega) \sum_{p|\tau_{m,i,n}} \hat{A}_{m,i,n}(\omega)\, e^{j\omega\,\tau_{m,i,n}} + \widehat{N}_m(\omega) \tag{13}$$

where, $\hat{S}_{i,l}(\omega)$ is the Fourier transform of the i^{th} source $s_i(t)$ while $\widehat{N}_k(\omega)$ and $\widehat{N}_m(\omega)$ are the Fourier transforms of the correlated and uncorrelated noise sources, respectively for the m^{th} channel. N_T is the number of target sources, K is the number of noise sources, and the inner summation index p, denotes summing the signal components.

1.3 The Phase Transform (PHAT)

The heart of SRP is the filter-and-sum (or delay-and-sum) beamforming operation, which results in noise power reduction proportional to the number of uncorrelated microphone channels used. Uncorrelated noise typically results from the independent (electronic) noise on each microphone channel. Correlated noise, on the other hand, results from coherent noise sources in the room, like sources outside the FOV, secondary targets and reverberation. Correlated noise presents greater challenges for beamforming than uncorrelated noise, and therefore will also be incorporated into this analysis. Approaches to deal with correlated noise from independent sources and reverberation have included various type of spectral weighing involving the generalized cross correlation (GCC).

If the noise spectrum is known, maximum likelihood weights can be developed to deemphasize low SNR spectral regions [33, 40]. If the noise spectrum is not known, a

phase transform (PHAT), can be applied that effectively whitens the signal spectrum [26, 33, 40, 41]. This approach is very popular when correlations are done for creating SRP likelihood functions or simply estimating time delays. Many claim that this is especially useful in reverberant environments [26]. It was shown in [33] that the PHAT is actually the optimal weighting strategy for minimizing the variance of the time delay estimate.

The general PHAT function is denoted as follows,

$$\hat{\theta}_{m,i}(\omega) = \frac{\hat{Y}_{m,i}(\omega)}{|\hat{Y}_{m,i}(\omega)|} \tag{14}$$

where, $\theta_{m,i}(\omega)$ is the weighting function aimed at emphasizing the true source over the undesired extrema and $\hat{Y}_{m,i}(\omega)$ is the signal spectrum described in equation9. Just as with the phase transform, this filter whitens the microphone signal spectrum. This whitening technique effectively flattens the signal spectrum. By whitening the microphone signals, SRP can be used effectively in microphone-array applications. The effect of PHAT on SRP output accuracy is better than other similar weighting functions under realistic (reverberant) operating conditions [42]. The hypothesis is that the SRP-PHAT will peak at the actual source location even when operating conditions are noisy and highly reverberant.

1.4 Partial whitening Transform: PHAT-β

While results from previous research work has shown that PHAT processing is optimal for SRP [33], there has not been considerable research to study how well targets of interest can be separated from noise peaks related to detection performance (especially at low SNR's and in presence of noise). In addition, there has been no detailed comparison between the nature of the signal bandwidth and the actual PHAT performance.

In radar and sonar systems where PHAT was primarily used, the spectrum for the signal of interest is mostly narrowband in nature. Under such conditions, PHAT has shown significant improvement in robustness compared to other weighting functions for use with SRP algorithm. However, the spectral content of speech signals fluctuates (a mixture of narrowband and broadband) and is subject to change with nature and type of the source.

For such a situation, the SRP weighting function discussed in [32], can be used to control the whitening effect on a part of the spectral range of the signal will be beneficial.

The research work presented in this thesis investigates the effect of a modified version of PHAT from [32] to parametrically control the level of whitening influence on the magnitude spectrum. This transform referred to as PHAT–β and defined as:

$$\hat{\theta}_{m,i}(\omega,\beta) = \frac{\hat{Y}_{m,i}(\omega)}{\left|\hat{Y}_{m,i}(\omega)\right|^{\beta}} \tag{15}$$

where, compared to equation10, β is the additional parameter that controls the extent of spectral whitening and can take values in the range $0 \leq \beta \leq 1$. When $\beta = 1$, equation11 becomes the conventional PHAT (equation10) where the normalized signal spectrum $\hat{\theta}_{m,i}(\omega,\beta)$ becomes 1 for all frequencies. When $\beta = 0$ the denominator is 1 and the PHAT-β has no effect on the original signal spectrum. Therefore, by varying β between 0 and 1, different levels of spectral normalization are achieved.

1.4.1 Expected effect of PHAT- β:

To obtain improvement in signal SNR, a matched filter weighting can be implemented to yield an optimal signal-to-noise ratio enhancement. But, for this a prior knowledge of the signal spectra is required for the filter design. This information is often not practical to obtain, especially in the case of human speech, where source and noise spectra change from frame to frame. The PHAT-β is expected to perform well in such situations, though the PHAT does not always guarantee an improvement in the overall SNR.

For wideband signals with significant non-uniformity over the spectrum, the PHAT tends to enhance SNR by increasing the signal energy over the spectrum more than that of the noise components. Also if strong resonances occur due to reverberation, the influence of 'β' is affected relative to other spectral components. On the other hand for narrowband signals, the PHAT increases the low-power regions of the original spectrum containing little or no signal energy, which can reduce the SNR.

The plots in Figures 2 to 7 show an example of the effect of change in β values of the modified PHAT transform discussed in this thesis in terms of its effect on the signal in time domain (Figures 2, 4, 6) and their PSD's (Figures 3, 5, 7) respectively. The signal used for generating the above plots was a 25ms segment from a voiced speech sample with the person uttering the alphabet: "a" in a single microphone channel at a sampling rate of 44.1 kHz.

The first graph (Figure 2) is the power distribution for frequencies within nyquist range, which is similar to a voiced signal spectrum with no PHAT weighting. The signal spectrum is a clear indication of voiced speech with relatively high energy in the lower end of the spectrum (below 6kHz). Figure 3 is an amplitude-time plot of the original source signal where the 'β' value was set at 0, i.e., no PHAT.

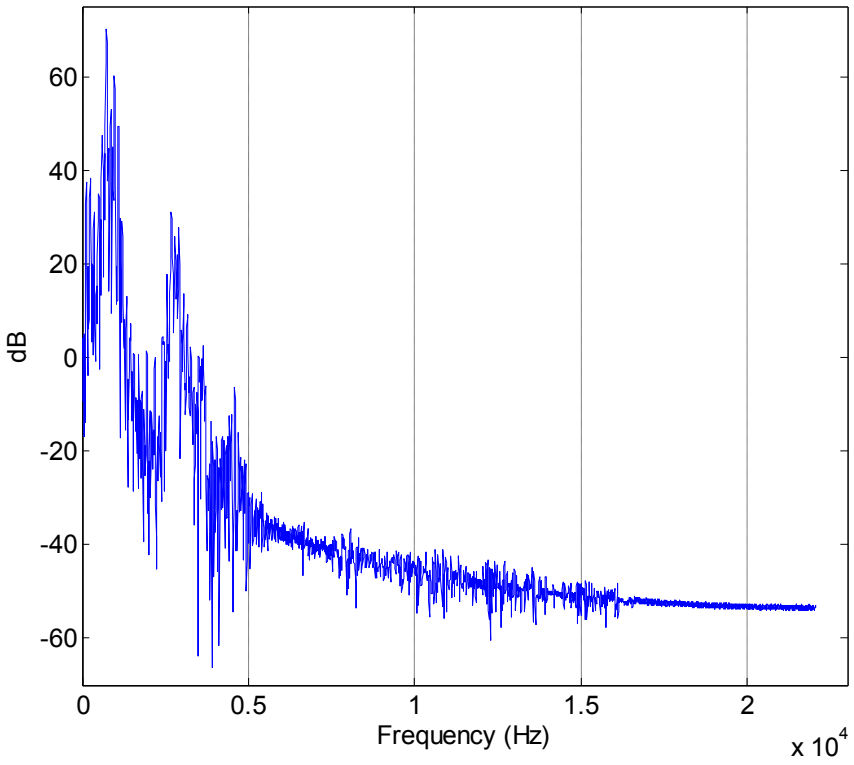


Figure 2: power distribution of the speech segment with β = 0
i.e., no PHAT

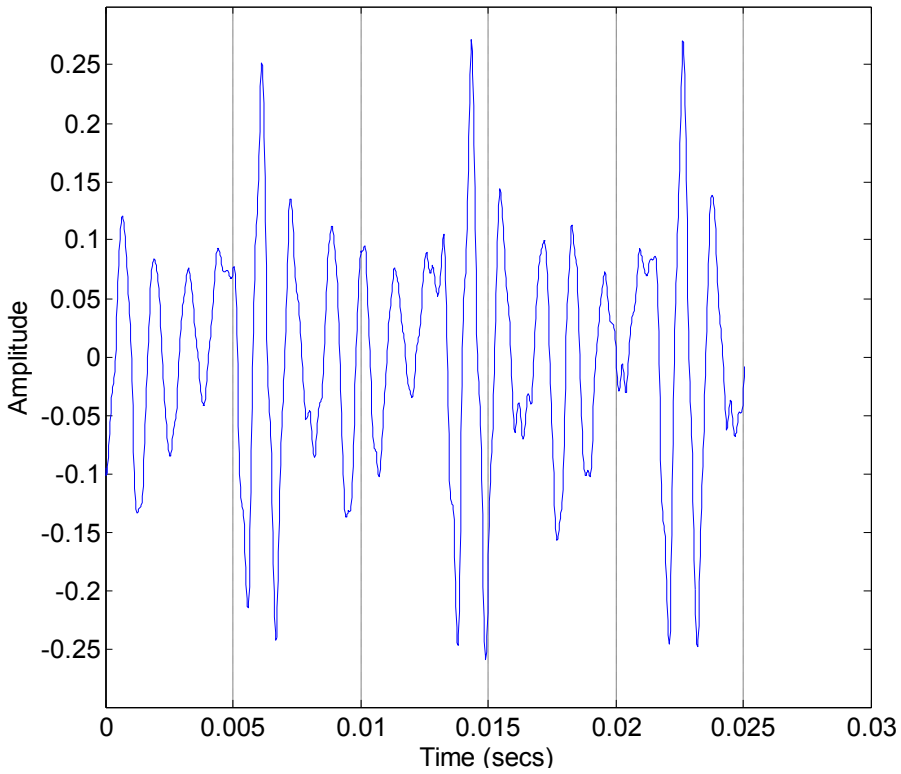


Figure 3: Time series plot of speech segment with β = 0
i.e., no PHAT

The effect of PHAT whitening (β = 1) is shown by the power distribution plot in Figure 4, which is similar to a white noise signal containing equal content of all frequencies within the Nyquist range. Compared to the original signal in figure 2, there is an equal distribution of power for all frequencies of interest due to the effect of setting β = 1. Even high frequency components beyond the voiced speech bandwidth range (noise) are emphasized.

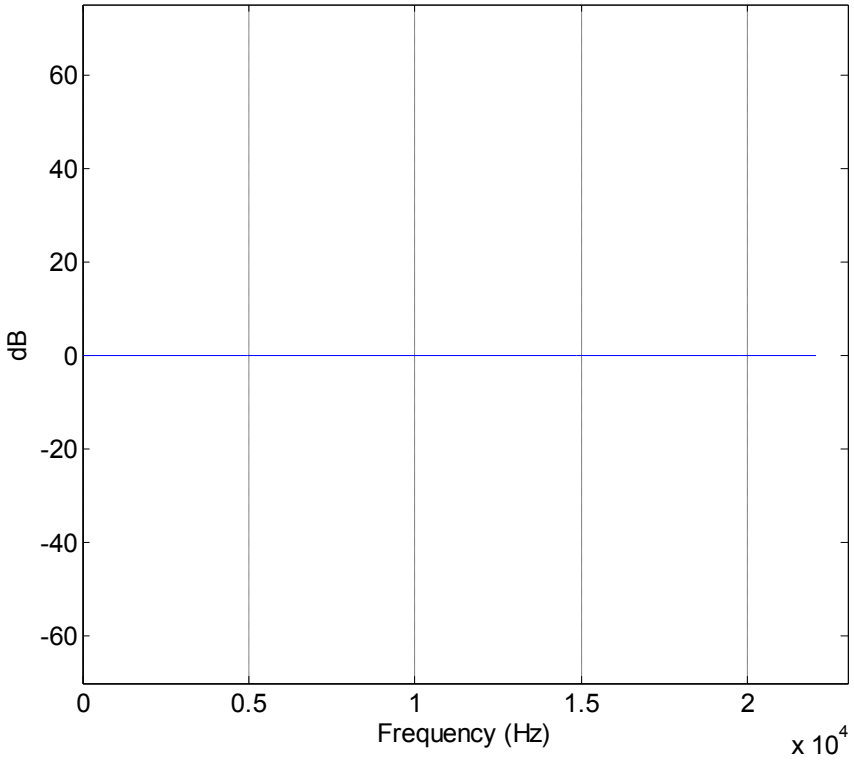


Figure 4: power distribution of speech segment with β = 1
i.e., after conventional PHAT transform, when all spectral components are normalized

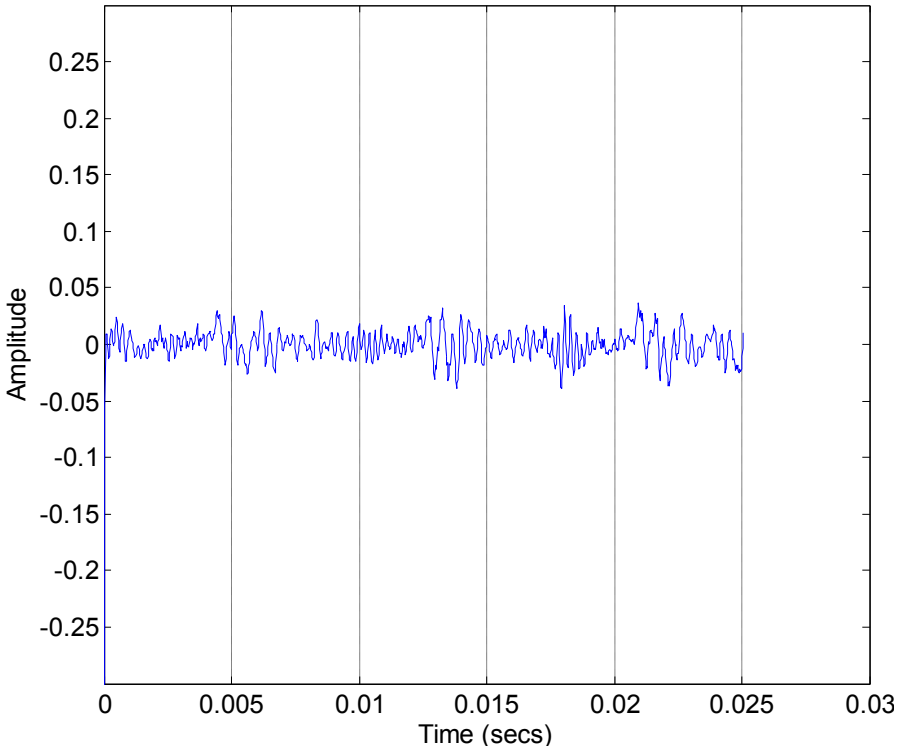


Figure 5: Time series plot of speech segment with β = 1
i.e., after conventional PHAT transform

The effect of PHAT-β transform (partial whitening transform), where $0 \leq \beta \leq 1$ is shown in the power distribution in Figure.6 where β was set at 0.6. Comparing the spectrum in figure 6 to figure 2 and 4, clearly shows the effect of controlling the whitening using β. The spectral region beyond 6 kHz has been emphasized relative to the frequencies of interest based on the level of whitening specified by β. The corresponding effect of PHAT-β on time signal is shown in Figure.7

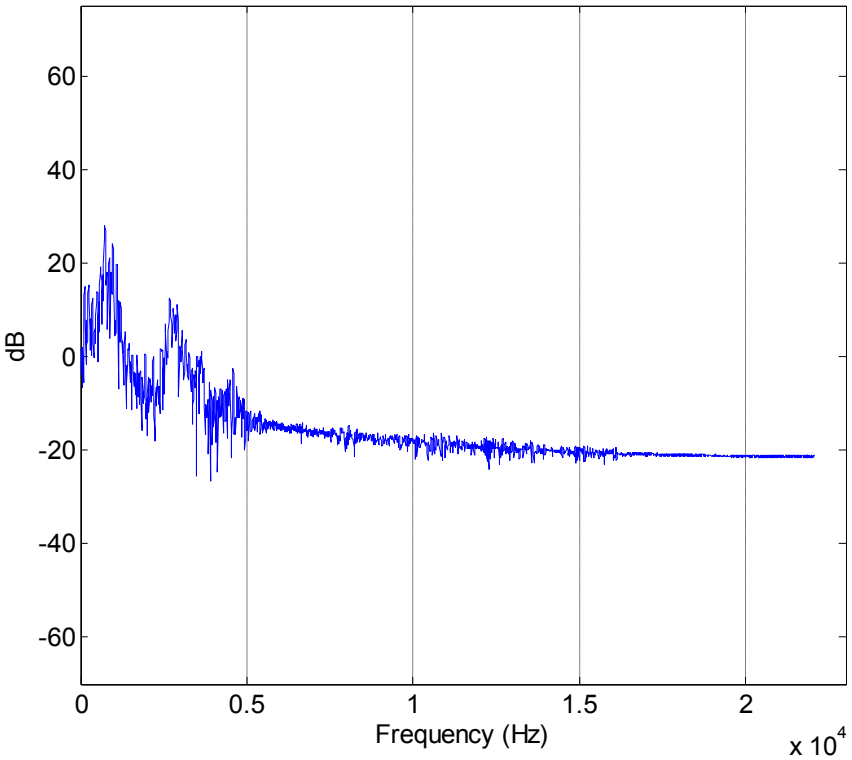


Figure 6: power distribution of Speech segment with β = 0.6
i.e., after partial PHAT transform

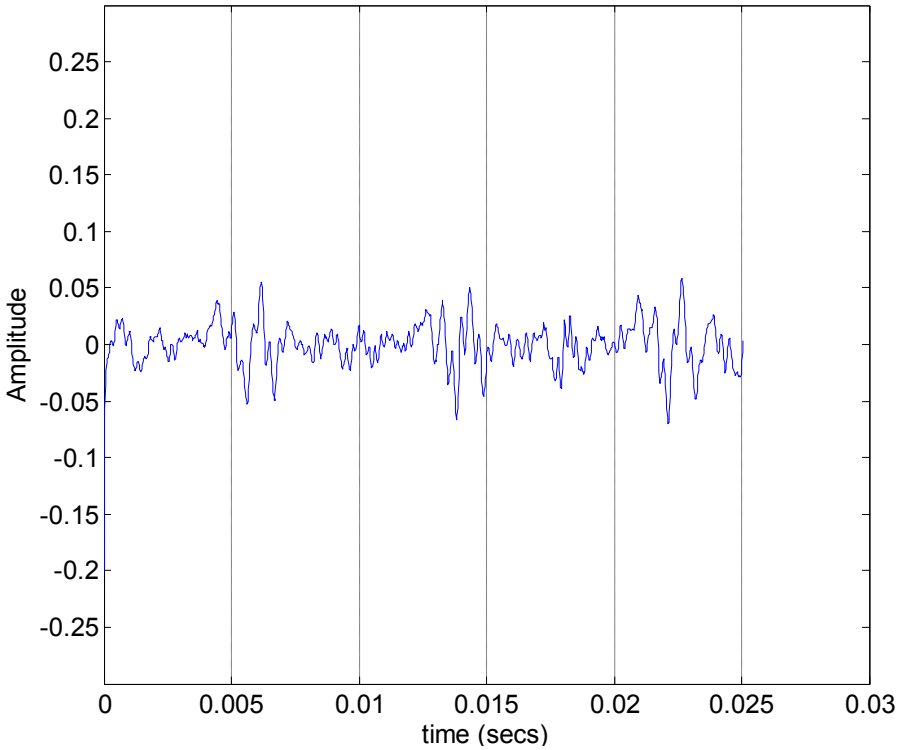


Figure 7: Time series plot of speech segment with β = 0.6
i.e., after partial PHAT transform

1.4.2 SSL improvement with PHAT- β:

The images in Figure 8 show the overall effect of '*β*' on SSL performance using SRP-PHAT. Each pair of images corresponds to SRP image obtained using a single value of '*β*' mentioned beneath the images for experimental data explained in chapter 4 for a narrowband signal sample at high SNR and for low room reverberation levels. The actual source location was at center of the black circle. The SRP images shown in Figure 8 were generated from experimental data described in chapter 3. The SRP images are shown for different values of *β*, with (a), (b), (c) showing the actual SRP intensity image and (d), (e), (f) are SRP images with threshold at '0' (all negative SRP values set to '0').

