2009

IMPACT OF MICROPHONE POSITIONAL ERRORS ON SPEECH INTELLIGIBILITY

Arulkumaran Muthukumarasamy
University of Kentucky, arul.171@gmail.com

ABSTRACT OF THESIS

IMPACT OF MICROPHONE POSITIONAL ERRORS ON SPEECH INTELLIGBILITY

The speech of a person speaking in a noisy environment can be enhanced through electronic beamforming using spatially distributed microphones. As this approach demands precise information about the microphone locations, its application is limited in places where microphones must be placed quickly or changed on a regular basis. Highly precise calibration or measurement process can be tedious and time consuming. In order to understand tolerable limits on the calibration process, the impact of microphone position error on the intelligibility is examined. Analytical expressions are derived by modeling the microphone position errors as a zero mean uniform distribution. Experiments and simulations were performed to show relationships between precision of the microphone location measurement and loss in intelligibility. A variety of microphone array configurations and distracting sources (other interfering speech and white noise) are considered. For speech near the threshold of intelligibility, the results show that microphone position errors with standard deviations less than 1.5cm can limit losses in intelligibility to within 10% of the maximum (perfect microphone placement) for all the microphone distributions examined. Of different array distributions experimented, the linear array tends to be more vulnerable whereas the non-uniform 3D array showed a robust performance to positional errors.

KEYWORDS: Speech intelligibility, Microphone array calibration, Delay-and-sum beamformer, Microphone positional errors, Speech intelligibility index

Arulkumaran Muthukumarasamy
June 10, 2009

IMPACT OF MICROPHONE POSITIONAL ERRORS ON SPEECH INTELLIGBILITY

By

Arulkumaran Muthukumarasamy

Dr. Kevin D. Donohue
Director of Thesis

Dr. Yu Ming Zhang
Director of Graduate Studies

June 10, 2009

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Unpublished theses submitted for the Masters degree and deposited in the University of Kentucky Library are as a rule open for inspection, but are to be used only with due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

Arulkumaran Muthukumarasamy

The Graduate School

University of Kentucky

2009

IMPACT OF MICROPHONE POSITIONAL ERRORS ON SPEECH INTELLGIBILITY

THESIS

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

By

Arulkumaran Muthukumarasamy

Lexington, Kentucky

Director: Dr. Kevin D. Donohue, Databeam Professor of

Electrical and Computer Engineering

Lexington, Kentucky

2009

DEDICATION

To my parents, relatives, and friends

ACKNOWLEDGEMENTS

Highly excited about the forth coming learning experiences, I commenced my odyssey. Although anxious and apprehensive about the uncertainties, I was ready to walk along the road. The journey turned out to be an enjoyable one. I would like to take this opportunity to express my gratitude to Dr. Kevin D. Donohue, for his solid motivation and guidance throughout this project. When I tumbled and fumbled you were always available to guide and encourage me by your example. Your constant advice and support helped me take the right decisions and made my work endearing. For every life there must be a path breaker and you have proved to be one for my success.

My thanks to my mother and father, Arularasi and Muthukumarasamy, Kannagi, Kasinathan, Ezhili and Arunrajen for their love, patience and showing immense confidence in my decisions and ideas. I would like to extend my greetings to all my friends for their care and inspiration when it was most needed. They exhibited an immense pleasure in assisting my endeavor and if not for them my stay in the university would not have been pleasanter.

I would also like to thank Dr. Lawrence Hassebrook, Dr. Robert Heath and Dr. Jens Hannemann for agreeing to take part in my defense committee and provide their valuable insight. Also, I must acknowledge the National Science Foundation *EPSCoR* program for funding in part this research work.

Thank you all.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Speech intelligibility

Even in today's modern multimedia society, speech is probably the most important and efficient means of individual communication. It is most often used to share information [1]. However, failure to understand the message at certain circumstances can be a result of several factors. A message spoken in Spanish to a listener who understands only Chinese may not be understood. Hence, a message has to be intelligent to be understood properly. An intelligent message in a language known to the listener could still be misunderstood if it is not audible or distorted by the environment [2].

Speech and Music are significantly different in their features. For example, at a cocktail party, people are talking with music running in the background. It would be hard to understand a particular person's speech unless the concentration is on his voice. The visual and gestural cues could be used to understand it even if only a fraction of speech is heard. But, in the mean time the music playing in the background might be recognized even in presence of noise with ease. Even if a fair amount of information is missed in the music, the brain is able to fill in the information due to the high degree of redundancy in music. However, since speech consists of a succession of sounds changing rapidly from instant to instant in intensity and frequency, it has less redundancy than music. Therefore it is hard to understand the normal speech even if some syllables are intelligible [3, 4].

Speech intelligibility is the measure of effectiveness of speech. It is defined as the degree to which the speech can be understood correctly by the listener [2, 5]. Intelligibility and speech quality are not equivalent. Speech quality refers to the quality of a reproduced speech signal with respect to amount of distortions and noise. A listener can completely understand a synthesized voice message which may be artificial and of low quality. A message may still be intelligible even if it lacks quality due to distortion [2, 6].

1.2 Factors governing Speech intelligibility

Speech intelligibility can be diminished or influenced by a number of acoustic, electronic and electromechanical factors [3]. It depends directly on the signal-to-noise ratio (SNR). It is quite complicated to deliver an intelligible speech to listeners in a real-world situation [1]. Many factors influence the speech and noise in a communication system, such as basic characteristics of speech and hearing, electrical and acoustic characteristics of the enclosure and behavioral conditions under which the communication takes place. These factors need to be considered to maintain the intelligibility in an enclosure [4]. The following figure shows the types of distortion that can be introduced in a communication system that governs intelligibility.

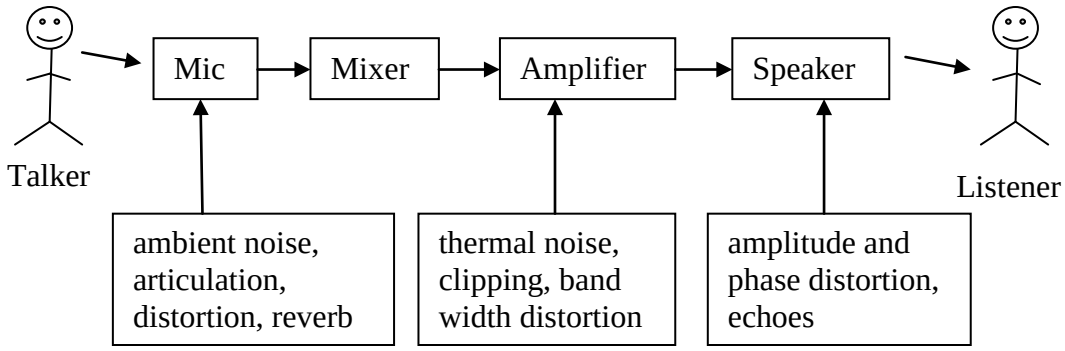


Figure 1: General speech communication enclosure with various distortions [2, 3]

There is a certain level of ambient background noise present in every acoustic environment. This intrusion of unwanted noise can mask the speech such that not all speech is available to the listener, thus reducing the SNR. This masking noise may be produced as a result of acoustical sources such as reverberation, ventilation or traffic. It may also arise electronically from thermal noises. Increasing the masking noise will clearly affect the intelligibility. Low frequency noise is more effective in masking as it masks both vowels and consonants unlike the high frequency noise which tends to primarily mask the consonant sounds. Competing human speech can also mask the desired speech, where the masking effect increases with the number and loudness of distracting voices [1, 3, 7].

Speech intelligibility is also affected significantly by a room's impulse response. Excessive reverberations and phase distortions contribute to the apparent background noise level which distorts the direct speech signal [1, 3]. Limitations in the bandwidth are also an important factor which affects intelligibility especially in telephonic conversations [8]. Intelligibility may also be affected by the predictability of the message, speaker's enunciation (accent) and also by the listener's hearing ability [3].

1.3 Measuring Intelligibility

With the development of telephone and other audio systems, speech intelligibility received a major attention from speech and audio processing researchers in the early part of the century. As a result, a subjective measure for intelligibility was proposed based on the use of physical speakers and listeners [1, 3, 6, 9]. This statistical procedure normally consists of a trained speaker reading out standardized word lists through the test system to a set of trained listeners. The percentage of recognized words or sentences is then taken as a measure of Speech intelligibility. But as these methods are time consuming, difficult to set up, and demand extensive statistical analysis, researchers opted for an automated, machine based test that quickly and easily estimates the intelligibility scores in speech systems. These objective measures are based on the physical parameters of the communication system to predict the intelligibility of those systems [1, 3, 6].

Subjective Measures

American National Standards Institute (ANSI) has approved a procedure for the subjective assessment measures as the standard *ANSI S3.2-1989, "Method for Measuring the Intelligibility of Speech over Communication Systems*". These subjective measures used trained talkers and listeners in their computations and are by far the most accurate and reliable methods for measuring intelligibility [3].