The results in Figure 8 show a clear improvement in SRP images with respect to reduction in noise peak values in the SRP image. However, for $\beta = 1$, there is increase in number and amplitude of false peaks that hurts SSL performance. The influence of PHAT and PHAT-*β*, on SSL performance for different situations is discussed in-detail in Chapters 4 & 5.

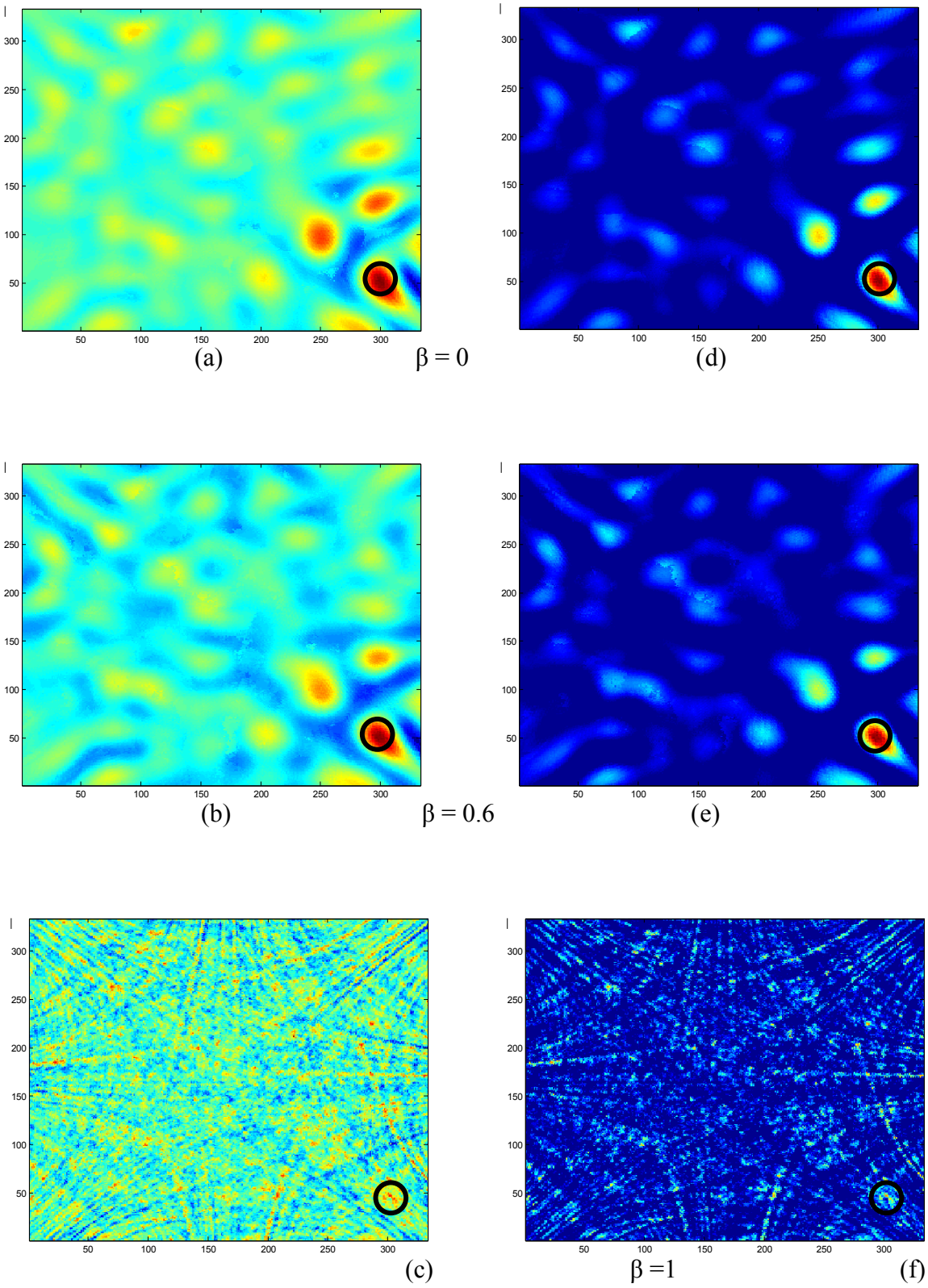


Figure 8: Effect of PHAT-β on SRP image

CHAPTER 3

Experimental setup and Design

This chapter examines the purpose and design of the experimental setup used to collect the data. The purpose of the experiment was to collect data for analysis in conditions similar to what was used to produce the simulations in [32]. It includes details about the test environment, the test signal types, noise levels, hardware setup and also details on the decisions taken during the implementation of SRP-PHAT-β.

1.5 Test environment

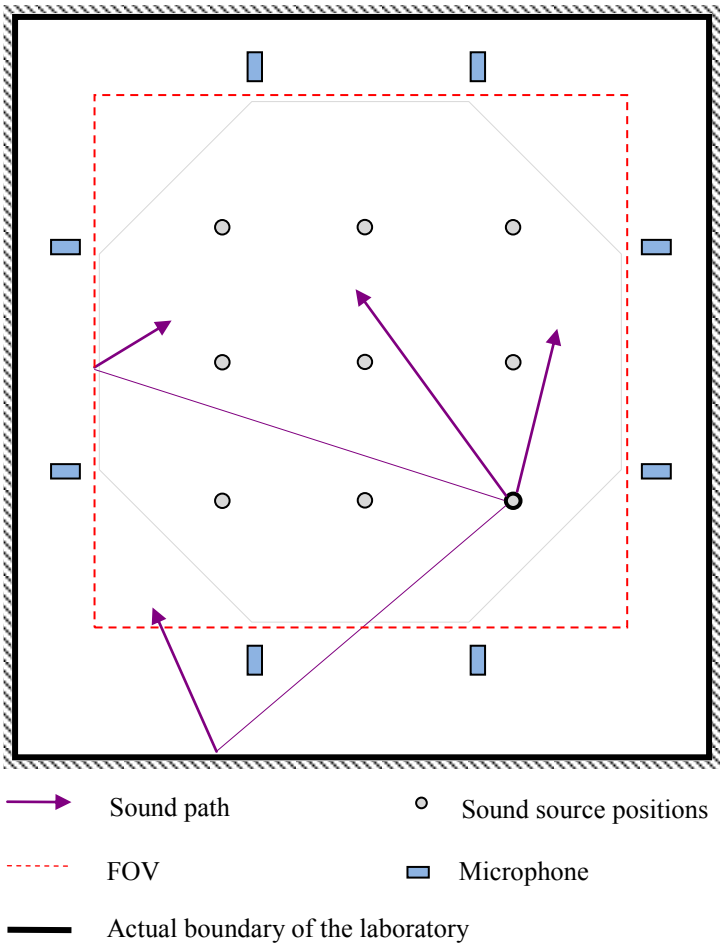


Figure 9: Test environment setup

The experimental room was set up for data collection at the Audio lab facility in the Center for Visualization and Virtual Environments at the University of Kentucky. Figure 9 represents the experiment space marking the FOV (dotted lines), on which the microphones constituting the array were mounted. A cage was built inside the laboratory (black line) with components from 80/20 Inc. *The Industrial Erector Set*. The data collection and processing was driven by two AMD dual-core computers running Ubuntu Linux. Each computer is connected to Delta 1010 card by M-Audio and supports 8 analog input channels and 8 analog outputs [43]. In addition, acoustic treatments can be mounted on the wall of the cage to realize various noise and reflectively properties such an 1.125 inch soundproof foam® (Chambersburg, PA) to reduce reverberation levels and plexi glass (high reverberation) were used. The dimensions of the room used to run the experiments for analysis were: 3.66m for both length and width, and 2.22m for the height. The average speed of sound was estimated using the measured delay of arrival between 2 microphones for sound from a predetermined source location. It was calculated at 346.2 m/s on the day of the experiment.

For the data collection, perimeter array geometry was used, consisting of 8 omni-directional microphones (EMC8000, Behringer) as shown in Figure 9, where the microphones formed an equilateral octagon of dimension 1.284m. Each microphone was placed at a height of 1.57m from floor level and 28cm perpendicular from the cage surfaces. The actual microphone positions were verified using a laser measuring device. These details are summarized in table 1 below.

Table 2: Summary of room setup for data acquisition

Room properties	**Parameters**
Length & Width	3.66m
Height	2.22m
Velocity of sound	$346.2ms^{-1}$
Mic array geometry	8 mics as vertices of an Equilateral octagon
Microphone spacing	1.284m
Source height	1.57m

During each data capture experiment, the sound source (speaker) was moved inside a fixed region within the FOV and placed at predetermined locations shown in Figure 9. At each source position, the sound source was placed along 2 orientations (the speaker facing 2 opposite directions) and data from all 8 microphones were recorded.

To vary the room reverberation levels, the material used for the room wall was switched between an acoustic foam (low reverberation) and plexi glass (high reverberation).

Soundproof Foam: While the acoustic foam provided increased absorption of multipath signals inside the FOV that would otherwise cause reverberation, depending on the thickness of the foam (1.125 inches for the experiment), low frequency components pass through the foam while others are attenuated. This also includes the noise from outside the FOV.

Plexi glass: Plexi glass walls act as excellent reflectors resulting in a worse case multipath scenario inside the FOV. Also, while the plexi glass effectively increases reverberant conditions inside FOV, it blocks noise from outside the FOV.

The reverberation time is defined as the time it takes for the acoustic pressure level to decay to one-thousandth of its former value, a 60 dB drop, also commonly referred to as the RT_{60} of the space. RT_{60} time for the experimental environments (foam and plexi) was measured using recordings from a white noise burst. In order to get accurate RT_{60} value white noise was played loud enough and long enough for the diffuse sound in the room reached steady state. The source should be about 2 meters away from the measurement mic so that the direct path does not dominate the recording. Then the white noise source was abruptly stopped but the recording continued until the sound levels fell below the noise floor. The beginning and ending parts of the recorded signal were used to estimate the signal power and noise floor power. The roll-off of sound from the room reverberation is found based on these 2 estimates. The slope of the roll-off is estimate in dB per second and the amount of time for a 60dB drop in sound is calculated as RT_{60} time. The RT_{60} time for foam was measured at 0.249 seconds while that of the plexi glass was measured to be 0.565 seconds.

1.6 Test signals used

1.6.1 Selection of signal types:

Two input signal types were used to drive the source speaker. One was impulse response to a Butterworth filter of order 4, with a lower 3dB cutoff at 400Hz and upper cutoff frequency at 600Hz for the narrowband signal, and 5600Hz for broadband signal. The Butterworth impulse response was chosen due to its maximally flat spectrum in the pass and stop bands for a uniform distribution of spectral power, while its impulse response is a causal signal with the appropriate phase spectrum. This signal generation resulted in an impulse-like signal from which performance for narrow and broadband signals could be inferred.

In addition to the impulse signal, a colored noise signal was generated from a white noise source using a band pass filter with a lower 3dB cutoff of 400Hz, and upper cutoff frequency of 600Hz for the narrowband signal, and 5600Hz for broadband signal. Colored noise was selected as a test signal because its power spectrum covered all frequencies in the range interest.

The selection of impulse and colored noise signal sources helps in analyzing the performance of in terms of a signal that is spread out in time (colored noise) and that which exists only for a small time interval (impulse). And, the broadband and narrowband variations help analyze performance in terms of signals that have different spectral characteristics. All signals were generated at a sampling rate of 32 kHz. They were later down sampled to 16 kHz for analysis to reduce the size of the actual audio data file storage in computer hard drive. The downsampling to 16 kHz did not affect the performance because the bandwidth of interest is in the range of 300 Hz to 6 kHz.

Table 3: Summary of signals used to drive the source

Signal type ↓	**Bandwidth**	
	Narrowband	*Broadband*
Impulse signal	400 Hz – 600 Hz	400 Hz – 5600 Hz
Colored noise	400 Hz – 600 Hz	400 Hz – 5600 Hz

1.6.2 Signal SNR

For a better understanding of the effect of β for signals with at different SNR levels, each test signal sequence was constructed with 6 segments of different SNR levels, each separated by a time interval of 1 sec and with a 3dB drop from the previous level. The waveform is as shown in Figure 10 below.

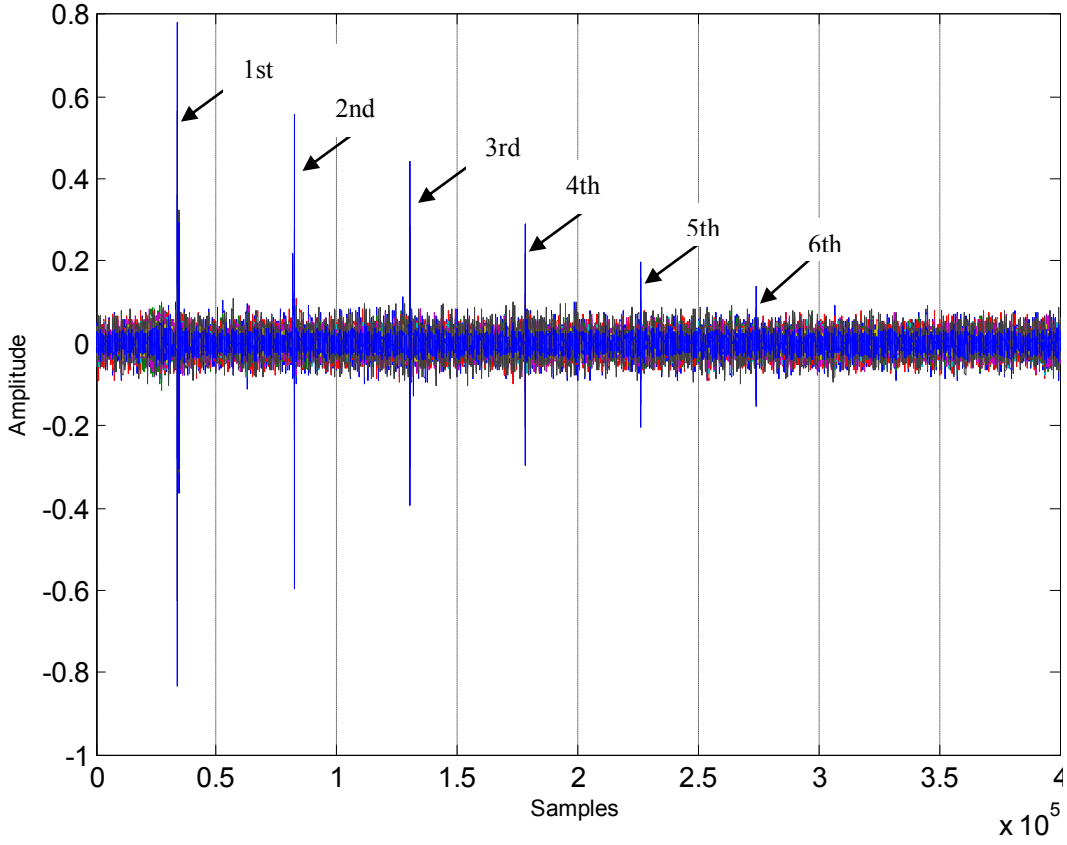
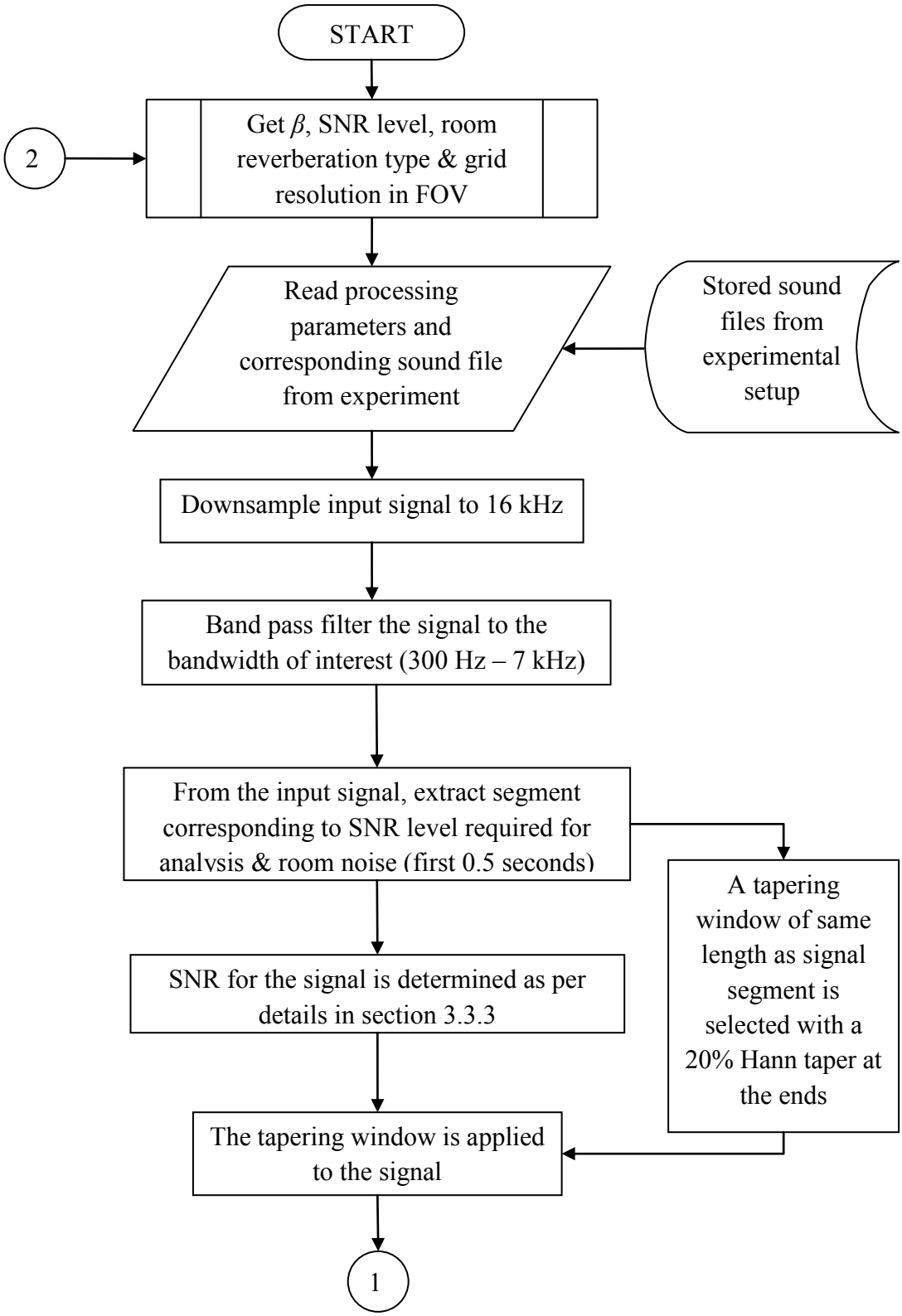
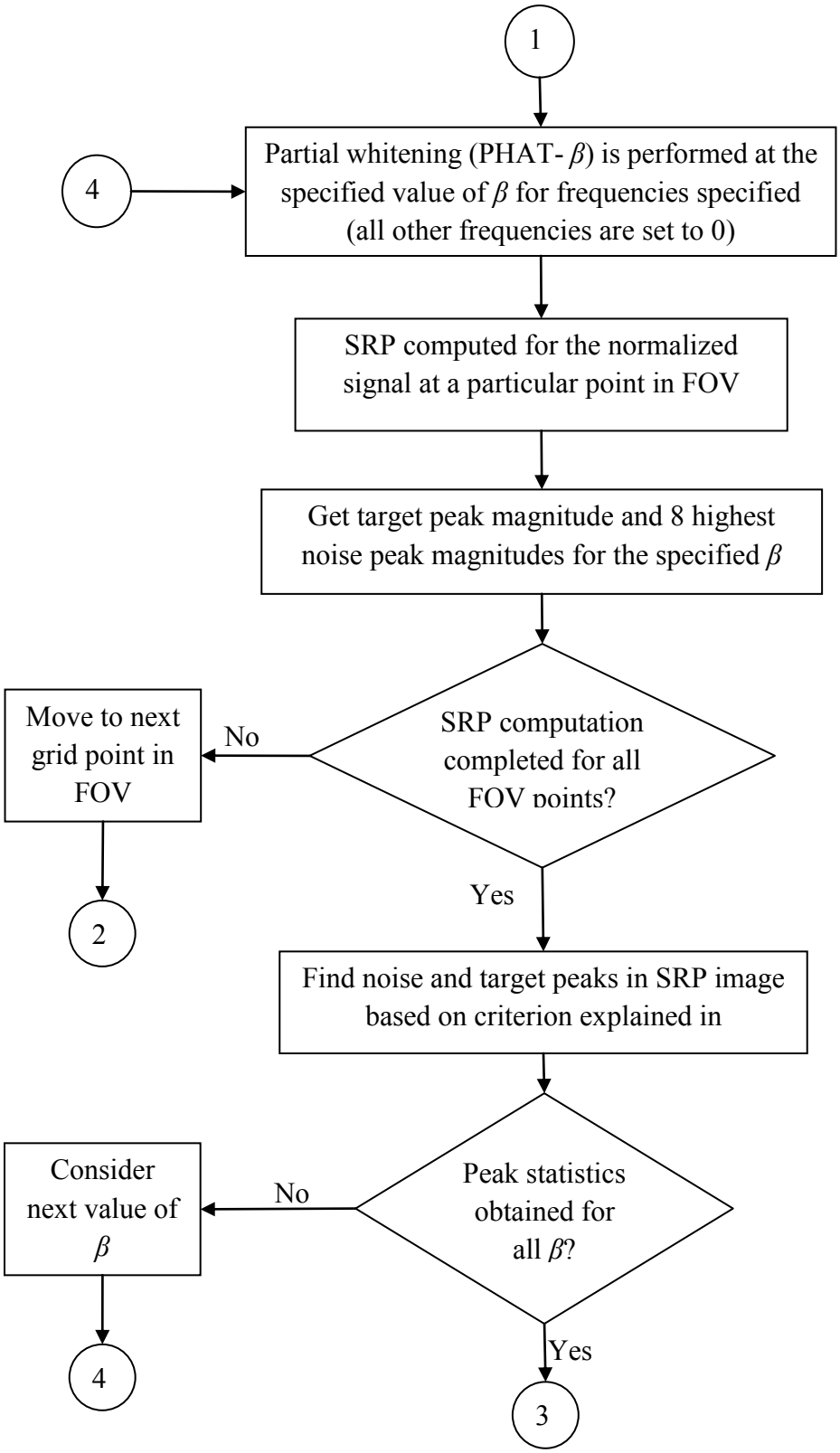


Figure 10: Input waveform

1.7 Algorithm implementation

The implementation of the SRP-PHAT-β algorithm is described in the flowchart below in figure 11 below.


START
2
Get β, SNR level, room reverberation type & grid resolution in FOV
Read processing parameters and corresponding sound file from experiment
Stored sound files from experimental setup
Downsample input signal to 16 kHz
Band pass filter the signal to the bandwidth of interest (300 Hz – 7 kHz)
From the input signal, extract segment corresponding to SNR level required for analysis & room noise (first 0.5 seconds)
A tapering window of same length as signal segment is selected with a 20% Hann taper at the ends
SNR for the signal is determined as per details in section 3.3.3
The tapering window is applied to the signal
1


1
4
Partial whitening (PHAT- β) is performed at the specified value of β for frequencies specified (all other frequencies are set to 0)
SRP computed for the normalized signal at a particular point in FOV
Get target peak magnitude and 8 highest noise peak magnitudes for the specified β
SRP computation completed for all FOV points?
No
Move to next grid point in FOV
2
Yes
Find noise and target peaks in SRP image based on criterion explained in
Peak statistics obtained for all β?
No
Consider next value of β
4
Yes
3

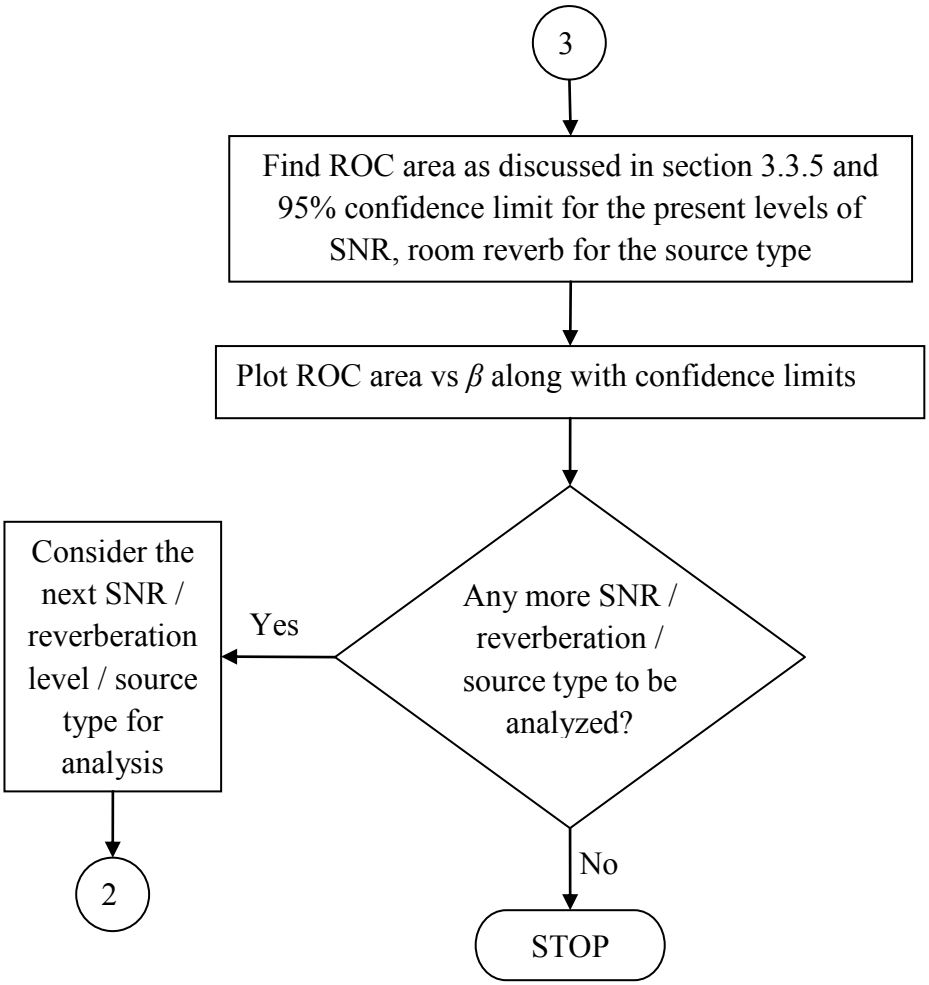


Figure 11: Flowchart for implementation of the SRP-PHAT- β

1.7.1 Analysis parameters

a) Grid spacing

The output of SRP is an array of values for each grid point inside the FOV. Selection of an appropriate grid resolution plays an important role in SSL accuracy by avoiding quantization errors [32]. For this thesis, the tolerance level for loss due to quantization error was set at 3dB. To ensure this limit will not exceed the 3dB limit for the frequencies of interest (300Hz – 5.4kHz), the grid resolution (Δ_{grid}) inside the FOV was computed considering the worst case frequency: f_h (highest frequency in the signal) and a spacing bound Δ_{grid} of 0.02m was set according to equation(15) from [32]:

$$\Delta_{grid} \leq \frac{0.4422 \cdot c}{\sqrt{d} \cdot f_h} \tag{16}$$

where, c is the velocity of sound measured and d = 2, is the number of coordinate dimensions where the source movement is considered.

b) β values used

The signals recorded using the microphone array was analyzed for β values between 0 & 1. Because the range of β values that showed significant improvement in performance of SRP were between 0.6 to 0.8, the analysis for this range included β increments of 0.05 in this range and at a 0.1 increment otherwise.

Table 4: Step size for β

	Step size for β increment	
	0.6 to 0.8	*otherwise*
Step size	0.05	0.1

c) Band pass filtering

The signal spectrum of interest is between 300 Hz to 5.6 kHz. So, the acquired signal is band pass filtered between 300 Hz and 7 kHz to remove high frequency components (>7 kHz) and eliminate the low frequency noise (< 300Hz). The effect of this filtering operation is evident in Figure.12, which shows the filtered version of the raw signal from Figure.10 indicating significant reduction in levels of background (room) noise.

As indicated in Figure 12, the statistics for room noise were computed based on signal segment from the first 0.5 seconds of the signal. This ensured that noise segment selected contains the steady state room noise.

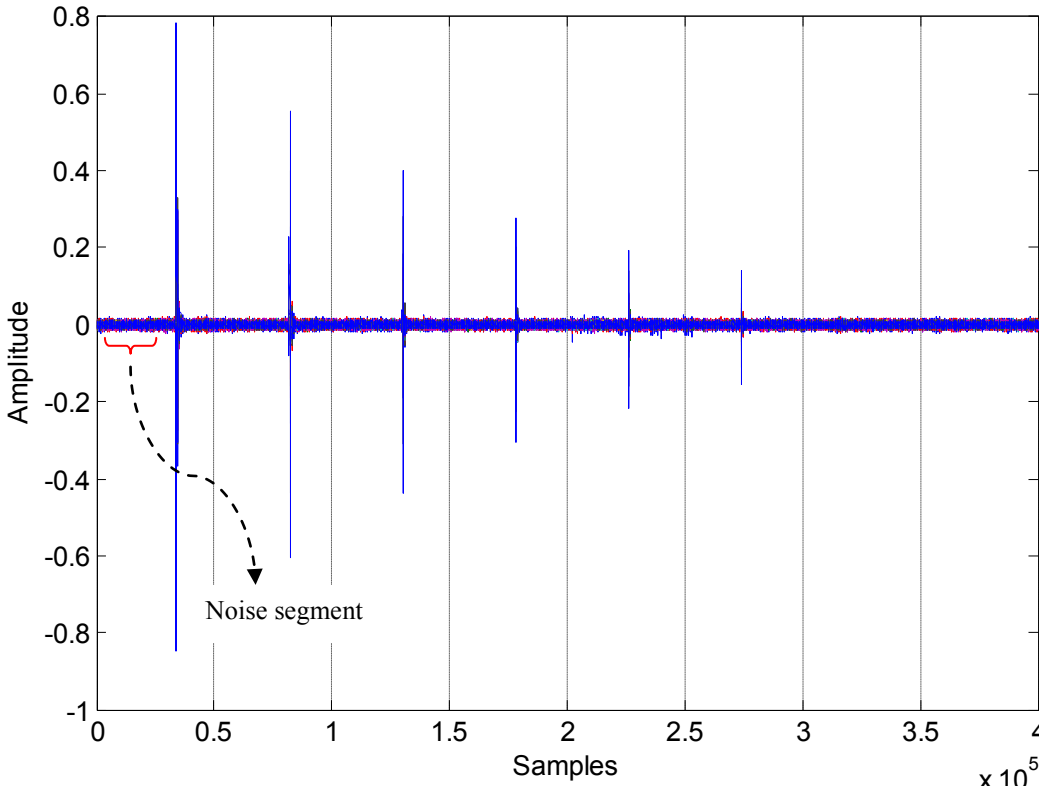


Figure 12: Band pass filtered signal

1.7.2 Tapering window

With prior knowledge of the time frames where the signal of interest existed, the signal segment is selected to contain the source sound. For all analysis in this thesis, the segment is selected as a window that is centered on the occurrence of maximum absolute signal amplitude corresponding to a particular SNR of interest.

The ends of the selected signal segment are tapered to remove abrupt discontinuities that could cause high frequency artifacts in the SRP image. The tapering is implemented by multiplying the signal segment $x_{m,i}(t)$ with a Hanning window $h_t(t)$, of length equal to the signal segment but with a 20% tapering at the 2 edges.

$$x_t(t;\ \overrightarrow{r_m}, \vec{r_i}) = x_{m,i}(t;\ \overrightarrow{r_m}, \vec{r_i}) * h_t(t) \tag{17}$$

The tapering effect on the signal is shown in Figure.14 and the un-tapered signal is in Figure.13. The reduction in pixilation due to tapering is clearly visible in SRP image of Figure.15 (right, compared to the one on left).