The subjective intelligibility measures generally differ on the usage of meaningful words or sentences during the evaluation of intelligibility. A variety of specialized word lists are in use for testing various aspects of speech communication. One of those

standardized word lists is the Modified Rhyme Test (MRT). It consists of 50 six-word lists of rhyming words constructed from a consonant-vowel-consonant sequence (see Table 1). The six words in each list differ only in the initial or final consonant sound. The talkers need to have good articulation and are trained to speak at consistent level. The listeners must have good discrimination and are familiar with all used test words and talker's voice. The talker and listener were given the whole list containing the words. The talker pronounces one of the six words in each list and the listeners identifies and marks the word they think the talker has spoken from the list. For example, suppose the talker pronounces the word 'Dent' from the first row in the list. The listeners have to circle/mark one of the words from that row that they think have been pronounced. This indicates their ability to differentiate the initial consonants. After the test is carried out the results are collected and analyzed statistically to indicate the errors in discriminating the initial and final consonant sounds [3, 6, 10].

Table 1: The first four rows of words in Modified Rhyme Test (MRT) [10]

Went	Sent	Bent	Dent	Tent	Rent
Hold	Cold	Told	Fold	Sold	Gold
Pat	Pad	Pan	Path	Pack	Pass
Lane	Lay	Late	Lake	Lace	Lame

ANSI standard specifies another similar method called the Diagnostic Rhyme Test (DRT). It consists of 96 rhyming pairs of words which differ by a single acoustic feature in initial consonants (see Table 2). The talker speaks one word at a time from the list and the listener has to mark the answering sheet with one of the two words he thinks is correct. For example, the talker chooses to test the feature 'Nasality' and hence pronounces the word 'Beat'. The listeners have to mark the word that they think was pronounced from the given list of words. The percentage of words that are correctly identified is then computed after the experiment. It has been suggested that consonants are more important for intelligibility than the vowels. These consonants are more sensitive to losses and additive impairments like noise, tones etc., as they are shorter in duration (10-100ms) and lesser in average power than the vowels. The final result of this

method provides valuable diagnostic information about the consonants that are hard to recognize and to be altered [3, 10, 11].

Table 2: The first four rows of word pairs in Diagnostic Rhyme Test (DRT) [10]

Voicing		*Nasality*		*Sustentation*		*Sibilation*		*Graveness*		*Compactness*	
Veal	Feel	Meat	Beat	Vee	Bee	Zee	Thee	Weed	Reed	Yield	Wield
Bean	Peen	Need	Deed	Sheet	Cheat	Cheap	Keep	Peak	Teak	Key	Tea
Gin	Chin	Mitt	Bit	Vill	Bill	Jilt	Gilt	Bid	Did	Hit	Fit
Dint	Tint	Nip	Dip	Thick	Tick	Sing	Thing	Fin	Thin	Gill	dill

Another word list called the Phonetically Balanced Word list (PB-50) is also used for measuring intelligibility subjectively. They contain monosyllabic test words in order to negate any influence of non-phonetic cues on the measured intelligibility. They were initially developed in Harvard University and the word lists mostly comprise of meaningless or jumble syllables [3, 6]. There are also other word lists available in practice such as Diagnostic Alliteration Test, Spelling Alphabet Test and Diagnostic Medical Consonant Test [3].

A set of percentage scores calculated from these measures shows the number of times the words were identified correctly by the listener which reflects the intelligibility of the system. The results are then adjusted mathematically to account for guessing. However, in real-time situations intelligibility is augmented as the speech consists of word flows or sentences [3, 10].

Objective Measures

The development of objective measures that predict intelligibility for various transmission channels began with the assumption that the intelligibility of speech signal is based on the sum of the weighted contributions from individual frequency bands. This idea was proposed between 1925 and 1930 by Fletcher and was later modeled by French and Steinberg in 1947 [1, 6]. It was described that the information content of a speech signal is not equally distributed along the frequency range of a speech signal. A model

was created where the response of the speech system is divided into twenty contiguous frequency bands each of them contributing to the intelligibility. The contributions of individual bands are summed to a total contribution that is defined by the *Articulation Index*. The Articulation Index ranges in value from zero to unity [3, 6].

The Articulation Index was the earliest attempt which uses the objective measures to predict the intelligibility in speech communication channel. Later two more measures called Speech Transmission Index (STI) and Speech intelligibility Index (SII) were introduced. These measures outmoded the Articulation Index since it did not effectively account for reverberation. The Speech Transmission Index was introduced in the early 1970's wherein the speech is modeled as an artificial test signal. The result of the analysis is an index ranging from 0 to 1. The STI accounts correctly for reverberation, noise, band-pass limiting and non-linear distortion. STI is standardized by *IEC standard 60268-16* (1998), and uses an amplitude modulation scheme to generate the test signal with speech like characteristics based on the concept that the speech can be described as a fundamental waveform that is modulated by low frequency systems and is analyzed for the modulation depth over the communication system. Reduction in the modulation depth results in the loss of intelligibility. Another method called Rapid Speech Transmission Index (RSTI) was developed as an alternative to the more complex STI measure. It used a speech as an excitation signal and measures only in two octave bands centered at 500Hz and 2 kHz respectively. RSTI has limitations as it does not account for system distortion and non-linear phase and amplitude [1, 3, 6].

Later in 1997 another objective measure called Speech Intelligibility Index was introduced, which estimates intelligibility using the physical parameters of the speech transmission channel. This method was proposed in the draft form as *ANSI s3.5 -1997*, "*American National Standards methods for Calculation of the Speech Intelligibility Index*". The SII also ranges from 0.0 (completely unintelligible) to 1.0 (perfect intelligibility). The SII accounts for band-pass limiting and noise but the effects of temporal and non-linear distortions are not directly included. SII demands higher computation but is the most robust and accurate of machine dependent intelligibility

measures. Under right conditions, it shows a good correlation with the subjective methods [3, 12, 13]. The SII model has been developed such that it predicts the average speech intelligibility for a desired speech-in-noise condition rather than the intelligibility of individual utterances/words [14].

1.4 Enhancement of Speech intelligibility

Research work on speech intelligibility has focused on augmenting intelligibility in multi-talker conditions and fluctuating noise sources. Various approaches have been introduced to enhance the intelligibility of speech under such conditions. Studies show that in a multi-talker environment, the intelligibility increases when the target and competing sources are spatially separated. Moreover, speech intelligibility is markedly increased as the number of competing sources decrease [15, 16]. Modifying the architecture of the enclosure such that it attenuates most of the noise was suggested for intelligibility enhancement. For example, intelligibility in a cockpit can be enhanced by insulating cockpits, muffling engines, widening broadcast bandwidth and using earplugs and headsets as an effort to match the ideal conditions.

Methods involving auditory processing and speech synthesis were also proposed to enhance intelligibility. For example, an approach using the Masking-Level Differences that measures the change in the masking effect of the noise in binaural hearing, relative to the change in the positions of signal and noise sources, was proposed to produce an improvement in the apparent SNR, without actually changing either the speech or the noise intensity [17]. Speech intelligibility can also be improved by slowing down the speech signals selectively and enhancing some important acoustic cues. For example, a speech synthesis method called Time Domain Pitch Synchronous Overlap-Add (TD-PSOLA) can be used to slow down speech by automatic pitch marking and later enhance the speech segments using algorithm of burst and fricative detection for improved intelligibility [18]. It can also be enhanced by using de-noising methods that typically increase the intelligibility of the signal by applying algorithms to suppress the

background noise such as Wiener filtering, spectral subtraction and minimum mean-square error log-spectral amplitude estimator [19-21].

However, the most common enhancement to speech intelligibility is made possible with the help of distributed microphone systems. Arrays are considered more advantageous than single distant microphones as it can reduce the room effects and additive noises in an enclosure. Arrays have been considered for various applications, such as talker localization, speech recognition and beamforming [22-24]**.** Beamforming techniques have been developed to steer the microphone array in order to receive signals from a desired direction, eliminating the signals from other directions to achieve substantial improvement in SNR of the output signal. This improvement in weighted SNR directly relates to a comparable improvement in speech intelligibility. It has been reported that at the threshold of intelligibility, every single dB improvement in SNR can increase the speech intelligibility by 10-15% [25]. This thesis mainly focuses on assessing array performance based on intelligibility and finding tolerances in microphone position errors in the beamforming process.

1.5 Motivation

Most array processing algorithms, especially beamforming methods, require the knowledge of the exact location of microphones and its geometry prior to data acquisition and processing. So in an effort to yield better intelligibility through beamforming, the three-dimensional positions of the microphones in the array need to be calibrated with the least possible error. For example, a delay-and-sum beamformer expects precise microphone positions (sub-centimeter accuracy) to estimate the source to microphone distances and to find required delays to time-align the microphone signals to beamform on a target location.