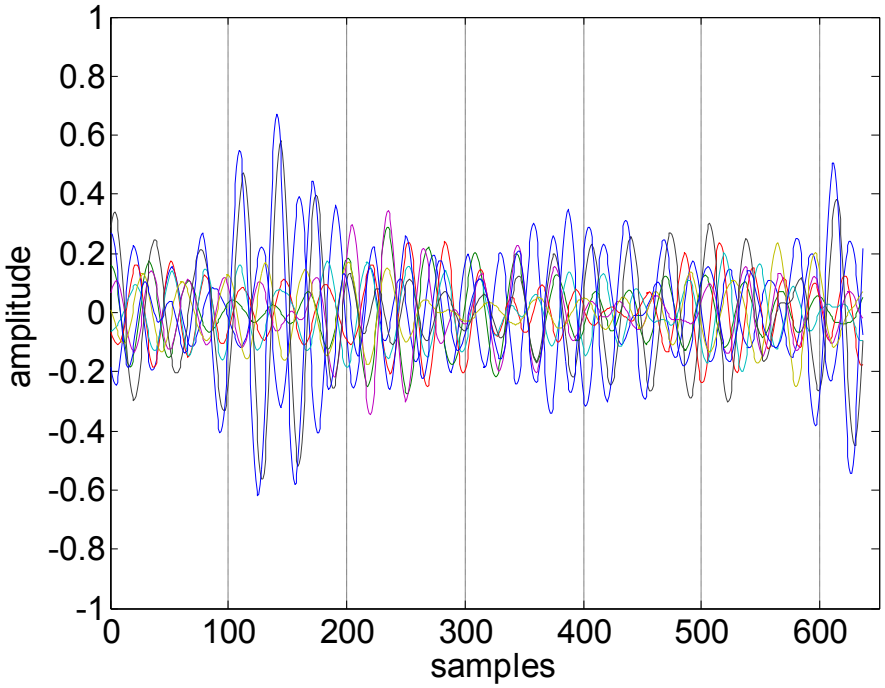


Figure 13: Selected segment before tapering

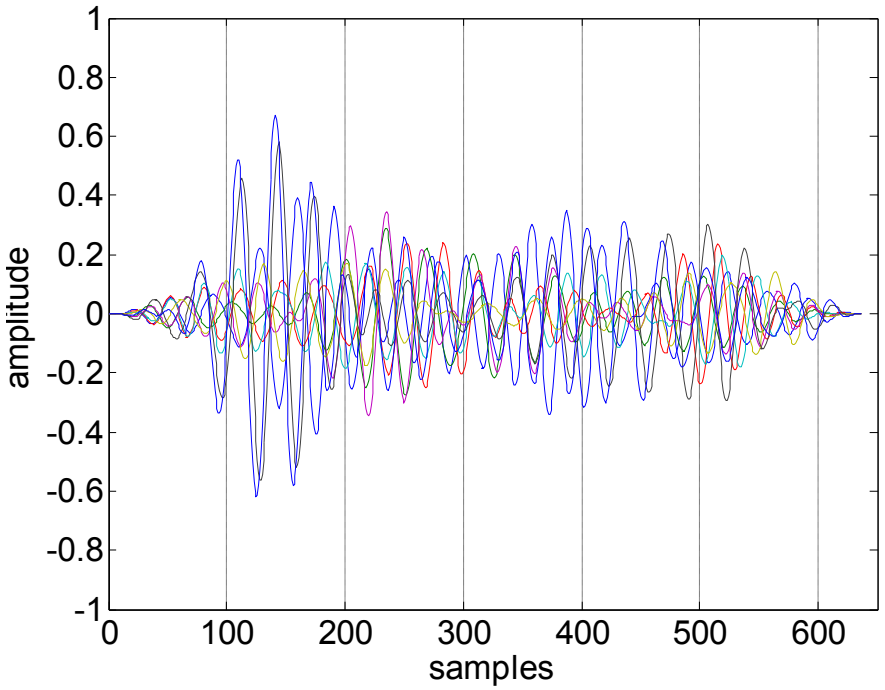
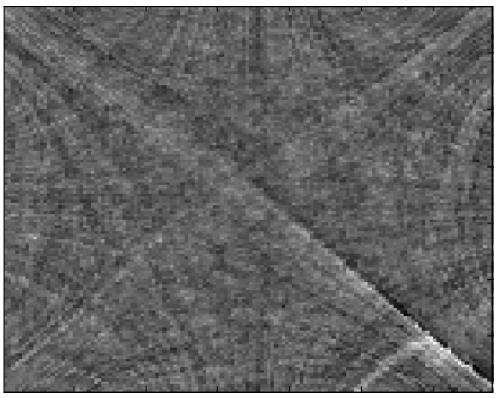
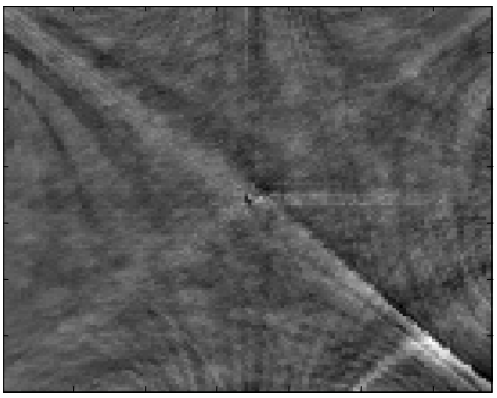


Figure 14: Signal segment after tapering at the ends



Pixilated SRP image before tapering



Tapering results in smoother SRP image

Figure 15: Effect of tapering on SRP

1.7.3 Signal SNR calculation

To calculate the signal SNR, the average power is computed for every signal segment before averaging over all channels. Consider $x_{m,i}(t)$ to be the signal from a source located at $\vec{r_i}$, received by a microphone located at $\overrightarrow{r_m}$. The signal envelope for the segment of interest is:

$$x_{env}(t) = \left|hilbert(x_{m,i}(t))\right| \tag{18}$$

Then RMS value of the signal envelope is determined:

$$x_{rms} = \sqrt[2]{mean(x_{env}(t))} \tag{19}$$

Using the statistics of room noise extracted from the first 0.5 seconds of the signal as shown in figure 12, the RMS value of noise is also estimated:

$$n_{env}(t) = |hilbert(n(t))| \tag{20}$$

$$n_{rms} = \sqrt[2]{mean(n_{env}(t))} \tag{21}$$

$$\text{Now, if } n_{rms} > 0, \quad SNR = \begin{cases} (\frac{x_{rms}}{n_{rms}})^2, & x_{rms} < n_{rms} \\ (\frac{x_{rms} - n_{rms}}{n_{rms}})^2, & x_{rms} \geq n_{rms} \end{cases}$$
$$\text{else, if } n_{rms} \leq 0, \text{ SNR} = \infty \tag{22}$$

1.7.4 Pixel classification: target vs. noise

Consider a case where the actual sound source was places inside the test environment as shown in the Figure 9. For analyzing the effect of β on are under ROC curves, the decision on classifying a peak detected as target or noise was made based on the decision criteria illustrated below and explained with example.

Target peak:

While computing the performance metrics, only positive peaks (local maxima) in the SRP image are considered as targets. So, pixels in SRP image either equal to or greater than their immediate neighborhood pixels, (strictly greater than at least one neighboring pixel) were considered as targets. A pixel closest to the actual target position is considered as the peak, and along the line connecting the peak to the original target position, none of the pixel values fell 6dB below the peak magnitude. Also, the pixels that lie on the gradient leading up to a local peak were not considered. If the above conditions were satisfied, the target peak height and location estimate error was recorded. Else, no target detection was considered and magnitude was set to zero [32].

In the Figure 16, the intensity values considered from the SRP image, are positive (≥ 0) as indicated by the colormap shown next to the SRP image. The pixel that was selected as target location is marked with a green circle on the bottom right part of Figure 15.

For pixels marked as ‘Case 1’ in the image, though they are positive and closer to the actual source location, they are not considered as pixels corresponding to actual target peak because they lie on the slope of the gradient leading to the actual target peak. This ensures that perturbations along the gradient leading to a target peak are not considered.

However, for local maxima (peaks) marked as ‘Case 2’, though they are not on the gradient leading to the actual peak, they are not considered as candidate for target peak because of their distance from actual source location.

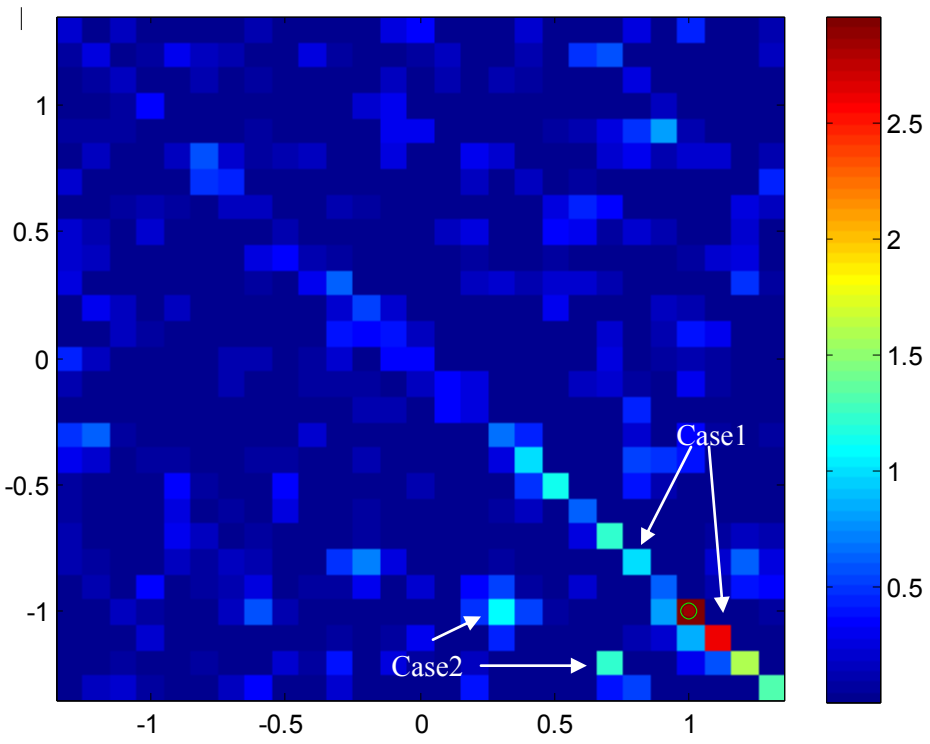


Figure 16: Example for decision logic for a target pixel

Noise peak:

A pixel in the immediate neighborhood of the detected target is not considered for noise peak. Also, pixels along the line connecting the detected target peak to the potential noise peak consisted of a negative value or were 6dB less than the target peak value. This ensured that variations along the gradients associated with the target peaks are not considered as noise peaks [32].

Figure 17 shows the SRP intensity distribution in the FOV. The range of power values represented is indicated in the colormap shown in the sidebar next to the image.

Pixels that lie in the immediate neighborhood of the detected target pixel are not considered as noise peaks (case1 in figure 17). For pixels that belong to case 2 (in figure 17), though they are not in the immediate target pixel neighborhood nor are on the gradient slope leading to a local maxima, their intensity level was not among the 8 highest peaks.

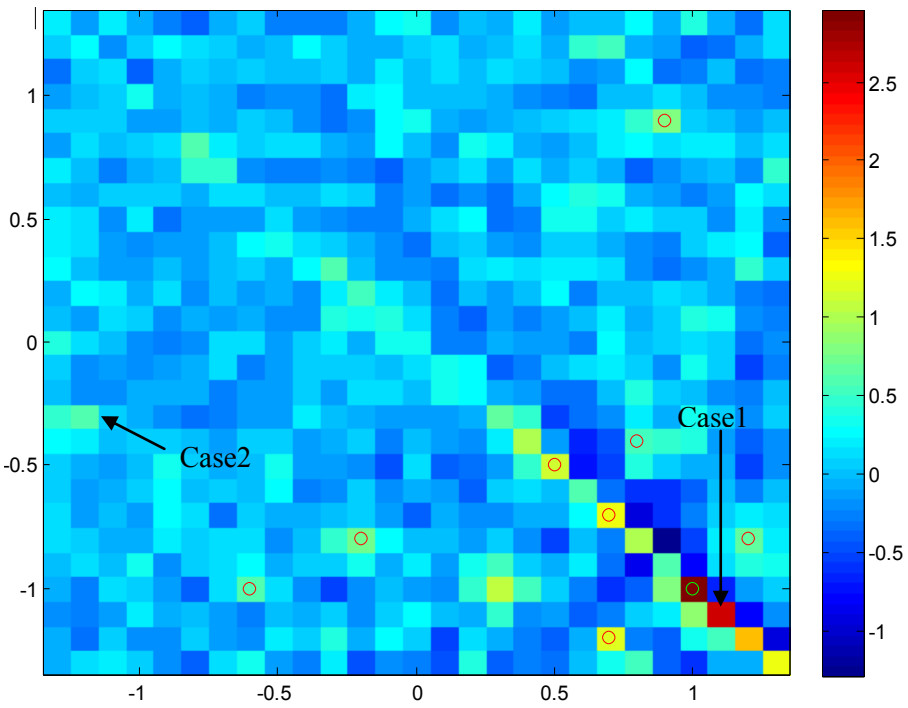


Figure 17: Example for decision logic for a noise pixel

1.7.5 Computing the ROC values

For all analysis in this thesis, the area under the ROC curve used to determine target detection performance. The ROC curve is a plot of probabilities of true (target peak) detection versus false-positive (noise peak) detection for all thresholds over the range of SRP values from the 2 classes (target & noise).

Given n_1 pixels from H_1, and n_0 pixels from H_0,The ROC area is estimated directly from the pixel amplitudes using the Wilcoxon statistic from [32]:

$$A_z = \frac{1}{n_0 n_1} \sum_{k=1}^{n_1} \sum_{l=1}^{n_0} C(S_{k|H_0},\, S_{i|H_1}) \tag{23a}$$

where, n_0 and n_1 are number of target and noise pixels & the value of:

$$C(S_{k,l|H_0}, S_{i,l|H_1}) = \begin{cases} 1 & for\ S_{k,l|H_0} < S_{i,l|H_1} \\ 0.5 & for\ S_{k,l|H_0} = S_{i,l|H_1} \\ 0 & for\ S_{k,l|H_0} > S_{i,l|H_1} \end{cases} \tag{23b}$$

To remove the dependency of A_z estimates calculated, the number of target and noise peaks considered were according to the ratio 1:8 (i.e. for every target detected, the 8 highest noise peaks in the FOV were considered for ROC analysis). This also doubles up as the worst case scenario as the 8 noise peaks selected will be the 8 highest peaks for that SRP image. Else, if all noise peaks were used, the low level noise peaks would result in very low false-positive ratio. This would in-turn cause higher A_z values, giving a false impression of a high ROC area.

To compute the 95% confidence limits for the ROC area for each case, the standard error statistic was calculated from the A_z estimate [36].

$$\sigma_{SE} \approx \sqrt{\frac{A_z(1-A_z) + (n_0-1)\left(Q_a - {A_z}^2\right) + (n_1-1)(Q_2 - {A_z}^2)}{n_0 n_1}} \tag{24a}$$

where,

$$Q_1 = \frac{A_z}{2-A_z} \text{ and } Q_2 = \frac{2{A_z}^2}{1+A_z} \tag{24b}$$

The results obtained and the discussions are explained in the following chapter.

CHAPTER 4

Results and Discussion

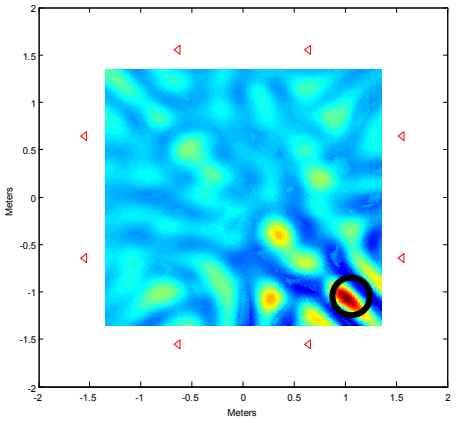
This chapter presents the experimental results and discusses the effect of β on a microphone array based SSL system performance for different test signals in the experimental setup discussed in Chapter 3. The results of β on SRP-PHAT images are presented in 4.1. The performance comparison between the area under ROC curve performance between the experiment and the simulations is presented in 4.2 along with similarities differences in ROC performance.

3.1 Results

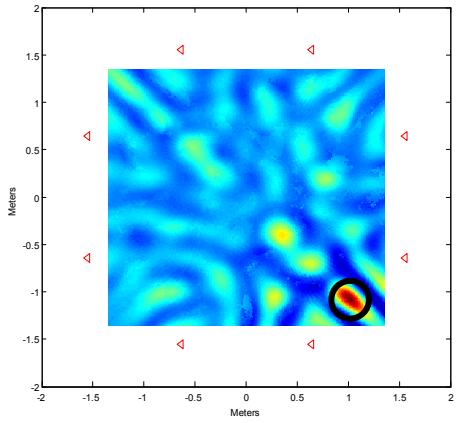
Figure 18 shows the SRP imaging results for a FOV containing a narrowband ((a), (b), (c)) and broadband signal source ((d), (e), (f)). The actual source location is at the center of black circle in the Figures. The microphone positions are indicated by small red triangles '◅' in the images. Each image shows the relative strengths of the target and noise peaks for β = 0, 0.6, and 1. The results presented in Figure 18 are for low room reverberation levels.

Consider the narrowband signal case (Figure 18 (a), (b), (c)), strong noise peaks are observed at non-target positions (due to partial coherences) at β = 0. As β increases to 0.6, there is significant reduction in noise peak amplitude in non-target locations as the partial coherence is reduced and the dominant noise peaks loose strength. At the same time, there is also an increase in the density of low level, fine-grained, noise peaks as β approaches 1. This confirms the results from simulation results in [32] that targets having a narrow signal spectrum degrade from the PHAT more than the broadband signals, due to enhancement of relative spectral components outside the narrowband signal range which contributes to noise peaks in SRP image and corrupts the target peak.

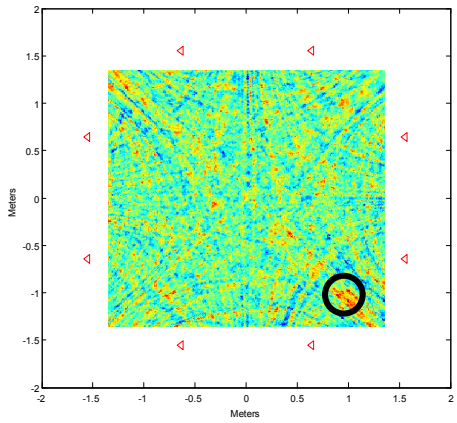
Narrowband



(a) β=0

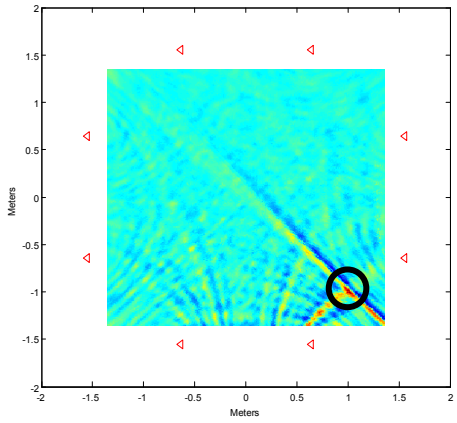


(b) β=0.6

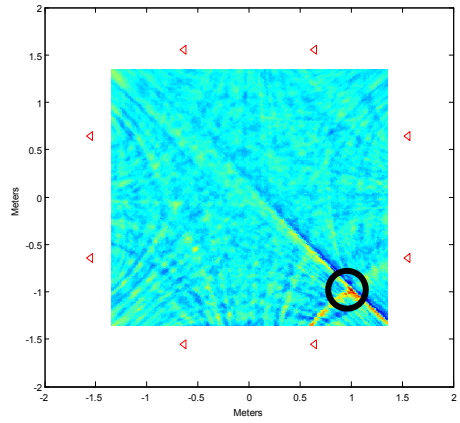


(c) β=1

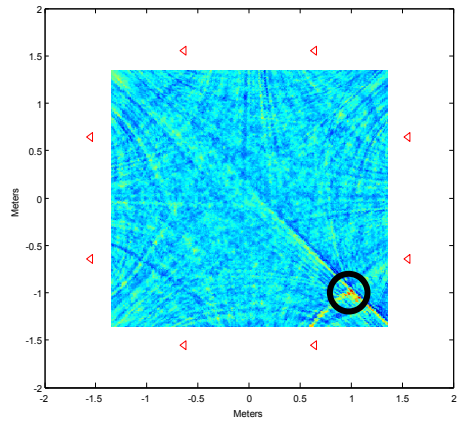
Broadband



(d) β=0



(e) β=0.6



(f) β=1

Figure 18: SRP images for narrowband and broadband signals for β = 0, 0.6 & 1

The influence of PHAT on the broadband target (Figure 18 (d), (e), (f)) is similar to the narrowband case for values of β upto 0.6 in terms of the influence on noise peak reduction. However, for β =1, the target peak strength appears to improve relative to increase in the noise peaks, whereas the narrowband source type shows performance degradation due to increase in intensity and number of noise peaks in SRP. The broadband signal exhibits this property primarily because the coherent target energy is distributed over most of the spectrum and the signal of interest gains from the PHAT. Hence, improvement in the noise performance for the low amplitude spectral regions also increases the signal power.

The effect of variation in β on SRP-PHAT is explained in detail in the following section with respect to target detection performance in noisy and reverberant conditions using the ROC curves.

3.2 Discussion of target detection performance

3.2.1 Analysis method

For analyzing the performance of PHAT- β in terms of target detection, results were assessed using area under the Receiver Operating Characteristics (ROC) curve [36-38] on acquired data. The computation of area under ROC curve and the 95% confidence limits is explained in chapter 3 (section 3.3.3).

An area under ROC (A_z) of 0.8 represents 80% probability that the target peak value will exceed any independent noise peak value selected. The curves obtained are analyzed and a subjective comparison of actual experimental results is done with respect to those of simulated data published in [32], to study the similarities and disparities in performance.

This following section presents a comparison of area under ROC vs. β plots for different signal types and operating conditions (reverberation and SNR). The relationship between β and its effect on ROC performance is discussed between the experiments conducted and those from simulations in [32].

3.2.2 Constant low reverberation (foam only) & different signal SNR

The Figures 19, 21, 23 & 25 show the variation in area under ROC curves for narrowband and broadband targets used in actual experiment under low room reverberation. The acoustic foam used on the walls absorbs most of the multipath signals and noise.

The range of β values resulting in improvement in performance is shown for different cases. The different SNR levels used in each ROC performance comparison is indicated in the legend.

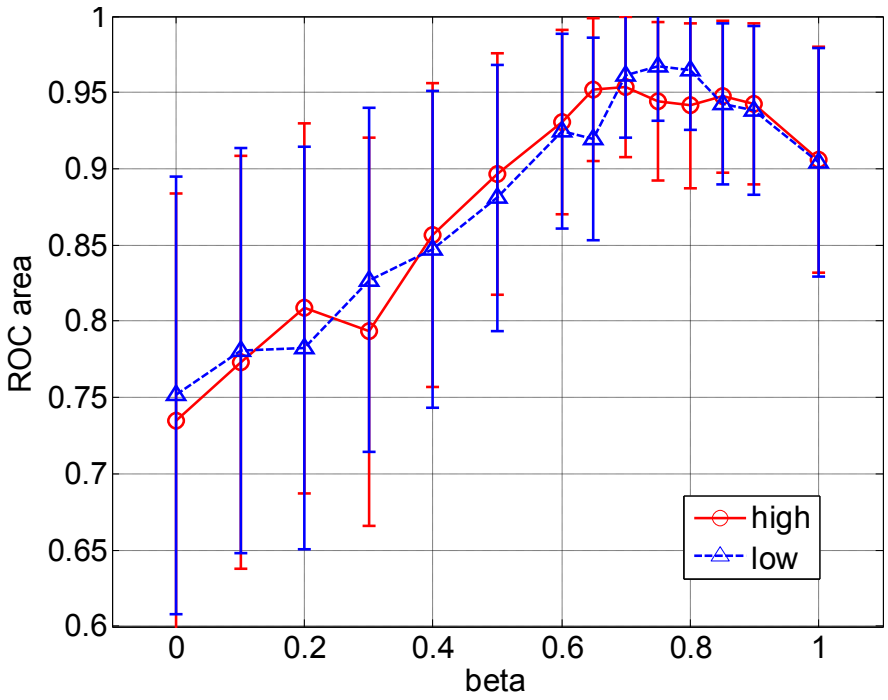


Figure 19: Broadband Colored noise : different SNR
Experiment under low room reverberation (foam)

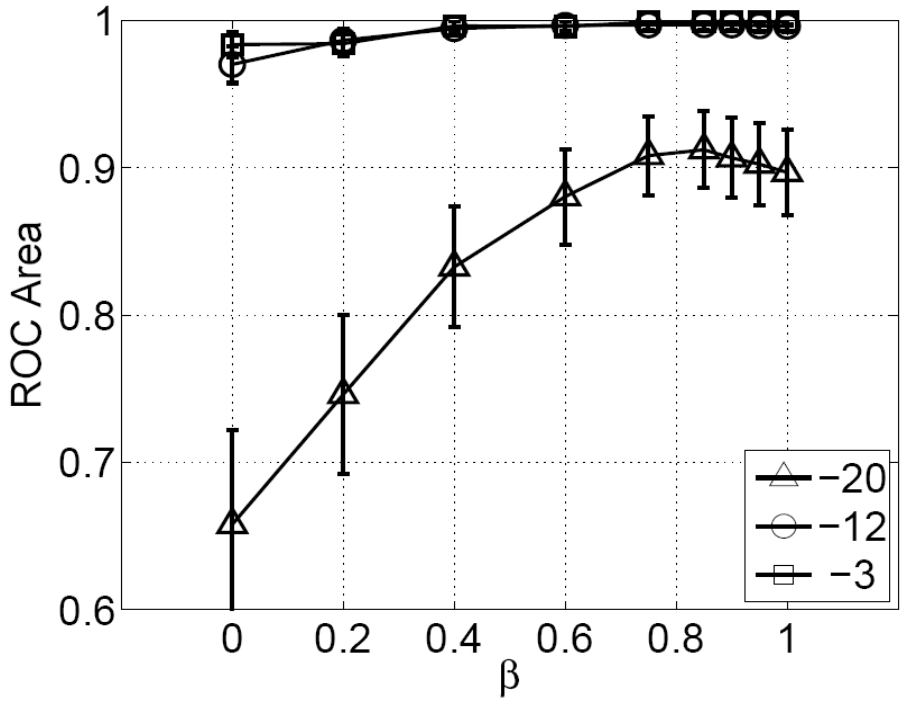


Figure 20: Broadband signal: different SNR
Figure adapted from [32] for simulation with room reflectivity set to 0.

Figures 19 shows the A_z estimate for a broadband colored noise signal for highest and lowest SNR signals when room reverberation is fixed (low when foam is used). Comparing it to results from simulations (Figure 20), where room reflectivity was set at 0:

The trend in the ROC curves is similar for experiment and simulation for all values of β, i.e., there is improvement in A_z value as β increases from 0 to 0.8. Beyond this, there is a small drop in performance as β increases closer to 1. But the positive influence of β (around 0.6-0.8) in improving detection performance is evident.

For a broadband signal with a wider spectrum, the loss in ROC values as β increases beyond 0.8 is not very dramatic because the increase in noise peak values with β is also accompanied by an increase in the target peak compensating the loss in ROC to an extent.

However, The expected variation in A_z performance between high and low SNR signals as in simulation results of Figure 20 is not present in Figure 19 because, for all simulated results, the room reverberation could be separated from the direct path signal for analysis. But for the experimental conditions, this is not possible and though acoustic foam was used, some reverberations still exist inside the FOV, especially for low frequency signals corresponding to the thickness of the acoustic foam used (1.125 inches).

From the ROC analysis, values for β suggested for use under similar operating conditions: 0.65 to 0.85

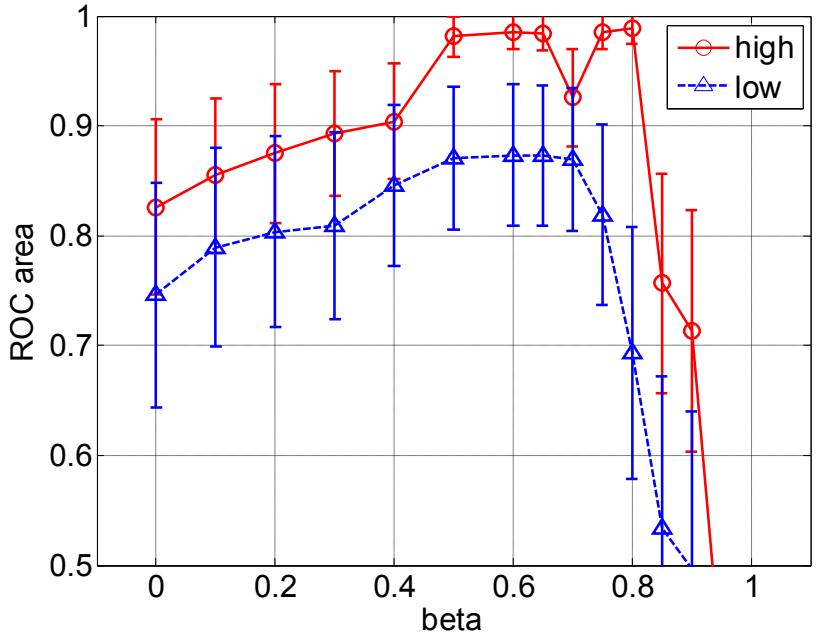


Figure 21: Narrowband Colored noise : different SNR
Experiment under low room reverberation (foam)

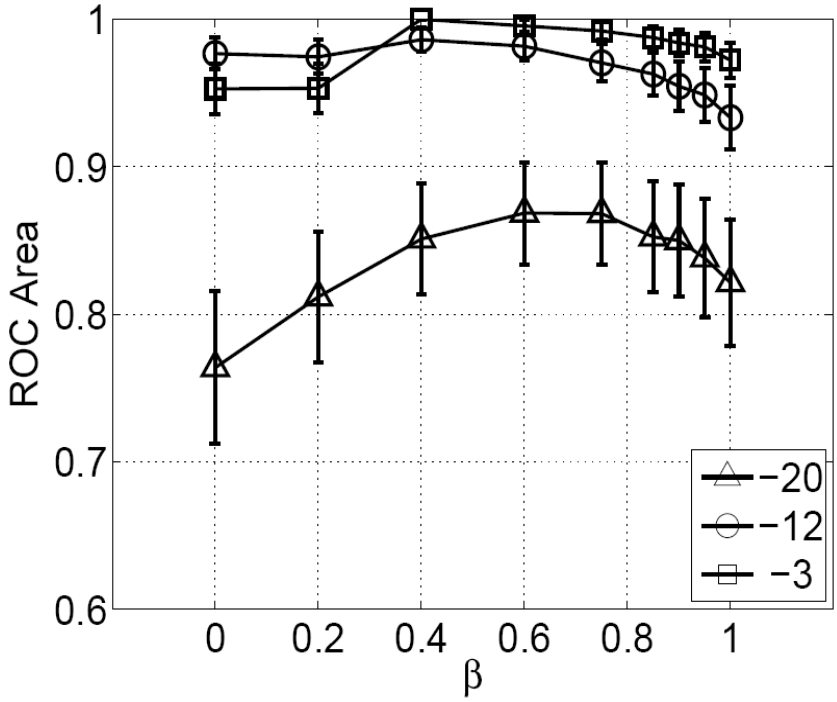


Figure 22: Narrowband signal : different SNR
Figure adapted from [32] for simulation with room reflectivity set to 0.

Figures 21, 22 show the effect of β area under ROC curve for a narrowband colored noise signal under room conditions discussed above.

The difference in ROC performance for high and low SNR signals in Figure 21 and the trend in variation of A_z with β agrees with those from Figure 22.

For higher β values the ROC area starts dropping for the narrowband signals because the signal content in the spectrum is lesser compared to broadband signals and the effect of spectral whitening starts emphasizing spectral components in the higher frequency range (noise).

The difference in performance is for $0.6 \leq \beta \leq 1$, where there is a more dramatic drop in A_z values for the experimental results than those of simulations. This can be explained by the fact that for narrowband signals, the number and intensity of noise peaks inside FOV increases dramatically as β approaches 1 (refer to Figure 18 (a), (b), (c)). In the results from simulation, this performance degradation for higher β is not dramatic because the narrowband signal had coherent energy extending to the Nyquist frequency (8 kHz). But in actual data, the coherence of the signal is lost due to the 16-bit quantization in the hardware and the noise floor of the amplifier. So, for β=1, the incoherent noise levels increase affecting the ROC performance for the target with a narrow spectral range.

Suggested values of β under similar conditions: 0.5 to 0.65

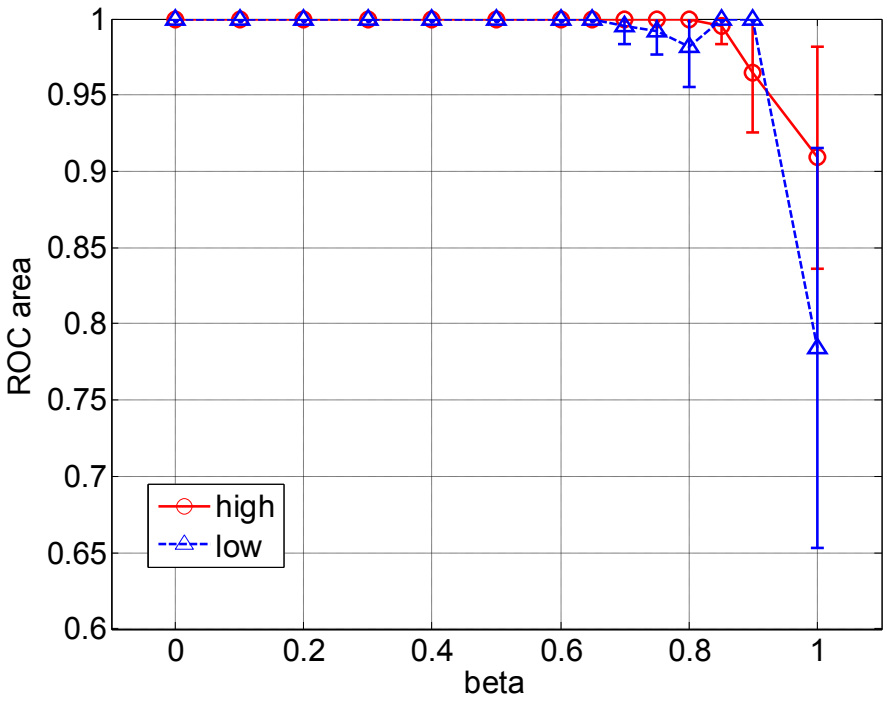


Figure 23: Narrowband impulse: different SNR
Experiment under low room reverberation (foam)

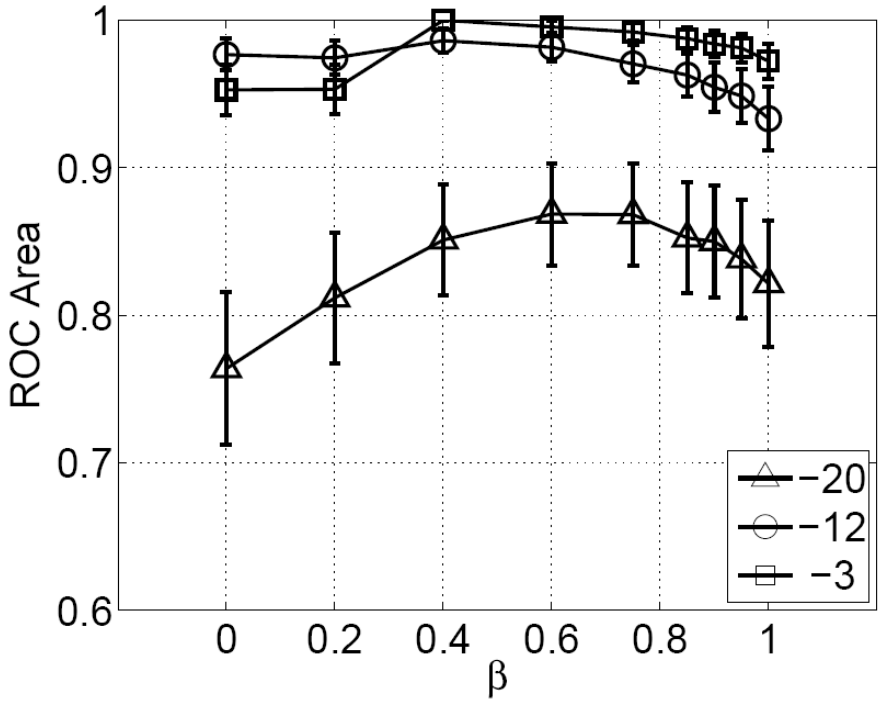


Figure 24: Narrowband impulse: different SNR
Figure adapted from [32] for simulation with room reflectivity set to 0.

From the ROC curve comparison between experimental results in Figure 23 with those of simulations in Figure 24 for a narrowband impulse signal, the simulation results predict a drop in ROC performance towards higher values of β. This trend is present in the experimental results too but the drop is much higher than those of simulations. This is again due to reason for a similar observation in results of figure 21 for a narrowband colored noise signal.

The increase in ROC performance for $0 \leq \beta \leq 0.6$ expected in the simulation results is not seen in the experimental results because experimental SNR was too high.

3.2.3 Constant high reverberation (plexi only) & different signal SNR

The experimental results in Figures 25, 27, 29 and 31, show the A_z estimates for narrowband and broadband targets in reverberant room conditions. The plexi has a high reflection coefficient and results in a highly reverberant condition inside the test environment. The signal SNR is due to the coherent noise and reverberation also adds to the partial coherences.

The range of β values resulting in improvement in performance is shown for different cases. The legends in the plots indicate the different signal SNR levels.