However, for larger arrays and applications involving quick placements of microphones, the determination of accurate estimates of the microphone positions is often challenging. Various measurement methods such as using hardware wirecloth, laser

devices or acoustic measurement are error prone at least on the order of centimeters. Even though the positions of microphones are estimated precisely with minimum error, small differences in the speed of sound and alteration in the position of microphones due to routine maintenance or human activity after calibration contribute to the error [26, 27]. These spatial precision errors translate into time delay errors between the microphone signals thus degrading the beamformer's output, in turn influencing the estimates of intelligibility. Hence, in order to understand the limits on the calibration process, the impact of these microphone positional errors on beamformer's response and intelligibility estimates should be examined.

1.6 Hypothesis

The non-precise microphone locations result in a loss of coherence (phase consistency) for signals arriving at each microphone, thus affecting the signal power at various frequencies on the beamformer's spectra. Analytical expressions are derived to show the impact of these location errors on the beamformed signal power. When the location error standard deviation increases over a particular value, the beamformer offers no enhancement to the target signal as a result of effective incoherent summation. Moreover, since the estimation of SII depends on the spectrum level of the signal and noise, power loss on the beamformed spectra leads to a decrease in SII estimate.

The main objective of this thesis is to understand the impacts of these location errors in the array on SNR and to propose tolerable limits on the amount of error on speech intelligibility. These location errors on microphone positions are modeled using random variables distributed uniformly in 3 dimensions, and their influence on Speech intelligibility are examined and compared to SNR. SII is used as the quantitative metric for estimating the intelligibility in the enclosure. Experiments and simulations are performed to present the relationships between the precision of microphone positions, SNR, and SII loss for a variety of microphone array geometries, target signals and distracting sources. The influence of different array distributions on their robustness to location error estimate was also examined.

1.7 Organization of the thesis

Chapter 2 presents the steps involved in computing SII using the ANSI s3.5 standard. It also gives an introduction to concepts of beamforming with respect to the delay-and-sum beamformer implementation for enhancing Speech intelligibility. The later section derives analytical expressions by modeling the microphone positional errors, to show its impact on the beamformer spectra and intelligibility.

Chapter 3 focuses on the implementation of the simulator that is used to demonstrate the impact of positional errors on SII. This chapter also discusses the details of analysis and the issues of variables used while performing the simulations and experiments. Moreover, it investigates the influence of array geometries and input SNR over intelligibility metrics using the results from the simulator.

Chapter 4 provides the specifications of the experimental setup which is used to collect data to assess the validity of the simulation results. It also presents the summative statistic results obtained from the analyses and proposes limits on tolerable error in microphone positions for speech intelligibility.

Chapter 5 summarizes the conclusions and also provides future directions for further research in the area.

CHAPTER 2

MODELING AND SENSITIVITY OF POSITIONAL ERRORS

This chapter gives a detailed description of the steps involved in the calculation of SII according to the ANSI s3.5 standard for a given speech- in-noise condition. The concepts of beamforming in microphone arrays and the mathematical modeling of a weighted delay-and-sum beamformer are discussed in Section 2.2. Section 2.3 mathematically models the microphone positional errors and derives analytical expressions to show its impact on frequency response of the beamformer and speech intelligibility.

2.1 Calculation of Speech Intelligibility Index

The SII is a measure of intelligibility that quantifies the proportion of audible and usable speech information for a listener [13]. For a given speech-in-noise condition, the SII calculation requires specific information about the speech spectrum, the noise spectrum and the auditory threshold. Both the speech and noise are filtered into frequency bands. The factor '*audibility*' is derived in each of the bands, indicating the proportion of speech cues that are audible in a given frequency band [13, 14]. The audibility of each band is then multiplied by the respective Band Importance Functions (BIF) value, and the SII is estimated by summing up the resulting values across the frequency bands. The speech and noise spectrum can be partitioned using any one of the procedures from the ANSI standard, each using a different number and size of frequency bands [12]. They are listed in descending order of accuracy as follows:

- Critical band, consisting of 21 bands
- One-third Octave band, consisting of 18 bands
- Equally Contributing band, consisting of 17 bands
- Octave band, consisting of 6 bands

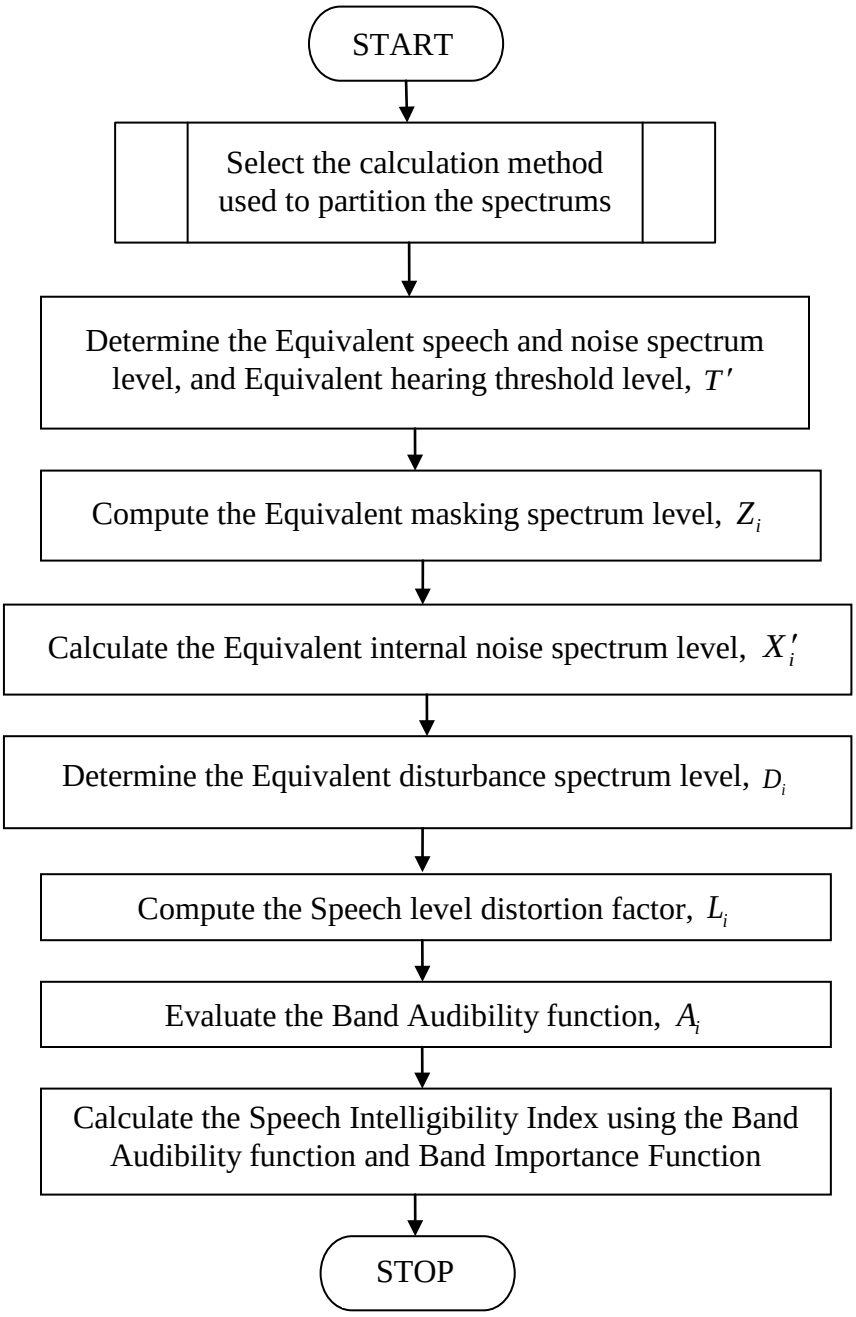


Figure 2: Flowchart describing the procedural steps involved in the estimation of SII

The flowchart in Fig. 2 discusses the variables to be estimated during the process of SII calculation. The first step is to select a band procedure to divide the speech and noise spectrums. Each band differs in some detail to compute SII although they are

conceptually the same. The SII standard gives the flexibility to choose any one of the bands depending on how specific the frequency measures are intended to be. Generally, when the measures are more frequency-specific (more bands), the SII computations are more accurate. The choice of any one procedure may also be influenced by the availability of the data as the Critical band features wide bandwidth (150 Hz to 8.5 kHz) whereas the Equally-contributing band needs lesser bandwidth (350 Hz to 5.8 kHz). The One-third octave method (bandwidth of 150 Hz to 8 kHz) is usually used as it corresponds with normal electro-acoustic analysis practices. Thus, in this thesis, the One-third octave band is assumed which divides the speech spectrum into 18 bands. The procedural steps involved in calculating the SII using ANSI S3.5 standard are described below [12]:

The Equivalent speech spectrum level, E', is the spectrum power levels of the target speech at each band center frequency given in Decibels (dB). The Equivalent noise spectrum level, N', is the spectrum power levels of the noise at the same band center frequency given in dB. Both these spectrum levels are based on the free-field levels. The free-field to eardrum transfer functions in Table 3 should be used if the speech is presented over the eardrum to project it into the free-field to yield the Equivalent speech and noise spectrum levels. The term noise includes both uncorrelated noise such as external noise, babble etc., as well as the noise correlated with the speech signal such as reverberation. For example, if the total signal received over a microphone in an array of *M* microphones is

$$y_m(t) = s_m(t) + n_m(t) \tag{1}$$

where $s_m(t)$ is the target signal and $n_m(t)$ is the noise signal at the m^{th} microphone. The target and noise signals are sent separately through a band-pass filter bank in order partition the spectrum into 18 bands based on the midband frequency given in the Table 3, and the spectrum power level in dB is estimated in each individual band i for the speech and noise using the power equations as below:

$$E_i = 10 \log \left\{ \lim_{T \to \infty} \frac{1}{T} \int_0^T \left(\left| s_{m,i}(t) \right| \right)^2 dt \right\} \tag{2a}$$

$$N_i = 10 \log \left\{ \lim_{T \to \infty} \frac{1}{T} \int_0^T \left(\left| n_{m,i}(t) \right| \right)^2 dt \right\} \tag{2b}$$

The minimum sound pressure level of the pure tone that is capable of evoking an auditory sensation at a specific frequency gives the pure-tone threshold level, which is determined by an appropriate psycho-acoustical method. The hearing threshold level is given as the difference between the pure-tone threshold level of a given ear at a specified frequency and the reference pure-tone threshold level. Then, the Equivalent hearing threshold level, T', for monaural listening is defined as the arithmetic average of the hearing threshold levels across the group of ears for which the SII calculations are performed. In general, for listeners in the 18-30 age groups, with no hearing loss, the equivalent hearing threshold is a hearing level of 0 dB across all frequencies. In case of the binaural listening, the value of the equivalent hearing threshold level for monaural listening should be decreased by 1.7dB.

The next step is to calculate the equivalent masking spectrum level, Z_i, which is defined as the sound pressure spectrum level in dB that appropriately accounts for the masking of speech produced by the equivalent noise. It comprises masking from within-band, out-of-band (spread of masking) and masking of one speech frequency by another (self-speech masking). In case of the one-third octave band method, the following parameters need to be calculated in order to determine the equivalent masking spectrum level. The self-speech masking spectrum level, V_i, calculates masking of higher speech frequencies by lower speech frequencies in conditions of severe low-pass or band-pass filtering, given in dB. This parameter is determined by subtracting a constant 24 dB (based on subjective testing) from the Equivalent speech spectrum level. The self-speech masking spectrum level, V_i is determined for each calculation band *i* using the equation

$$V_i = E_i' - 24\text{dB} \tag{3}$$

where E_i' is the Equivalent speech spectrum level. For the one-third octave band procedure the index *i* runs from 1 to 18.

For each calculation band *i*, the value of variable B_i in dB is determined which gives the larger of the equivalent noise spectrum level, N_i', or the self-speech masking spectrum level, V_i, which can be expressed as

$$B_i = \begin{cases} N_i', if\ N_i' > V_i \\ V_i, if\ V_i > N_i' \end{cases} \tag{4}$$

The next step is to determine C_i which is the slope per octave (doubling of frequency) of the upward spread of masking in dB/octave for each calculation band. For one-third octave frequency bands, the slope C_i is calculated using the relation given below:

$$C_i = -80\text{dB} + 0.6[B_i + 10\log F_i - 6.353\text{dB}] \tag{5}$$

where B_i is obtained from Eq. (4) and F_i is the nominal midband frequency of the one-third octave band in Hz as listed in Table 3. For the lowest frequency calculation band the equivalent masking spectrum level Z_i is equal to B_i. For all but the lower frequency calculation band, the equivalent masking spectrum level Z_i is determined using the equation

$$Z_i = 10\log\{10^{0.1N_i'} + \sum_{k}^{i-1} 10^{0.1[B_k + 3.32C_k \log(0.89F_i/F_k)]}\} \tag{6}$$

where N_i' is equivalent noise spectrum level, B_k is the same as B_i, F_i is the nominal one-third octave midband frequency as in Table 3 and F_k is the nominal midband frequency for frequency bank *k* as in Table 3.

The reference internal noise spectrum level in the ear of the listener is standardized by ANSI and is listed in Table 3. The reference internal noise spectrum level increased by the equivalent hearing threshold level would give us the equivalent internal noise spectrum level, X_i' in dB. It is given by the equation

$$X_i' = X_i + T_i' \tag{7}$$

where X_i is the reference internal noise spectrum level listed in the Table 3 and T_i' is the estimated equivalent hearing threshold level.

The Equivalent disturbance spectrum level D_i is estimated as the larger of the equivalent masking spectrum level Z_i and the equivalent internal noise spectrum level X_i'.

$$D_i = \begin{cases} Z_i, if\ Z_i > X_i' \\ X_i, if\ X_i' > Z_i \end{cases} \tag{8}$$

The speech level distortion factor accounts for the decrease in the intelligibility of speech at high presentation levels. It reaches unity when there is no distortion due to presentation level. Its value decreases to a minimum of zero at high presentation levels. The speech level distortion factor L_i is computed using the equation:

$$L_i = 1 - (E_i' - U_i - 10\text{dB})/160\text{dB} \tag{9}$$

where E_i' the Equivalent Speech Spectrum level and U_i is the standard speech spectrum level at the normal vocal effort found in Table 3. Eq. 9 is developed from the data given in reference and the constant (10) is the difference between 72.35 dB and overall level standard speech at normal vocal effort (62.35 dB from Table 3). In the event that a different vocal effort such as raised or loud is used, this constant can be modified according to the new overall level of the standard speech at the stated vocal effort. A value of one should be used if the calculated value of distortion factor exceeds one. A temporary variable K_i is calculated as follows:

$$K_i = (E_i' - D_i + 15\text{dB}) / 30\text{dB} \tag{10}$$

where E_i' and D_i are the Equivalent speech spectrum level and equivalent noise spectrum level respectively. The value of K_i should be limited between the interval [0, 1]. If the estimated value K_i is greater than 1 then it should be set to 1. If it is negative then the value of K_i is set to 0.

The Band Audibility function is calculated using the below equation:

$$A_i = L_i K_i \tag{11}$$

where L_i is the speech level distortion factor calculated from Eq. (9) and K_i is the value computed in Eq. (10).

The Speech intelligibility Index is then estimated as:

$$SII = \sum_{i=1}^{n} I_i A_i \tag{12}$$

where I_i is the Band Importance Functions as listed in Table 3 and A_i is the Band Audibility function computed using Eq. (11).

Table 3: One-third octave band SII procedure – frequency bands, standard speech spectra, internal noise and free field to eardrum transfer function (Adapted from ANSI s3.5-1997 [12])

Band No.	Frequency Band			Standard speech spectrum level for stated vocal effort, dB				Reference internal noise spectrum level, dB	Free-field to eardrum transfer function, dB
	Nominal midband freq(Hz)	Bandwidth adj, dB	Band Importance	Normal	Raised	Loud	Shout		
1	160	15.65	0.0083	32.41	33.81	35.29	30.77	0.60	0.00
2	200	16.65	0.0095	34.48	33.92	37.76	36.65	-1.70	0.50
3	250	17.65	0.0150	34.75	38.98	41.55	42.50	-3.90	1.00
4	315	18.65	0.0289	33.98	38.57	43.78	46.51	-6.10	1.40
5	400	19.65	0.0440	34.59	39.11	43.30	47.40	-8.20	1.50
6	500	20.65	0.0578	34.27	40.15	44.85	49.24	-9.70	1.80
7	630	21.65	0.0653	32.06	38.78	45.55	51.21	-10.80	2.40
8	800	22.65	0.0711	28.30	36.37	44.05	51.44	-11.90	3.10
9	1000	23.65	0.0818	25.01	33.86	42.16	51.31	-12.50	2.60
10	1250	24.65	0.0844	23.00	31.89	40.53	49.63	-13.50	3.00
11	1600	25.65	0.0882	20.15	28.58	37.70	47.65	-15.40	6.10
12	2000	26.65	0.0898	17.32	25.32	34.39	44.32	-17.70	12.00
13	2500	27.65	0.0868	13.18	22.35	30.98	40.80	-21.20	16.80
14	3150	28.65	0.0844	11.55	20.15	28.21	38.13	-24.20	15.00
15	4000	29.65	0.0771	9.33	16.78	25.41	34.41	-25.90	14.30
16	5000	30.65	0.0527	5.31	11.47	18.35	28.24	-23.60	10.70
17	6300	31.65	0.0364	2.59	7.67	13.87	23.45	-15.80	6.40
18	8000	32.65	0.0185	1.13	5.07	11.39	20.72	-7.10	1.80
Overall SPL, dB				62.35	68.34	74.85	82.30		

2.2 Beamforming using Microphone arrays:

Microphone array processing generally refers to combined processing of the signals obtained from spatially separated coherent sensors. These arrays are commonly used to enhance the SNR in a noisy environment with the help of spatial filtering. Spatial filtering is a technique by which a signal from a desired direction can be received, eliminating the signals from all other directions. Beamforming is a versatile approach of spatial filtering [29-31]. Beamforming algorithmically steers the microphone array in order to receive signals from the desired direction (look direction) and attenuate the signals from other directions. Beamforming methods can be broadly classified into two types, fixed and adaptive beamforming techniques. Delay-and-sum Beamforming and Filtered Delay-and-sum Beamforming belong to the fixed Beamforming category and Frost's Beamformer and the Griffiths-Jim Beamformer belong to adaptive Beamforming. Beamforming can be applied to both source-signal capture and localization of sound sources [30-33].