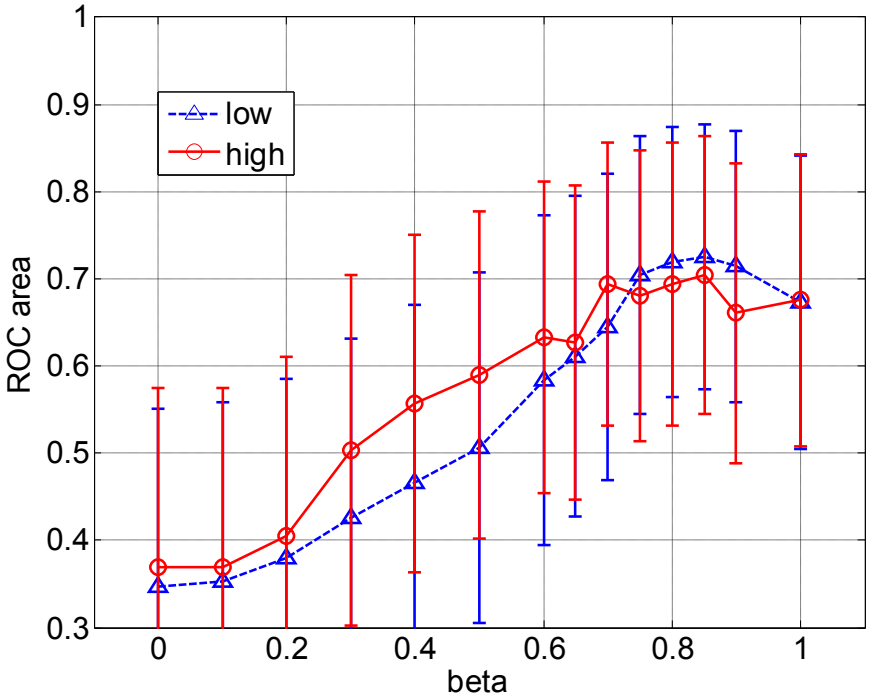


Figure 25: Broadband Colored noise : different SNR
Experiment under high room reverberation (plexi)

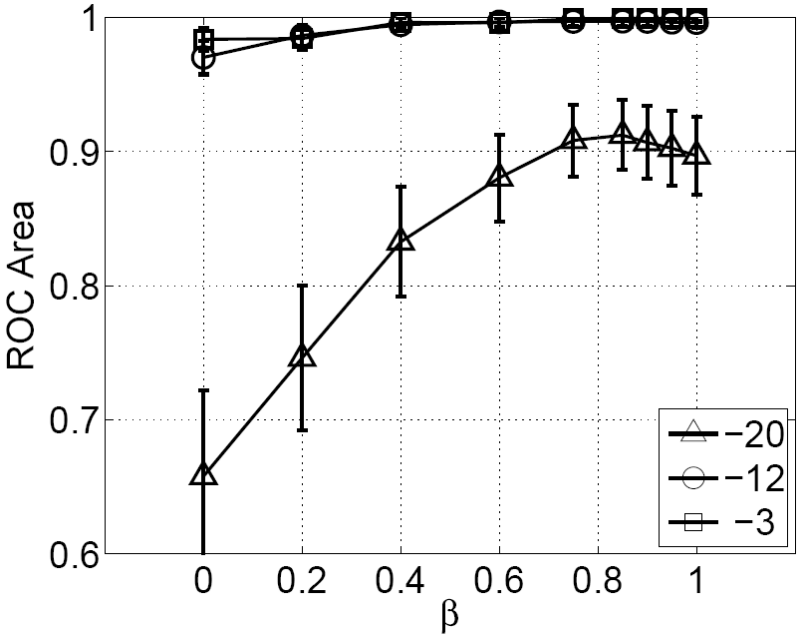


Figure 26: Broadband signal : different SNR
Figure adapted from [32] for simulation with room reflectivity set to 0.

Figure 25 shows the effect of change in β on area under ROC curve for a broadband Colored noise signal for a highly reverberant room environment.

Compared to the simulated results in Figure 26, the effect of β on ROC performance is similar in the Figure 25 with respect to the improvement in ROC values as β approaches 0.8. Beyond this point, there is a slight drop in ROC value due to increase in noise peak values as a result of the whitening effect of PHAT.

The plot in Figure 26 presents results from a simulated room with no reverberation. So, comparing the ROC curves of Figure 26 with the experimental results in Figure 25, it can be observed that there is an overall reduction in ROC values when there is high level of room reverberation.

Suggested values of β under similar conditions: 0.65 to 0.85

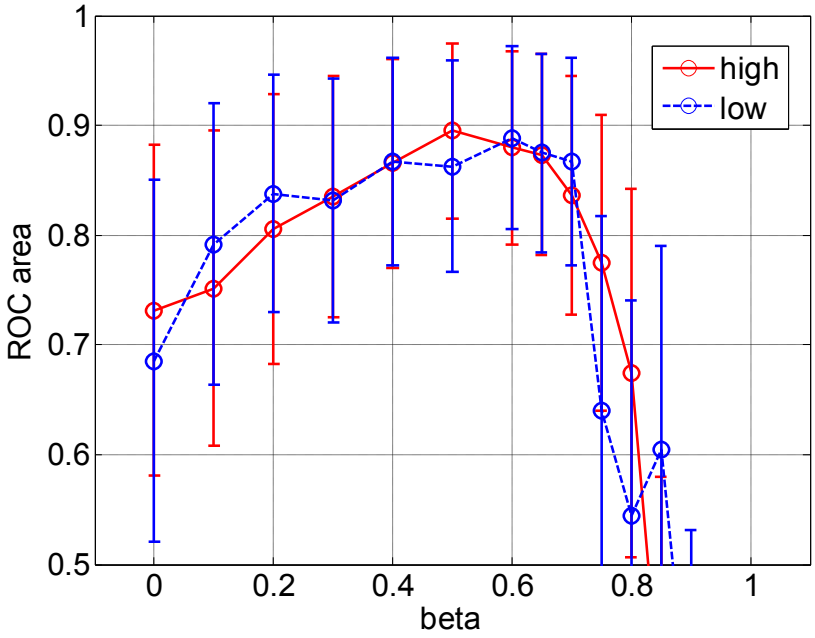


Figure 27: Narrowband Colored noise : different SNR
Experiment under low room reverberation (plexi)

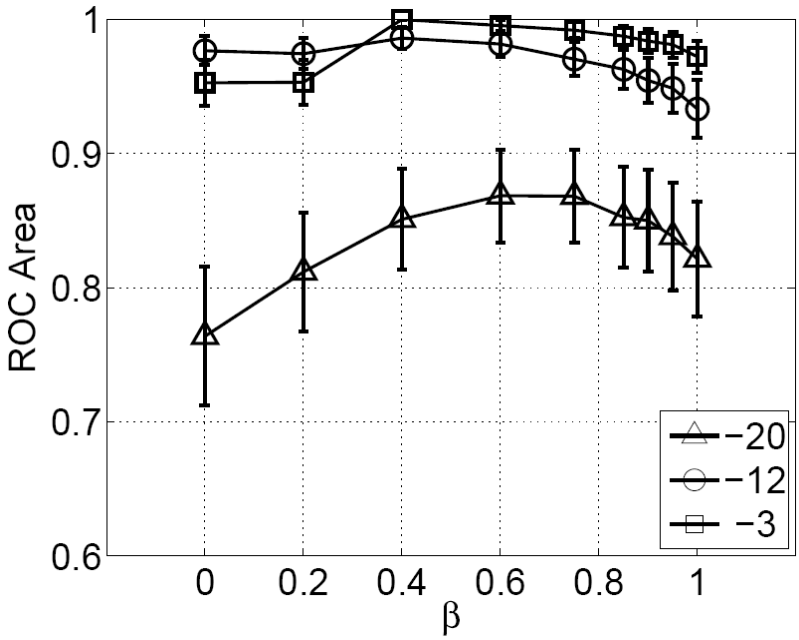


Figure 28: Narrowband signal : different SNR
Figure adapted from [32] for simulation with room reflectivity set to 0.

A_z values in Figure 27 show the effect of β for a narrowband colored noise signal under highly reverberant room conditions.

There is a strong agreement in the general trend in ROC performance variation with β upto 0.6. For higher β values the ROC area starts dropping for the narrowband signals because the signal content in the spectrum is lesser compared to broadband signals and is due to the effect of spectral whitening (PHAT) that starts emphasizing the higher frequency components (noise).

The difference in performance however is for $0.6 \leq \beta \leq 1$, where there is a very dramatic drop in A_z values for the experimental results than those of simulations. This can be explained by the fact that for narrowband signals, the number and intensity of noise peaks inside FOV increases dramatically as β approaches 1 (refer to Figure 18 (a), (b), (c)). This in turn affects the ROC performance as observed previously from results of narrowband signals in low reverberation environment in figures 21, 23.

Similar to the Figure 25, there is a significant drop in the experimental ROC performance because the room reverberation levels are higher when plexi glass is used.

Suggested values of β under similar conditions: 0.5 to 0.65

3.2.4 Constant signal SNR (lowest) & different reverberation levels

Experimental results for ROC area for narrowband and broadband targets with high and low levels of room reverberation are shown in Figures 29, 31. The range of β values resulting in improvement in performance is shown for different cases. The reverberation levels for each ROC curve is indicated in the legend. Also, these plots were generated for source signals with lowest SNR levels.

The reverberation levels in the experiment and simulated results are indicated in the legend.

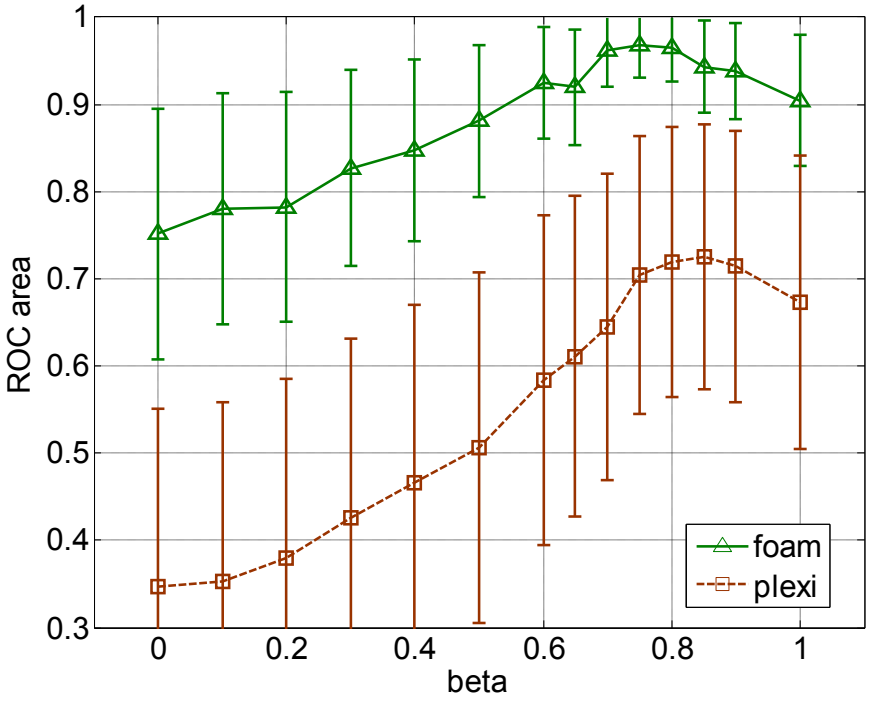


Figure 29: Broadband colored noise : different reverberation
Experiment under fixed SNR

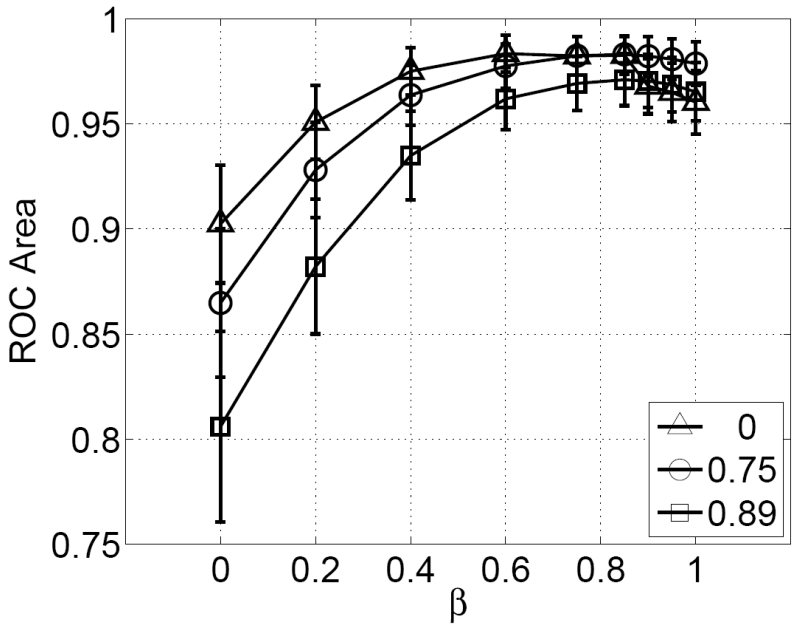


Figure 30: Broadband signal : different reverberation
Figure adapted from [32] for simulation of low SNR source.

Figures 29 shows the A_z estimate for a broadband colored noise signal for high (plexi) and low (foam) reverberation levels for fixed SNR fixed. Figure 30 shows ROC performance variation for similar conditions but from simulations in [32].

The trend in the ROC curves is similar in both reverberation conditions for all values of β, i.e., there is improvement in A_z value as β increases from 0 to 0.8. Beyond this, there is a small drop in performance as β increases closer to 1. But the positive influence of β (around 0.6-0.8) in improving detection performance is evident.

Also, there is a clear difference in the ROC values between foam (low reverberation) & plexi (high reverberation) as expected. The use of plexi glass increases reverberations inside the FOV, which has a more detrimental impact on target detection than room noise because reverberant noise is correlated with the target signal of interest. Hence, the ROC improvement due to PHAT β in a high reverberation case is more beneficial especially in normal talking scenarios.

The initial values of A_z for the plexi (high reverberation) case is less than 0.5 for β upto 0.6. This could be due to the increased levels of correlated noise under high reverberation which affects ROC performance. But as β ≥0.6, the A_z performance improves and this is a strong indicator for the effectiveness of PHAT-β in target improving detection.

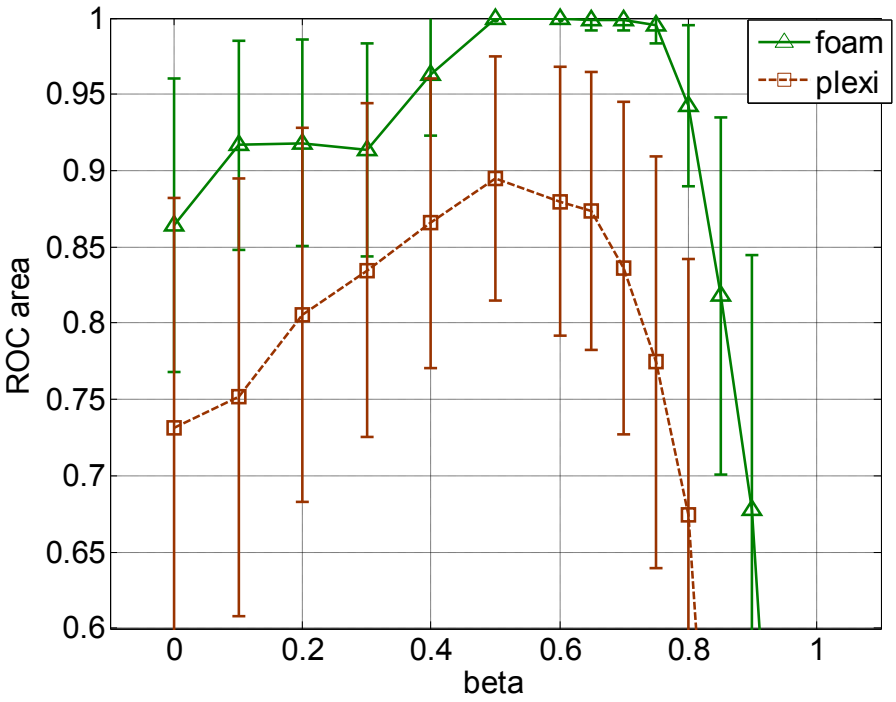


Figure 31: Narrowband colored noise: different reverberation
Experiment under fixed SNR

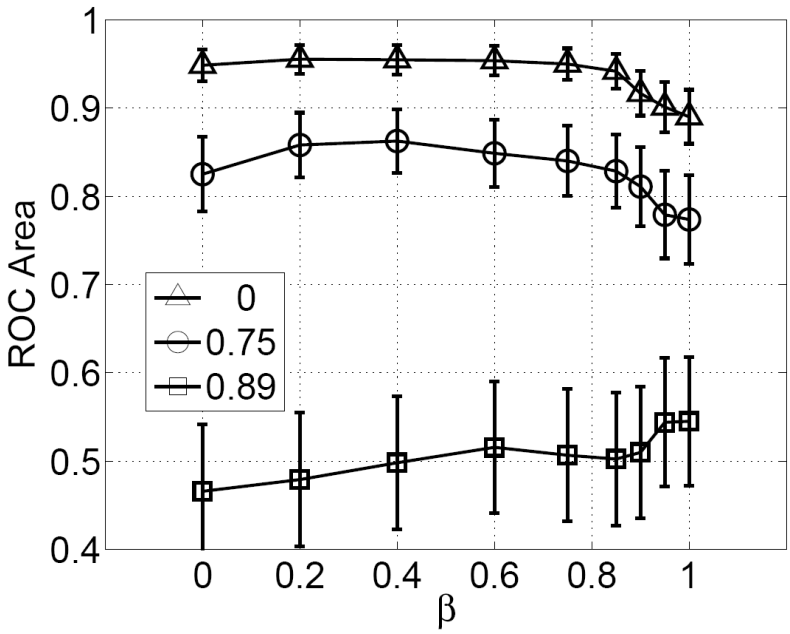


Figure 32: Narrowband signal : different reverberation
Figure adapted from [32] for simulation of low SNR source.

ROC area values in Figure 31 show the effect of β for a narrowband colored noise signal under different levels of room reverberation levels from the experiment.

There is a strong agreement in the general trend in ROC performance variation with β for both high and low room reverberation conditions. For higher β values the ROC area starts dropping for the narrowband signals which is the trend observed in the experimental results presented before (Figures 21, 23, 25).

When compared to the ROC curves in Figure 32, the performance results in Figure 31 (experiment) indicate a more dramatic influence of PHAT-β on performance improvement both in low and highly reverberant conditions, particularly for higher reverberation condition. This is a strong factor in support of not choosing the conventional PHAT (β = 1).

To summarize the above analysis, it is clear that partial PHAT weighting with β improves ROC performance low reverberation conditions (foam only) and more importantly in high reverberation conditions (plexi only) by almost 20%.

- This gain in ROC performance is evident for β in range of 0 to 0.6 for both narrowband and broadband source types.
- For β values greater than 0.6:
    - The ROC area starts dropping drastically for the narrowband. This is because the signal bandwidth is small and whitening of the spectrum for $\beta > 0.6$ starts emphasizing spectral components in the higher frequency range. Hence the ROC values fall more drastically when compared to the results from simulations of [32].
    - On the other hand, the broadband signal has a wider spectrum and the gain in noise peaks happens along with gain in target peaks. So, ROC performance with β improves significantly till $\beta = 0.8$ beyond which there is a slight roll of in ROC as β approaches 1 due to increase in high frequency (noise) peak values by PHAT which causes a small drop in ROC area values.

The analysis of the results clearly demonstrates that reverberation has a more detrimental impact on target detection than uncorrelated room noise because reverberant noise is correlated with the target. This increases the variance at the actual target position, and also at non–target positions by increasing the overall noise power through additional energy from multi–path signals. Hence, the ROC improvement due to PHAT β in a high reverberation case is more beneficial especially in normal talking scenarios.

Also, the ROC curves for the experimental cases were a little different when compared to those for simulations in [32] because:

- Overall effect of using acoustic foam on the walls is not equivalent to the simulated room with "0" reverberation because the foam does not effectively block all reverberations.

- The microphone placement in the simulation was exactly at the required locations whereas in actual experimental situation these measurements were not precise. Even a small error (few centimeters) in microphone placement would cause location error for sources with high frequency content.
- The sources used in the experiments were directional while the results presented in [32] are based on simulation of an omni-directional source. To offset the effect of the source directionality on experimental analysis, the experiments were conducted with sound sources facing 2 opposite orientations at each test position in the FOV during data collection.

CHAPTER 5

Conclusions and Future Work

The chapter focuses on the contributions made and the inferences derived based on the results of the experimental evaluation of SRP-PHAT-β on signals of varying characteristics under different operating situations. The outcomes of the experiments and a summary of the performance of SRP-PHAT-β are discussed in Section 5.1. Section 5.2 describes the possible future research directions that could be followed in developing a comprehensive system for Speaker source detection and localization.

5.1 Summary

The thesis used an ROC area analysis for assessing the detection performance of SSL processes for real experimental data. Comparisons between ROC performance for real data and those of simulated conditions highlight performance sensitivities in the ROC area statistics. The analysis reemphasized the performance gains offered by β in SRP-PHAT and at the same time, suggesting the need for a proper choice of β based on operating conditions to achieve optimal performance.

It has been shown that is the use of conventional PHAT (β = 1) improves SRP performance compared to not using PHAT at all (β = 0) [27, 34, 40] and also that PHAT is an optimal weighting approach for SRP under noisy and reverberant conditions [42]. But, the analysis results from Chapter 4 on narrow and broadband signal sources showed a consistent loss in detection performance when using the conventional PHAT (β = 1) under noisy and reverberant conditions. The losses were most significant under high reverberation and for narrowband signal types. A significant performance improvement was consistently seen for broadband targets especially in reverberant conditions.

The PHAT-β allowed for a parametric variation on the amount of influence given to the original spectral amplitude on the final coherent power values. It is a useful parameter to vary the impact of the PHAT based on operating conditions and a way to make the performance of the PHAT more robust over a range of narrow and broadband target sources under different operating conditions.

The values of β suggested for use under different conditions based on analysis in Chapter 4 are summarized in Table 4 below.

Table 5: Suggested β values

	High SNR	**Low SNR**	**High reverberation**	**Low reverberation**
Narrowband	0.6	0.55-0.6	0.55-0.6	0.6
Broadband	0.65-0.8	0.65-0.8	0.65-0.75	0.7

Based on experimental results listed in the table above, for a source signal which is complex in nature like human speech (varying combination of narrowband and broadband components), β values between 0.55-0.65 could be used for robust performance enhancement under different application environments.

5.2 Future work

Since the overall objective of this thesis was to evaluate and emphasize the effectiveness of partial weighting factor: β on SRP-PHAT, the experimental setup and test conditions used were similar to those in the simulations described in [32]. For a more comprehensive performance evaluation, different experimental setups could be investigated, such as:

- changes in the number of microphones in the array
- microphone spacing (logarithmic spacing, etc.)
- other microphone geometries (planar array , 3D array)
- even the low SNR signals used for the experiment were strong enough which resulted in small changes in ROC performance with SNR variation. Even lower values of signal SNR could be tested for a better comparison of the effect of SNR on target detection performance.

APPENDICES

Appendix A: Acoustic signal modeling

This appendix gives an overview of the theory and concepts on acoustic signal propagation and the different factors to be considered for implementation of sound source location techniques.

1 Sound Propagation

Sound waves propagate through an air medium by the movement of molecules along the direction of propagation. These are referred to as compressional waves. The wave equation for acoustic waves propagating in a homogeneous and lossless medium is given by:

$$\nabla^2 s(t,r) = \frac{1}{c^2}\frac{\partial^2}{\partial t^2} s(t,r) = 0 \tag{25}$$

where, *s(t, r)* represents the sound pressure at a time instant '*t*' for a point in space with Cartesian coordinates $r = [x, y, z]^T$. Here, ∇^2 is the Laplacian operator and '*T*' is the transpose operator. The variable '*c*' is the speed of sound, which depends on the pressure and density of the medium, and is constant for a given wave type and medium.

In general, for a far-field sound source, waves propagate as spherical waves, with the amplitude decaying at a rate proportional to the distance from the source [44]. These properties result in complex mathematical analysis of propagating signals, which is a major issue in array processing for a near-field source. However, at a sufficiently long distance from the source, acoustic waves may be considered as plane waves, considerably simplifying the analysis.

The solution to the wave equation for a plane wave is:

$$s(t,r) = Ae^{j(\Omega t - k^T r)} \tag{26}$$

where, A is the wave amplitude, the angular frequency is $\Omega = 2\pi F$ (F is the real frequency in Hertz), and k is the wave number vector, which is a function of the speed and direction of the wave propagation.

2 Acoustic Noise Field

Noise field (or *background noise field*) is generally considered as the acoustic field in the absence of information transmission. It constitutes unwanted or disturbing acoustic waves introduced by man-made and natural sources. Hence, depending on the correlation between noise signals at distinct spatial locations, the following common categories of noise fields can exist for microphone-array applications that affect their performance [45].

Coherent versus Incoherent Noise Field

A coherent noise field corresponds to noise signals propagating from their source without undergoing reflection, dispersion or dissipation. These are characterized by high-correlation with the direct path signals. In general, a source in open air environment without obstacles to sound propagation causes coherent noise field.

On the other hand, an incoherent noise field is characterized by uncorrelated noise signals. An example of incoherent noise is electrical noise in microphones, which is considered to follow a random distribution.

Diffuse Noise Field

Noise signals propagating in all directions simultaneously, with almost equal energy and low spatial correlation make up a diffuse noise field. A perfectly diffuse sound field is typically generated by distant, uncorrelated sources of random noise over all directions. Many noise environments, such as car cabin, office environment, etc, to a certain extent, can be characterized by a diffuse noise field.

Background Noise

In urban environment, noise is omnipresent. Most background noise is generated by traffic movement, air circulation systems in public places. High levels of background noise reduce the intelligibility in perception of a sound source such as that from a human speaker.

Background noise generally degrades the performance of SSL systems depending on the SNR.

The above mentioned acoustic disturbances are considered to arrive from all directions, and so they are generally characterized as surrounding noise sources. Although different noise characteristics can be attributed to different types of sources, background noise has higher levels of low-frequency content. Also, background noise commonly displays a nearly Gaussian distribution. By exploiting the *a priori* knowledge about the spectral content of the source signal, if available, noise suppression techniques can be implemented to target a specific characteristic for improvement in SSL performance.

3 Reverberation

Reverberation is common in acoustic signals that propagate in lightly damped enclosures. In closed environments, the source signal is reflected by walls, floors, ceilings and objects inside the room. These multiple reflections are added to the direct-path speech signal component (after some attenuation and phase shift). Hence, the signal impinging on the microphones is affected by reverberation. In most cases, room reverberation is characterized by the reverberation time, *RT*60, which is the time required for reverberation energy to decay by 60dB. It is dependent on the room size and furnishings, as well as the reflection coefficients of the constituents of the room [46].

The position of the source and the acoustic sensors in the room and their relative distances also define the strength of reverberation. The intelligibility of captured signal is considerably reduced in a highly reverberant environment. Additionally, performances of conventional noise reduction algorithms are greatly affected in reverberant conditions. De-reverberation can be performed by identifying the reverberant channel effects and compensating for them [47]. Reverberation effects are given equal weight age for evaluating SRP-PHAT-β algorithm to that of acoustic noise conditions.

4 Localized interference

In public places, the desired source signal may be corrupted with noise from neighboring sources, also referred to as *"cocktail party effect"* [48]. This effect is mostly encountered in situations involving human speech as signal source. It can be countered by making use of a multi-microphone-based system, where the spatial separation between the

desired source and the acoustic interference sources is exploited to improve SSL performance [6, 23, 35, 36].

5 Acoustic coupling effects

The far-end signal emitted by the source propagates in the environment and is captured by the microphones in the same way as other interfering signals. Hence, acoustic feedback constitutes another source of disturbance in the case of a source-loudspeaker setup. This 'noise' situation is easily handled using a reference signal at the loudspeaker for the tuning of echo suppression systems. Also, since the relative positioning of the microphones and loudspeakers in most situations is known, it helps in maximal utilization of spatial filtering techniques [26].

6 Acoustic Room Modeling

Acoustic room modeling is commonly used to simulate the propagation of source signals in a typical room. This is accomplished by convolving source signal sequence with simulated room impulse responses for specific room characteristics and positions of the speaker and microphone.

The modeling of sound fields in reverberant rooms can be performed by solving the wave equation for boundary conditions defined by the enclosure limits. This approach results in a description of the spatial distribution of acoustic field in the room by identifying the enclosure's spatial and spectral modes. However, it does not provide a direct reconstruction of an impulse response [49]. A number of geometric approaches have been developed for modeling the sound propagation in a room such as image methods, ray tracing, and beam tracing. The principle of the image method first introduced in [50] has been predominantly used, since all reflections up to a given order or reverberation time are modeled easily. However, its computational complexity grows exponentially.

7 Acoustic Array Properties

Acoustic sensor arrays consist of a set of acoustic sensors placed at different locations in order to receive a signal carried by propagating waves. Sensor arrays are commonly considered as spatially sampled versions of continuous sensors, also referred to as *apertures*. From this perspective, sensor array fundamentals can conveniently be derived from continuous aperture principles by means of the sampling theory.

Continuous Aperture

A continuous aperture is an extended finite area over which signal energy is gathered. The two important terminologies used in the study of continuous aperture are the *aperture function* and the *directivity pattern*.

The aperture function*:* defines the response of a spatial position along the aperture to a propagating wave. The aperture function, takes values between zero and one inside the region where the sensor integrates the field and is null outside the aperture area [45].

The directivity pattern*:* also known as *beam pattern* or *aperture smoothing function,* corresponds to the aperture response as a function of frequency and direction of arrival. The directivity pattern corresponding to a uniform aperture function is illustrated in Figure 42.

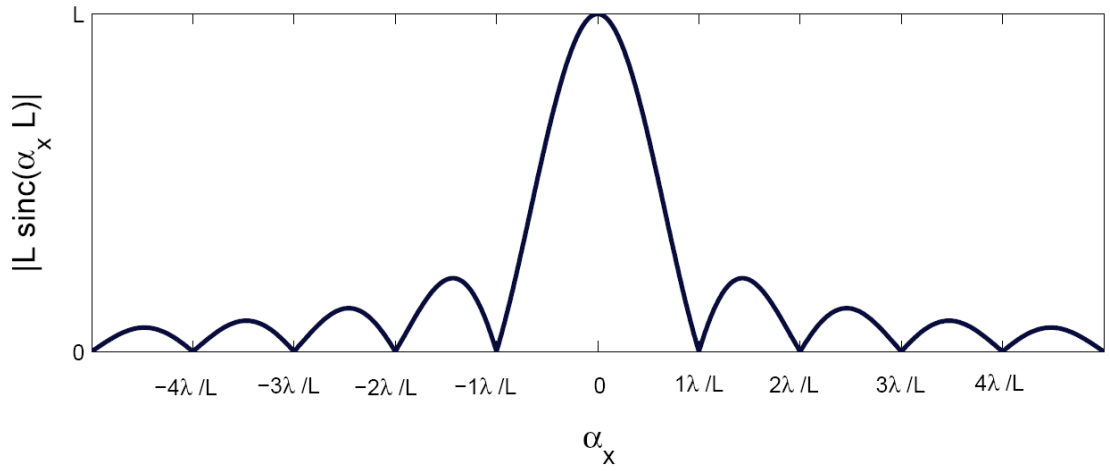


Figure 33: Directivity pattern of a linear aperture
Figure adapted from: Yermeche, Z., Soft-Constrained Sub band beamforming for Speech Enhancement

From the Figure 33, the zeros in the directivity pattern are located at $\alpha_x = i(\lambda/L)$, where i is an integer and the beam width of main lobe is $2\lambda/L = 2c/(F_L)$. Thus, when the aperture length is constant, the main lobe is wider for lower frequencies and vice versa.

With respect to the horizontal directivity pattern (i.e., $\varphi = \pi/2$), whose polar plot is shown in Figure 34, It can be seen that higher the frequency, i.e., L/λ higher (right), narrower is the main beam width.

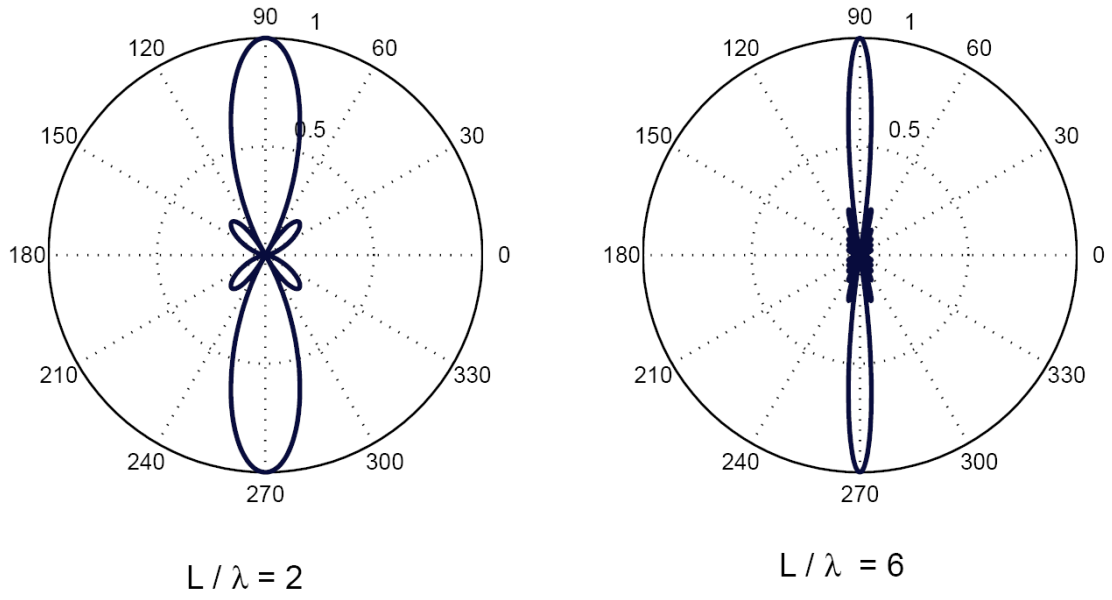


Figure 34: Polar plot of the directivity pattern of a linear aperture
as a function of the horizontal direction θ, with L/λ = 2 (left) and L/λ = 6 (right).
Figure adapted from: Yermeche, Z., Soft-Constrained Sub band beamforming for Speech Enhancement

Linear Sensor Array

A sensor array can be viewed as a continuous aperture excited at a certain finite number of points. In the case of equally weighted sensors, increasing the number of sensors results in reduced energy in the side lobes. On the other hand, for a fixed number of sensors, the beam width of the main lobe is inversely proportional to the sensor spacing *d*. Taking all the above factors into consideration, a proper selection of sensor array parameters helps avoid the effect of '*Spatial Aliasing*' described in the following section.

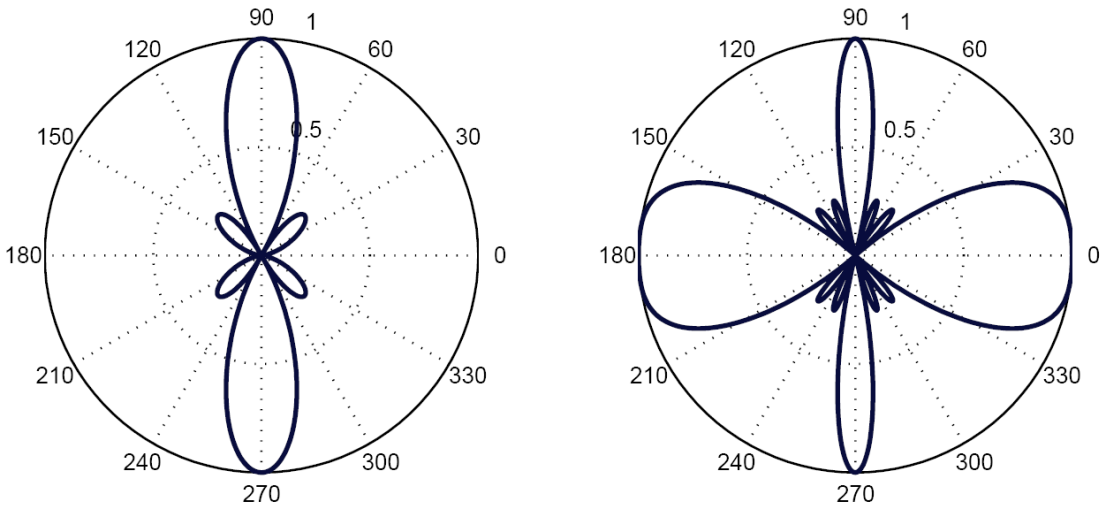
Spatial Aliasing

Similar to the concept of temporal sampling in any continuous-time signal, spatial sampling can result in aliasing [45]. Spatial aliasing appears as spurious lobes in the

directivity pattern, as illustrated in figure 35. The requirement to fulfill the spatial sampling theorem, so as to avoid spatial aliasing, is given by:

$$d < \frac{\lambda_{min}}{2} \tag{27}$$

where, λ_{min} is the minimum wavelength in the propagating signal. For example, the critical spacing distance required for processing signals within the human speech bandwidth (300Hz to 5.4kHz) is approx d ≈ 5 cm.



d = λ/2 d = λ

Figure 35: Polar plot of the directivity pattern of a linear sensor array with four elements, as a function of the horizontal direction θ; with a critical spatial sampling, d = λ/2 (left) and aliasing effects for d = λ (right).
Figure adapted from: Yermeche, Z., Soft-Constrained Sub band beamforming for Speech Enhancement

Appendix B: Review of different SSL techniques

This appendix gives a brief introduction into the existing approaches for sound source localization, their merits and the factors that affect their performance.