In a delay-and-sum beamformer, the signals received from the microphones are time aligned to adjust for the differences in the path length for the signal to reach the microphones from the source. These time aligned signals are weighted and summed together to get the output signal. All the noise signals that remain misaligned get attenuated when the signals are added. Instead of directly summing the time aligned signals, filtering the time aligned signals would achieve better attenuation of the interfering signals [22, 31, 33]. Adaptive Beamforming techniques aim to adjust the array processing parameters dynamically according to an optimization criterion either on sample by sample basis or frame by frame basis [30].

The performance of the beamformer can be usually determined using various metrics such as Direct to Reverberant ratio (DRR), Directivity index, SNR improvement and Intelligibility measures [34, 35]. For an impulse response, DRR can be given as the ratio of the direct path energy to the reverberant path energy. It is mainly used to quantify reverberant suppression in an enclosure rather than intelligibility related effects.

Directivity index gives the ratio of the array output power due to source in the target direction to the output power due to sound arriving from all other directions. But it cannot be used to assess the array for use in speech enhancement as it is a narrow band performance metric. Improvements in the output SNR with respect to input SNR are also used to assess the performance of the array. Another method called intelligibility averaged gain which is based on the well known Articulation Index is also used to measure the array performance. However, since the AI method was outdated by more recent and reliable measures of intelligibility like SII, SII is used to assess the performance of the array in this work. The SII is generally a frequency-weighted SNR metric which is calculated using the method discussed in the previous section.

2.2.1 Delay and Sum Beamformer

The delay and sum beamformer is based on the idea that the outputs of all the sensors are the same except that they are delayed by a different amount. The size of the delays is determined by the direction (far-field) or point (near-field) at which the microphone array is steered. Here, the array is focused on a source at a near-field point (spherical wavefronts) inside the array. The direction of propagation of the source to each microphone varies in case of a near-field source and thus delay is related to the distance between the source and the microphones in the array [32, 36].Consider an array of *N* microphones and sound sources at different spatial locations distributed in a 3-D space. Let $u_i(t;\vec{r}_i)$ be the pressure wave resulting from the i^{th} sound source located at known position $\vec{r}_i$, where $\vec{r}_i$ is the vector denoting the coordinates of $x, y,$ and z axis. Then, the waveform received at the m^{th} microphone located at $\vec{r}_m$ is given by [37, 38]:

$$v_m(t;\vec{r}_m,\vec{r}_i) = \beta_m u_i(t+\tau_m;\vec{r}_i) + n_m(t) \tag{13}$$

where $u_i(t+\tau_m;\vec{r}_i)$ is the delayed version of the source signal located at $\vec{r}_i$, τ_m is the direct path time delay to the m^{th} microphone from $\vec{r}_i$, β_m is the signal attenuation at the

m^{th} microphone located at $\vec{r}_m$ and $n_m(t)$ represents all the uncorrelated additive noise sources.

The below figure shows that for an array of *N* microphones, a delayed version of the source signal $u_i(t)$ exists in each microphone channel. The delayed versions of $u_i(t;\vec{r}_i)$ can be time-aligned using the actual delays (τ_m) and applying the weights (α_m) to the signals received at each microphone. The resulting signals can be summed together so that it reinforces the desired speech signal while the unwanted off-axis noise signals are combined in a more unpredictable and non-coherent fashion. Generally, the SNR of the total output signal is greater than (or at worst, equal to) that of any individual microphone's signal [33, 39].

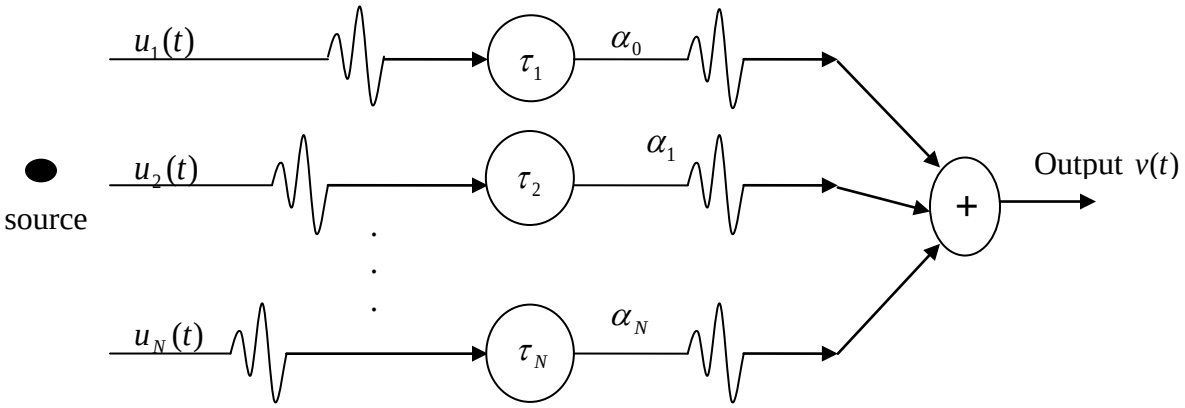


Figure 3: Delay and Sum Beamformer

By time-aligning the microphone signals, i.e., the delayed versions of $u_i(t;\vec{r}_i)$, the resulting signals can be summed together so that all copies add constructively while the uncorrelated noise signals present in $n_m(t)$ cancel [31, 32]. Assuming that the positions of the source (near-field) and the microphones in the array are known, the actual delays τ_m can be estimated for each of the microphones using the distance between the microphones to the source. The distance of the microphones from the position of the source can be computed as:

$$d_m = \left| \vec{r}_i - \vec{r}_m \right| \tag{14}$$

where $\vec{r}_i$ and $\vec{r}_m$ are the vector coordinates of the source and microphone positions in 3-D.

Using the speed of sound *c*, the time difference of arrivals of signal between the microphones can be computed. The closest microphone to the sound source is chosen as the reference microphone and receives zero delay. All other microphones receive a delay τ_m equal to the time difference of arrival as:

$$\tau_m = \frac{(\max(d_m) - d_m)}{c} \tag{15}$$

where d_m is the distance between the m^{th} microphone and the source of interest from Eq. (12), c is the speed of sound. If the positions of the signal source and microphones are not known, the delays between the microphone signals can be found using the cross correlation between the microphone signals. The delays usually correspond to the maximum value of the correlation between the microphone signals.

Once the actual delays τ_m are known, each received signal at the microphones can be appropriately delayed. The individual microphone signals are then weighed by a factor α_m before summing up. These weights can be chosen to be either uniform or variant using several methods based on frequency or delays of the microphone signals [30, 37]. The beamformed output of the i^{th} source is then the sum of N scaled copies of the signal $u_i(t)$ with N uncorrelated additive noise sources [37]:

$$b_i(t; r_i) = \sum_{m=1}^{N} \alpha_m u_i(t; \vec{r}_i) + n_m(t - \tau_m) \tag{16}$$

Separating the noise term from Eq. (16), the beamformed output which has the maximum possible target SNR can be given by the equation:

$$b_i(t;\vec{r}_i) = \sum_{m=1}^{N} \alpha_m u_i(t;\vec{r}_i) \tag{17}$$

However, the major disadvantage of delay-and-sum beamforming systems is that a large number of sensors are required to improve the SNR. Delay- and-sum beamforming results in a 3 dB increase in the output SNR for every doubling number of microphones (assuming all microphones have equal SNR values) [30, 37]. This improvement in SNR directly relates to a significant improvement in speech intelligibility. Also, the beamformer seeks only to enhance the signal in the direction to which the array is currently steered. Another limitation of the delay-and-sum beamformer is its inability to adapt to changing noise conditions and reverberations [33]. An adaptive beamformer such as the Griffiths-Jim Beamformer can be used in such cases for improved performance. However, this thesis considers only the delay-and-sum beamformer for analysis as most other array-beamforming methods are variations or extensions of this basic beamformer.