1 Sound Source Localization

Speaker localization is of particular interest in applications that require information of the source position. Based on the localized speaker position, the microphone array can be steered in the corresponding direction to continue tracking the source if it's in motion. This approach is appropriate for a moving source (e.g. video-conferencing), where the source position estimate is input to a video-system [26]. Localization systems are also used in a multi-speaker scenario to enhance speech from a particular source with respect to other sources in the area of interest.

Existing source localization procedures may be loosely grouped into three general categories:

a. approaches employing time-difference of arrival (TDOA) information
b. techniques adopting high-resolution spectral estimation concepts and
c. principle of maximizing the steered response power (SRP) of a beamformer.

These broad classifications are delineated by the application environment and method of estimation. The *first* category includes procedures which calculate source locations from a set of delay estimates measured across various combinations of microphones in an array of sensors. The *second* method refers to any localization scheme relying upon an application of the signal correlation matrix for source position estimation. The *last approach* refers to any situation where the location estimate is derived directly from a filtered, weighted and summed version of the signal data received at the sensors.

2 Time Difference of Arrival – TDOA

The most widely used source localization approach exploits time-difference of arrival (TDOA) information. With this localization strategy, a two-step procedure is adapted. Time delay estimate (TDE) of the signals from a point source, relative to pairs of spatially distinct microphones is determined. This value along with knowledge of the microphone positions is used to determine an estimate for the source location. A specific delay can be mapped to a number of different spatial points along a hyperbolic curve, as illustrated in Figure 43. The curves are then intersected in some optimal sense to arrive at a source location estimate. A number of variations on this principle have been developed [22]. They differ considerably in the method of derivation, the extent of their applicability (2-D vs. 3-D, near vs. distant sources, etc.) and the means of arriving at the solution. Primarily because of their computational practicality and reasonable performance under amicable conditions, the bulk of *passive* talker localization systems in use are TDOA-based.

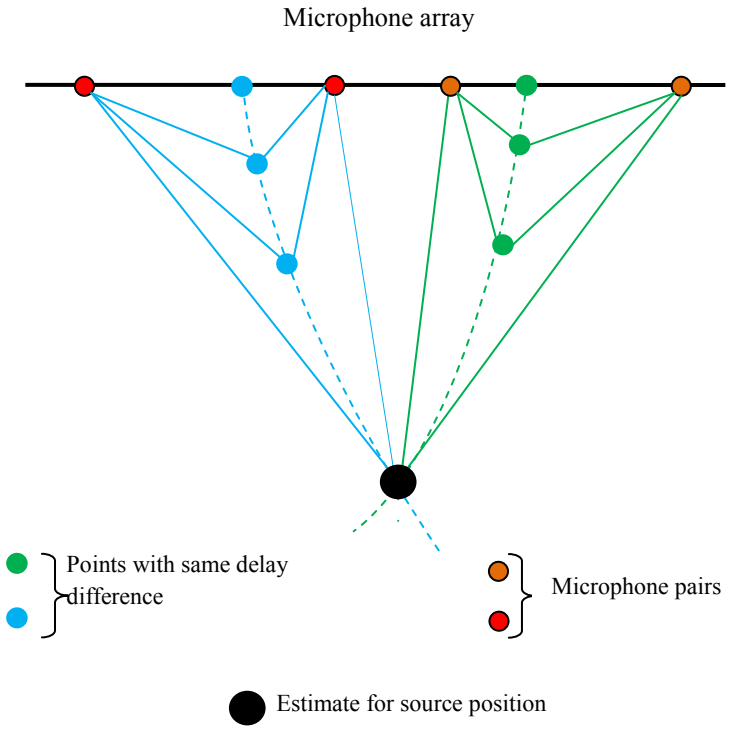


Figure 36: Sound source location using TDOA on a microphone array.

Acquiring good TDE of the received speech signals is essential in achieving effective speaker localization. The two major sources of signal degradation which complicate this estimation problem are background noise and channel multi-path due to room reverberations. The noise-alone case has been addressed at length and is well understood and applied successfully to speech source localization in low-multipath environments [30-32]. However, once room reverberations rise above minimal levels, TDOA based methods begin to exhibit dramatic performance degradations and become unreliable [31].

Another common limitation is the inability to accommodate multi-source scenarios. The algorithms assume a single source model. While some TDOA-based methods are used to track several individuals by operating at short analysis intervals, the presence of multiple simultaneous talkers, excessive ambient noise, or moderate to high reverberation levels in the acoustic field typically result in poor TDOA results and subsequently, unreliable location estimates. A TDOA based locator in such an environment would require a means for evaluating the validity and accuracy of the delay and location estimates.

3 High Resolution Spectral Estimation

This second categorization of location estimation techniques includes the modern beamforming methods adapted from the field of high-resolution spectral analysis like: autoregressive (AR) modeling, minimum variance (MV) spectral estimation, and a variety of Eigen-analysis based techniques (e.g. MUSIC algorithm) [45, 51]. Though the above approaches have been successfully implemented in a variety of applications, they possess certain restrictions that limit their effectiveness with the speech-source localization problem.

The high-resolution approaches discussed above are based on calculation of the spatio-spectral correlation matrix derived from the signals received at the sensors. This matrix is obtained from ensemble average of the signals over an interval in which the sources and noise are assumed to be statistically stationary. For practical situations, fulfilling these conditions is difficult and this contributes to performance degradation [26].

With regard to the localization problem, these methods were developed in the context of far-field plane waves coupled with the use of a linear microphone-array. Though the AR model and Eigen-analysis approaches are limited to the far-field, uniform linear array situation, the MV and MUSIC algorithms have shown to be extendible to the case of general array geometries and near-field sources [26]. As far as computational expense is concerned, a search of the entire location space is required. While the computation required at each iteration is lesser compared to steered-beamformer, the situation is compounded if a complex source model is adapted. Additionally, it should be noted that these high-resolution methods are all designed for narrowband signals. They can be extended to wideband signals, including speech, either through simple serial application of the narrowband methods or more sophisticated generalizations of these approaches [26]. Either of these routes increases the computational requirements considerably.

These algorithms tend to be significantly less robust to source and sensor modeling errors than conventional beamforming methods [26]. The incorporated models typically assume ideal source radiators, uniform sensor channel characteristics, and exact knowledge of the sensor positions. The sensitivity of these high-resolution methods to the modeling assumptions can be reduced, but at the cost of performance. Additionally, signal coherence, (reverberant conditions) is detrimental to algorithmic performance, particularly with the Eigen-analysis approaches.

4 Steered-Beamformer-Based Locators

This family of SSL approaches uses passive arrays for which the system input is an acoustic signal produced by the source. The optimal Maximum Likelihood (ML) location estimator in this situation amounts to a focused beamformer which steers the array to various locations and searches for a peak in output power (*focalization)* [44, 52]. Theoretical and practical variance bounds obtained via focalization are detailed in and the steered-beamformer approach has been extended to the case of multiple-signal sources in [26].

The simplest type of steered response is obtained using the output of a delay-and-sum beamformer. This is what is most often referred to as a conventional beamformer. Delay-and-sum beamformer applies time shifts to the array signals to compensate for the propagation delays in the arrival of the source signal at each microphone. These signals are time-aligned and summed together to form a single output signal. More sophisticated beam formers operate filters on the array signals as well as time alignment. The derivation of the filters in these filter-and-sum beam formers distinguishes one method from another.

Beamforming techniques have been applied to both source-signal capture and source localization. If the location of the source is known, then a beamformer can be focused on the source, and its output becomes an enhanced version of the inputs from the microphones. If the location of the source is not known, then a beamformer can be used to scan, or *steer*, over a predefined spatial region by adjusting its steering delays. The output of a beamformer, when used in this way, is known as the *steered response*. The steered response power (SRP) may peak under a variety of circumstances, but with favorable conditions, it is maximized when the steering delays match the propagation delays. By predicting the properties of the propagating waves, these steering delays can be mapped to a location, which should coincide with the location of the source.

Due to the efficiency and satisfactory performance of other methods, SRP has rarely been applied to the talker localization problem. The use of standard iterative optimization methods, such as steepest descent and Newton-Raphson, for this process was addressed by [53]. A shortcoming of each of these approaches is that the objective function to be minimized does not have a strong global peak and frequently contains several local maxima. As a result, this genre of efficient search methods is often inaccurate and extremely sensitive to the initial search location. In [5] an optimization method appropriate for a multi-modal objective function, Stochastic Region Contraction (SRC), while improving the robustness of the location estimate, the resulted in increased computation requirement compared to other less robust SSL techniques. The above factors have been major reasons to prohibit its use in the majority of practical, real-time source locators.

Furthermore, the steered response of a conventional beamformer is highly dependent on the spectral content of the source signal. Many optimal derivations are based on *prior*

knowledge of the spectral content of the back ground noise, as well as the source signal [26]. In the presence of significant reverberation, the noise and source signals are highly correlated, increasing estimation error. Furthermore, in nearly all array-applications, little or nothing is known about the source signal. Hence, such optimal estimators are not very practical in realistic speech-array environments.

The beamforming principle may be used as foundation for source localization by steering the array to various spatial points to find the peak in the output power. Localization methods based on the maximization of the steered response power (SRP) of a beamformer have been shown to be robust [26]. However, they present a high dependency on the spectral content of the source signal, which in most practical situations is unknown. The following chapters discuss a modified version of the SRP-PHAT algorithm [32] for use in sound source detection applications.

REFERENCES

1. L. Rabiner and B. Juang, Fundamentals of Speech Recognition. 1993: Prentice-Hall.
2. T. S. Huang, Multimedia/multimodal signal processing, analysis, and understanding, in First International Symposium on Control, Communications and Signal Processing. 2004.
3. M. Coen, Design principles for intelligent environments, in Proc. Conf. Artificial Intelligence. 1998.
4. Guillaume Lathoud, Jean-Marc Odobez, Daniel Gatica-Perez, AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking" Proc. of the Joint Work shop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI). 2005.
5. Hoang Tran Huy Do, Real-Time SRP-PHAT Source Location Implementations on a Large-Aperture Microphone Array,. 2007, Brown University.
6. Corporation, A.E. http://www.andreaelectronics.com/Buy/ProductDesc/Superbeam_Array.htm_ [cited]
7. Microsoft Inc. http://research.microsoft.com/users/ivantash/MicrophoneArrayProject.aspx http://windowsvistablog.com/blogs/windowsvista/archive/2007/09/24/using-a-mic-array-for-sound-capture.aspx [cited]
8. Brandstein, M., Adcock, J. and Silverman, H. , Microphone array localization error estimation with application to sensor placement. J. Acoust. Soc. Am., 1996. 99(6): p. 3807-3816.
9. Aarabi.P, The fusion of distributed microphone arrays for sound localization. EURASIP Journal of Applied Signal Processing (Special Issue on Sensor Networks), 2003.
10. J. L. Flanagan and H. F Silverman, Workshop on Microphone Arrays: Theory, Design & Application. 1994, CAIP Center, Rutgers University.

11. Rabinkin, D.V., Optimum sensor placement for microphone arrays. 1998, RUTGERS THE STATE UNIVERSITY OF NEW JERSEY.
12. C. H. Knapp and G. C. Carter, The generalized correlation method for estima tion of time delay. IEEE Trans. Acoust. Speech Signal Process, 1976. ASSP-24: p. 320-327.
13. M. Brandstein, J.A., and H. Silverman, A closed-form location estimator for use with room environment microphone arrays. IEEE Trans. Speech Audio Proc, 1997. 5: p. 45-50.
14. Gehrig, T.N., K.; Ekenel, H.K.; Klee, U.; McDonough, J. Kalman filters for audio-video source localization. in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 2005.
15. Ping Zou; Zheng Huang; Jianhua Lu, Passive stationary target positioning using adaptive particle filter with TDOA and FDOA measurements. in Joint Conference of the 10th Asia-Pacific Conference on Communications and the 5th International Symposium on Multi-Dimensional Mobile Communications Proceedings. 2004.
16. Julier, S. and Uhlmann, J, A new extension of the Kalman filter to nonlinear systems. in In Proceedings of AeroSense: the 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls. Multi Sensor Fusion, Tracking and Resource Management 1997: SPIE.
17. LaViola, J. A comparison of Unscented and Extended Kalman Filtering for estimating quaternion motion. in 2003 American Control Conference. 2003: IEEE Press.
18. Ward, D., Lehmann, E., and Williamson, R, Particle filtering algorithms for tracking an acoustic source in a reverberant environment. . IEEE Transactions Speech and Audio Processing 2003. 11(6).
19. Vermaak, J. and Blake, A, Nonlinear filtering for speaker tracking in noisy and reverberant environments. in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2001.
20. Larocque, J., Reilly, J., and Ng, W, Particle filters for tracking an unknown number of sources. IEEE Transactions on Signal Processing, 2002. 50(12).

21. Lehmann, E. Importance sampling particle filter for robust acoustic source localisation and tracking in reverberant environments. in Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA). 2005. Piscataway, NJ, USA.
22. Peterson, J. M. ; Kyriakakis, Chris, Hybrid Algorithm for Robust, Real-Time Source Localization in Reverberant Environments 2006, USC - Immersive Audio Laboratory.
23. Barry D. Van Even and Kevin M. Buckley, Beamforming: A versatile approach to spatial filtering. IEEE Acoustics, Speech, and Signal Processing Magazine 1988: p. 4-24.
24. Shi, G.; Aarabi, P, Robust speech recognition using near-field superdirective beamforming with post-filtering, in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing. 2003.
25. J. Fisher, T.D., W. T. Freeman and P. Viola, Learning Joint Statistical Models for Audio-Visual Fusion and Segregation, in Neural Information Processing Systems (NIPS). 2000.
26. Brandstein, M. and D. Ward, Microphone Arrays - Signal Processing Techniques and Applications., ed. Springer. 2001.
27. DiBiase, J.H., A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays. 2000, Brown University.
28. Tianshuang Qiu; Hongyu Wang, An Eckart-weighted adaptive time delay estimation method. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1996. 44(9): p. 2332-2335.
29. Kuhn, J., Detection performance of the smooth coherence transform (SCOT). IEEE Trans. Acoust. Speech and Signal Processing, 1978. 3: p. 678-683.
30. S. Bedard, B.C., and A. Stephenne, Effects of room reverberation on time-delay estimation performance, in IEEE Int. Conf. Acoust Speech, Signal Processing (ICASSP-94). 1994.
31. M. Brandstein and H. Silverman, A robust method for speech signal time-delay estimation in reverberant rooms, in IEEE Int. Conf. Acoust., Speed Signal Processing (ICASSP-97). 1997: Munich, Germany. p. 375-378.

32. Kevin Donohue, J.H., Hank Dietz, Performance of Phase Transform for Detecting Sound Sources with Microphone Arrays in Reverberant and Noisy Environments. 2007, Center for Visualization and Virtual Environments, University of Kentucky.
33. T. Gustafsson, B. Rao, M. Triverdi, Source localization in reverberant environments: modeling and statistical analysis. IEEE Transactions on Speech and Audio Processing, 2003. 11(6): p. 791-803.
34. J. H. DiBiase, H. Silverman, M. S. Brandstein, Robust Localization in Reverberant Rooms, in Microphone Arrays, Signal Processing Techniques and Applications. 2001, Springer Verlag, Berlin. p. 157-180.
35. K.D. Donohue, A.A., J. Hannemann, Audio signal delay estimation using partial whitening, in Proc. of the IEEE, Southeastcon 2007. March 2007. p. 466-471.
36. J.A.Hanley, B.J.McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982. 143 (1): p. 29-36.
37. Metz, C.E. Basic principles of ROC analysis. in Seminars in Nuclear Medicine 1978, pp. .
38. Trees, H.L.Van, Detection, Estimation, and Modulation Theory. Vol. 1. 2001: Wiley & Sons.
39. P. Svaizer, M. Matassoni, M. Omologo, Acoustic source location in a three-dimensional space using crosspower spectrum phase, in IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-97). 1997: Munich, Germany. p. 231-234.
40. B. Mungamuru, P.Arabi, Enhanced sound localization. IEEE Trans. on Systems, Man and Cybernetics, 2004. 34(3): p. 1526-1540.
41. Michael S. Brandstein, H. Silverman, A practical methodology for speech sourc localization with microphone arrays. Computer, Speech, and Language, 1997. 11: p. 91-126.
42. Kevin W. Wilson, T.D., Learning a Precedence Effect-like Weighting Function for the Generalized Cross-Correlation Framework. IEEE Journal on of Speech and Audio Processing, 2006.

43. Donohue, D.K. http://www.engr.uky.edu/~donohue/audio/Examples/Examples.html [cited; Audio lab setup information].
44. Ziomek, L., Fundamentals of Acoustic Field Theory and Space-Time Signal Processing. 1995: CRC Press.
45. Johnson, D., Dan E. Dudgeon, Array Signal Processing - Concepts and Techniques. 1993: Prentice Hall.
46. Jacobsen, F., The Sound Field in a Reverberation Room. 2006, Technical University of Denmark.
47. Scott M. Griebel and Michael S. Brandstein, Microphone Array Speech Dereverberation Using Coarse Channel Modeling, in IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings. 2001. p. 201-204.
48. Kai Yu, Boling Xu, Mingyang Dai, Chongzhi Yu, Suppressing cocktail party noise for speech acquisition, in 5th International Conference on Signal Processing Proceedings, 2000. 2000. p. 831-835.
49. Kuttruff, H., Room Acoustics. 3 ed. 1991: Elseiver Applied Science.
50. Jont B. Allen and David A. Berkley, Image Method for Efficiently Simulating Small-Room Acoustics. Journal of Acoustical Society of America, 1979. 65: p. 943–950.
51. Haykin, S., Adaptive Filter Theory. 2 ed. 1991: Prentice Hall.
52. W.J. Bangs and P. M. Schultheiss, Space-time processing for optimal parameter estimation. Signal Processing (J. Griffiths, P. Stocklin, and C. V. Schooneveld, eds.). 1973: Academic Press.
53. Wax, M.; Kailath, T., Optimum localization of multiple sources by passive arrays. IEEE Trans. Acoust., Speech, Signal Processing, 1983. ASSP-31: p. 1210-1217.

VITA

Anand Ramamurthy was born on 5th April 1984 in Chennai, India. He received his Bachelor's Degree in Electronics and Communication Engineering from Anna University, India in the year 2005. In pursuit of his higher education he attended The College of Engineering at University of Kentucky, Lexington. His research interests are in Signal and Image processing and he was a Graduate research assistant in the Audio Lab at the Center for Visualization and Virtual Environments, University of Kentucky.