2.3 Calibration of Microphone Positions

Microphone array systems have the ability to provide quality acquisition and enhancement of speech from individual targets in a multi target environment. Such large aperture arrays are now becoming feasible in common environments, as the cost of supporting computing technology is diminishing. However, calibration of such arrays is an important issue to be considered while modeling these arrays [26, 27]. Many array processing methods require knowledge of microphone locations prior to data acquisition and processing. The three-dimensional position coordinates of the microphones in the array have to be estimated with the least possible error. For example, a delay-and-sum beamformer needs precise microphone positions (sub-centimeter accuracy) to estimate the source to microphone distances, and find required delays to time align a signal located at a point of interest [27, 40]. So, in order to yield better intelligibility through beamforming methods, the microphone positions need to be calibrated precisely.

However, the challenging problem is deriving accurate estimates of the microphone positions in an array. In real-time, direct or remote measurements to estimate the positions of microphones are prone to errors at least on the order of centimeters [26, 27]. The microphone arrays are usually built in different configurations and various methods are in practice to calibrate their microphone positions. The conventional methods such as using hardware wirecloth or foam mounts are still used to measure the exact positions of the microphones in an array. These methods have the tendency to "bow" and are difficult to implement with sufficient accuracy for larger arrays. Even the calibration by a laser transit is simply too error-prone, time consuming and tedious [26, 27]. The recent automatic calibration systems using the acoustic signals also involve some errors in calibration while estimating the positions of the microphones.

In addition to the errors caused by the direct, remote or acoustic calibrations, several other factors also contribute to the error. Small differences in speed of sound can contribute measurable error when the distances to the array are sufficiently large. Another source of error might be due to the small shifts in the positions of the microphones after the physical measurements were taken. Sometimes the panels on which the microphones are mounted could be altered during routine maintenance or by human activity which may cause some errors and demand recalibration [26, 27]. These measurement errors affect the SNR which in turn degrades Speech intelligibility. Analytical expressions are derived in the next section to show the impact of these positional errors on the beamformed spectra.

2.3.1 Modeling location errors:

The delay-and-sum beamformer is assumed to beamform the target speaker at a known location in the array. The delayed individual microphone signals can be given as in Equation (13). To obtain the output of the beamformer, the delays have to be estimated and each microphone signal is shifted according to the delay. The microphones are weighed based on the reciprocal of their distance to the source and are scaled such that

the closest microphone gets a weight of 1. Therefore, for a given set weights, α_m, and number of microphones *N*, the output of the delay-and-sum beamformer:

$$b_i(t;\vec{r}_i) = \sum_{m=1}^{N} \alpha_m v_m(t - \hat{\tau}_m;\vec{r}_i) \tag{18}$$

where, $\hat{\tau}_m$ is an estimate of the actual delay computed from time-delay measurements.

Any error in the computed geometric distances due to non-precise microphone positions will affect the estimated delays. Also, inaccuracy in the estimation of speed of sound may also contribute to the error in these estimated delays. To investigate the errors in calibrating the position of microphone placements and its relationship to output of the beamformer, a delay estimation error $\hat{e}_m$ is given as:

$$\hat{e}_m = \hat{\tau}_m - \tau_m \tag{19}$$

where τ_m is the true delay and $\hat{e}_m$ is a uniformly distributed random variable with zero mean and variance $\sigma_m^{\,2}$. For a zero error in estimated delay, the individual microphone signals are delayed appropriately and the beamformer's output would have N scaled copies of the source signal $u_i(t;\vec{r}_i)$ as in Eq. (17). However, as a result of these estimation errors, $\hat{e}_m$, the delayed version of the microphone signals do not align from the source signal, and thus the output of the beamformer with microphone position errors is given by

$$\hat{b}_i(t;\vec{r}_i) = \sum_{m=1}^{N} \alpha_m u_i((t - \hat{e}_m);\vec{r}_i) \tag{20}$$

The impact of this error on the gain of the main lobe of the array beam field is better seen in the frequency domain, so the Fourier transform of Eq. (20) becomes:

$$\hat{b}_i(\omega;\vec{r}_i) = \sum_{m=1}^{N} \alpha_m u_i(\omega) e^{-j2\pi f \hat{e}_m} \tag{21}$$

To show the spectral power loss due to the precision errors in microphone placements, convert the frequencies to wavelengths (λ), using the speed of sound '*c*', to obtain

$$\hat{b}_i(\omega;\vec{r}_i) = u_i(\omega) \sum_{m=1}^{N} \alpha_m e^{-j2\pi c\left(\frac{\hat{e}_m}{\lambda}\right)} \tag{22a}$$

Equation (22) shows that the exponential term in the beamformer's response depends on the error in the estimated delay relative to the wavelength of the source. To be more consistent with the errors with respect to distance, the delay estimation error $\hat{e}_m$ which is estimated in time is used to introduce a new variable called spatial distance positional errors $\hat{E}_m$ in the Eq. 22(a) such that $\hat{E}_m = c\hat{e}_m$. Then, the Eq. 22(a) can be rewritten as

$$\hat{b}_i(\omega;\vec{r}_i) = u_i(\omega) \sum_{m=1}^{N} \alpha_m e^{-j2\pi\left(\frac{\hat{E}_m}{\lambda}\right)} \tag{22b}$$

Note that if this spatial error is larger than, or on the order of the wavelength, the mean of the exponential summation is close to zero. This scaling down is ideally what happens to sound sources that are not at the location of interest. However at the location of interest, spatial positional errors result in power loss in the main lobe of the beam field. This can be quantified by taking the expected value of Eq. (22b) over the error terms from the array to result in

$$\mathrm{E}\left[\hat{b}_i(\omega)\right] = u_i(\omega) \sum_{m=1}^{N} \alpha_m \mathrm{E}\left[e^{-j2\pi\left(\frac{\hat{E}_m}{\lambda}\right)} \right] \tag{23}$$

If there is no precision error in the microphone placement, the expected value becomes one. Once the distributions of the precision errors are known, the expected value of Eq. (23) can be obtained. In the case of a zero mean uniform distribution with standard deviation σ, the expected value becomes [24, 41]:

$$\mathrm{E}\left[\exp(-j2\pi\left(\frac{\hat{E}_m}{\lambda}\right)\right] = \mathrm{sinc}\left(\pi\frac{\sqrt{12}\sigma}{\lambda}\right) \tag{24}$$

Equation (24) predicts the scaling/attenuation of the beamformed target signal based on the signal wavelength and the standard deviation of the spatial positional errors. Figure 4 plots the relationship between the sinc function and σ over wavelength. The expected value (sinc function) reaches a maximum of 1 when there is no error (σ/λ = 0). The figure indicates that the expected value goes to 0 when σ approaches the wavelength divided by √12 or approximately a quarter of the wavelength (σ/λ = ±0.288). The figure also suggests that the expected value never reaches 1, once σ/λ reaches ±0.288 (1/√12) i.e., after σ reaches the quarter wavelength. Therefore, for frequencies beyond this quarter wavelength limit for the standard deviation of the microphone location precision, the beamformer offers no enhancement for the target signal. Eq. (24) mainly predicts the signal loss at various frequencies for a given microphone precision error.

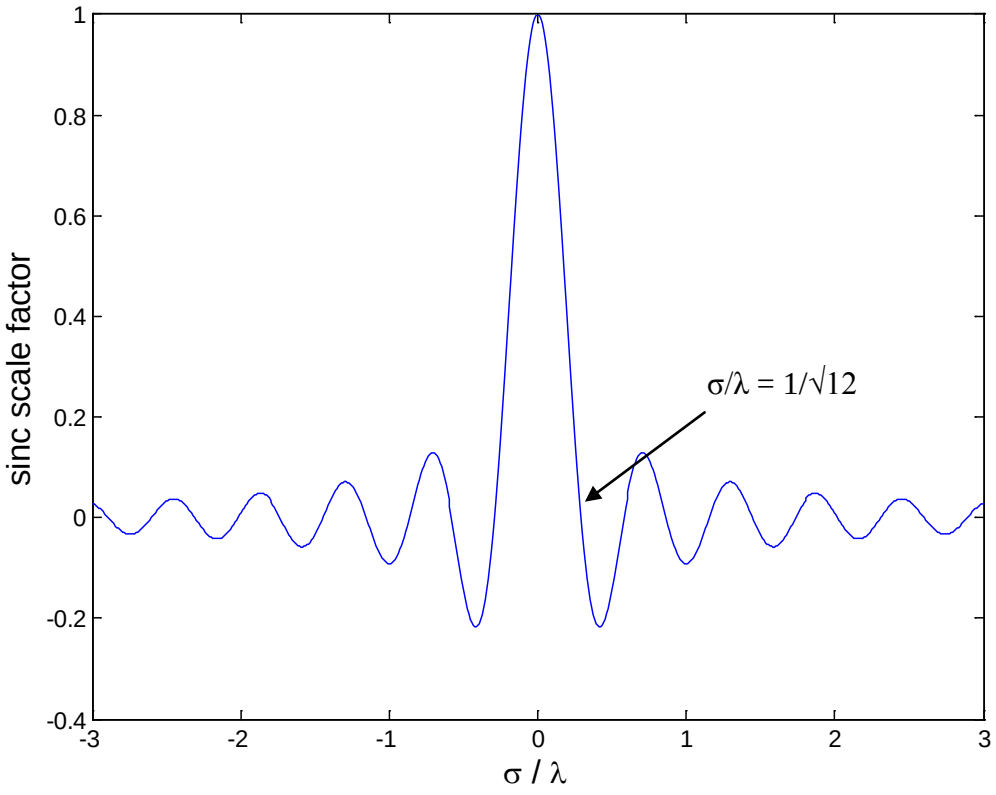
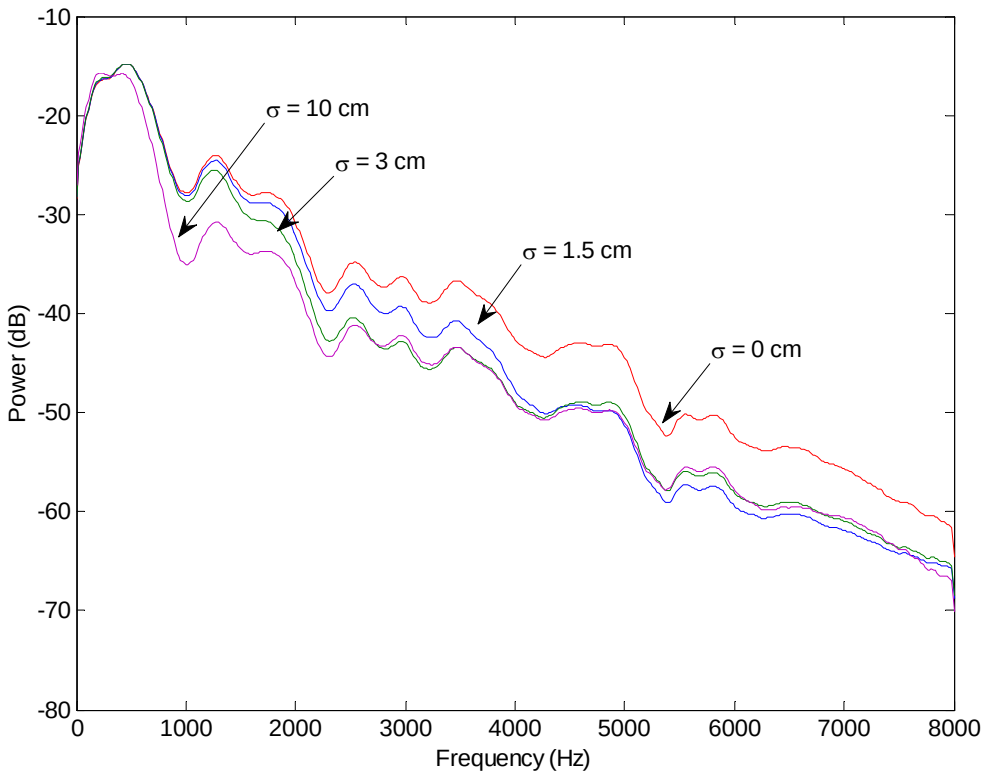


Figure 4: sinc function vs. σ/λ

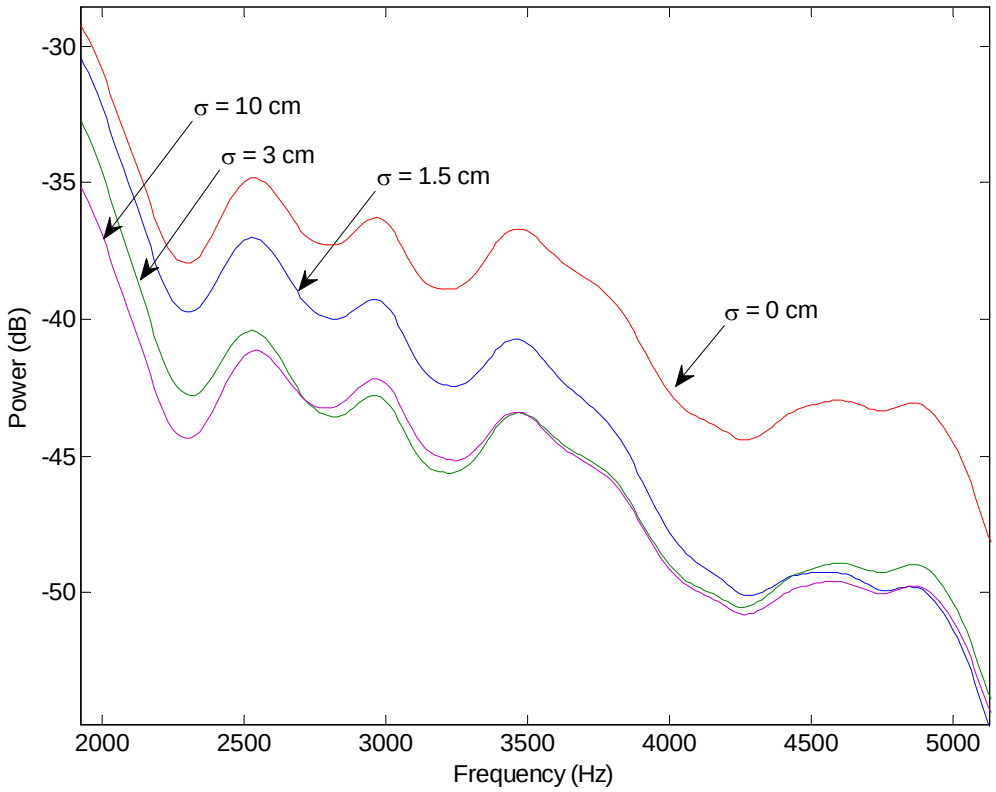
2.3.2 Sensitivity of precision error in Microphone position

To illustrate the frequency sensitivity of the location error on beamformer power gain, the power spectra of the beamformed signal for a range of standard deviations on the location errors are plotted. The simulation of the microphone positions and sound propagation gives complete control over the location error. The array for the simulation was a planar microphone geometry consisting of 16 microphones equally spaced on a Cartesian grid at 1.2m, above a 3.6x3.6x2.1m field of view (FOV). A speech signal (a male speaker single-microphone recording), positioned centrally and 1.1 m above the floor within the FOV, was simulated over the array using [41, 42]. The speed of sound was assumed to be 345 m/s with no reverberations and an air attenuation of -3.28e-5 dB per meter-Hz. Power spectra of the beamformed signal for several standard deviations of microphone position error are shown in Fig. 5.



(a)

Figure 5: Normalized power spectrum of Beamformed signals for given Precision error standard deviation σ (a) Normal version (b) Zoomed-in version



(b)

Figure 5, continued

The delay error and the microphone location errors are well associated to each other. Even though the random variables are used to define the delay estimation error in Equation (19), it reflects the probable spatial positional errors on the microphone placement. These placement errors lead to the error in the delay estimates. Therefore, the uniformly distributed random variables can be introduced on the microphone positions to result in a random error in the delay estimates. The plots in Fig. 5 compare power spectra of the beamformed signal with no error to those with standard deviations of 1.5 cm, 3 cm, and 10 cm. The figure illustrates the impact of the microphone location error on the loss of power as a function of frequency, due to the sinc function given in Eq. (24). For error using a 10 cm standard deviation for position error, Eq. (24) predicts (using the quarter wavelength approximation at a sound speed of 345 m/s) that wavelengths less than 0.4

meters (frequencies greater than 862 Hz) will not benefit from the coherent summation of the beamformer. Thus for the range shown, the 10 cm error is the lower limit for beamformer performance, since the error is so larger relative to the relevant wavelengths that coherence is not utilized for a power gain. Similarly, the 0 error spectrum is the upper limit for all frequencies.

Introduction of the precision error in microphone positions tends to portray a low pass effect on the spectrum. It can be seen that as the standard deviation of the precision error increases, it starts to affect the middle frequency components of the beamformer's signal spectra. But once the error reaches a particular value, the power loss decreases as a result of coherence and this trend is frequency dependent as shown in Fig. 5(a). The power for the 1.5 cm standard deviation drops from the upper limit to the lower limit over the range shown in Fig. 5. The error begins to affect the spectrum starting from around 1500 Hz. This corresponds to wavelengths approximately 15 times larger than the standard deviation and a scale factor of 0.92 from the sinc function of Eq. (24). As the frequency approaches 5000 Hz, the 1.5 cm standard deviation beamformer signal merges with the 10 cm standard deviation beamformer signal. This corresponds to a wavelength 4.6 times larger than the standard deviation of the error and a scale factor of 0.3 from the sinc function.

2.3.3 *Impact on SNR and Intelligibility*

This section examines the effect of microphone location errors on the SNR and speech intelligibility. The spectral power losses due to positional errors lead to a decrease in the SNR and also the estimate of SII. As shown in Fig. 6, the middle frequency bands (1500-5000 Hz) correspond to the most significant weights in BIF while estimating the SII. From previous analyses it can be seen that for location errors with 1.5 cm standard deviation, the beamformer offers no enhancement for frequencies greater than 5750 Hz. Hence, as error standard deviation increases greater than 1.5 cm, it results in significant power loss in this middle frequency range. So, precision errors on the order of few centimeters are expected to result in a considerable SII loss for beamformed speech in

noise. The power loss in the beamformer's spectra also degrades the output SNR of the beamformer. A loss of 1-3 dB in the output SNR can also be expected due to these few centimeter positional errors. These results are incorporated into the SII metric to explore the intelligibility loss for different arrays through various simulations and experiments with different masker sources.

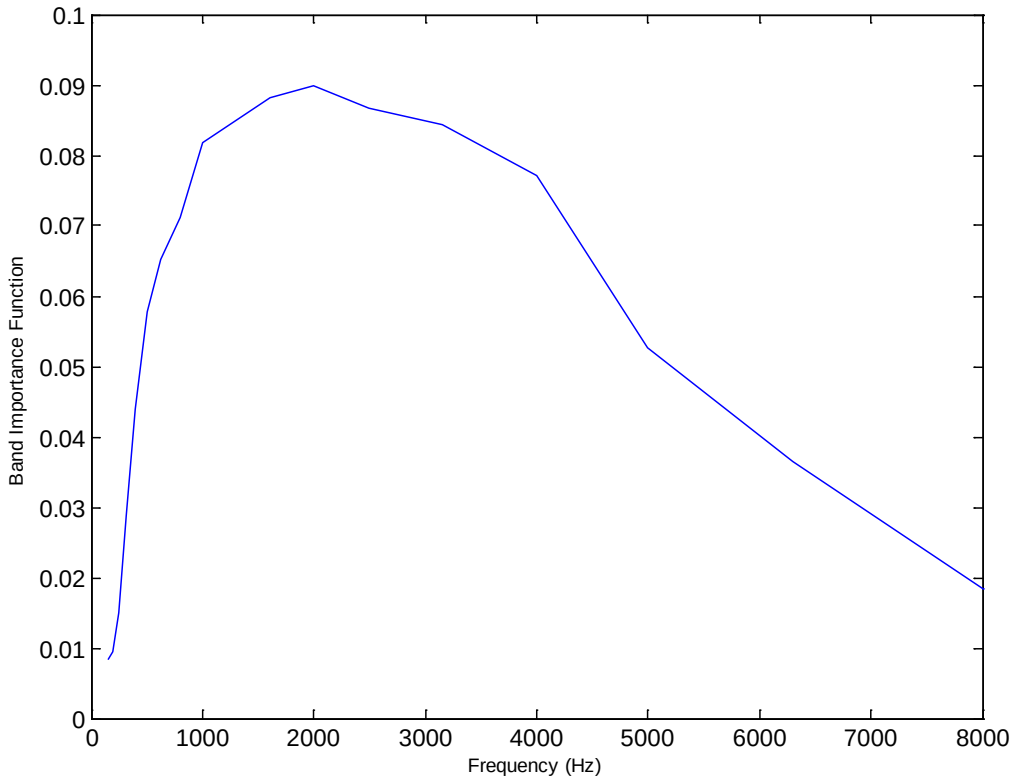


Figure 6: Band importance functions for an average speech according to One-third Octave band method [12]

The experiments and simulations are performed based on the discussions from this chapter to propose tolerable limits on positional errors over intelligibility loss, for a variety of array configurations. The simulator design, experimental setup used for data collection and the design constraints and parameters involved during the experiments and simulations are discussed in the following chapters.

CHAPTER 3

SIMULATOR DESIGN

The implementation of the simulator and the variables/parameters involved during the simulation are discussed in this chapter. The simulator is used to demonstrate the impact of positional errors on SII and propose tolerable limits on positional errors over intelligibility loss, for a variety of array configurations. It involves Monte Carlo runs using the single-channel speaker recordings with a complete control on the positions of the sound sources and microphones. The test signal types, array configurations and the procedural steps involved during the simulations are described in the initial sections of the chapter. The choice of random variable distributions and the effects of speech pauses on the performance metric and calculation of SNR are explained in the latter sections. Finally, the influence of array geometries and input SNR on intelligibility metrics is discussed using the results obtained from the simulations.

3.1 Test signal sources

The single-microphone low-noise recordings were made to serve as the sources for highly flexible simulations of array recordings where the source position can be controlled. These single-speaker single-microphone speech recordings were made using a single omnidirectional measurement microphone (EMC8000, BEHRINGER International GmbH) in a relatively quiet office environment at the Audio Systems lab facility in the Center for Visualization and Virtual Environments (CVVE), University of Kentucky. The speaker was approximately 0.23 to 0.46 meters from the microphone and acoustic treatments (foam) were placed behind the microphones on two sides to limit reverberations. Goldwave [43] was used to reduce the low level noise and room modes through post-filtering performance using an spectral envelope noise reduction algorithm. The recordings were saved as a 16-bit mono wave file sampled at 44.1 kHz for about 20 seconds. The speakers were selected from the native English speaking students, staff and professors from CVVE. The speakers were asked to read a script dominated by words

used in intelligibility studies with children. These wave files were then used to simulate speech recording over a microphone array during simulation analyses.

3.2 Microphone array configurations

The simulations in this work consider 4 microphone spatial distributions: linear, planar, perimeter and a non-uniform spread over 3 dimensions as shown in Fig. 7. Microphones were equally spaced for the linear and perimeter distributions, whereas an irregular spacing of microphones formed the planar and non-uniform 3-D array distributions. The microphones are simulated to scan a Field of view (FOV) which defines the spatial limits for the focal point of the beamformer. For all simulations described in this section the dimensions of the FOV were: 3.6m for both length and width and 2.2m for the height.

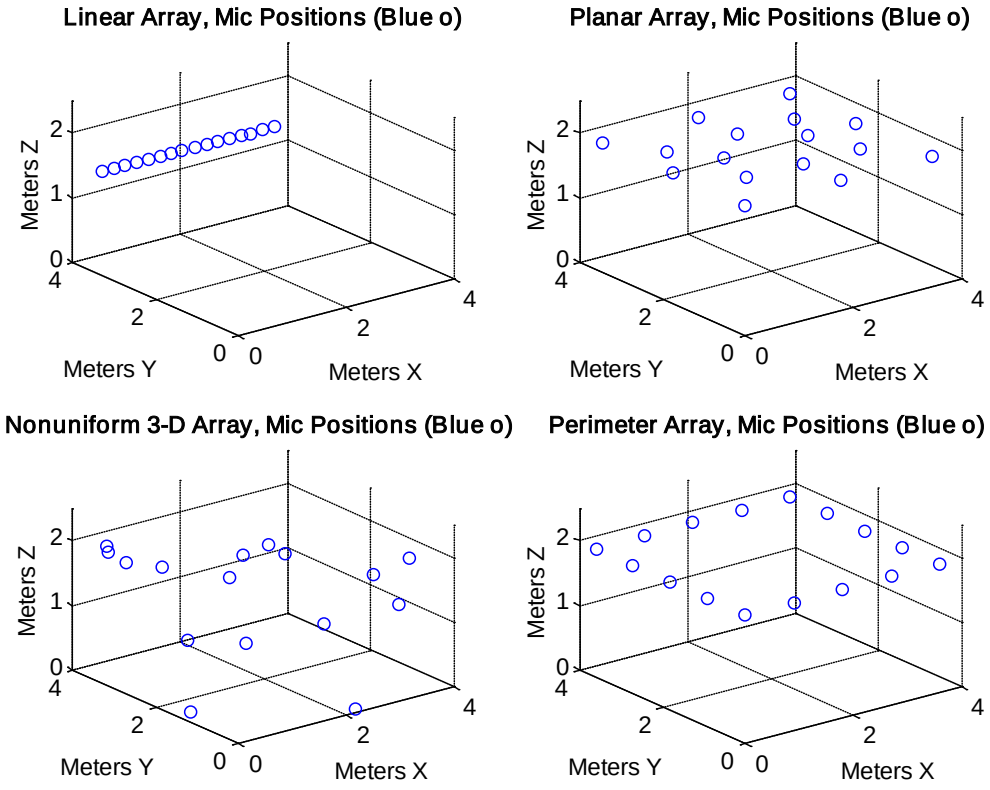


Figure 7: Microphone array distributions

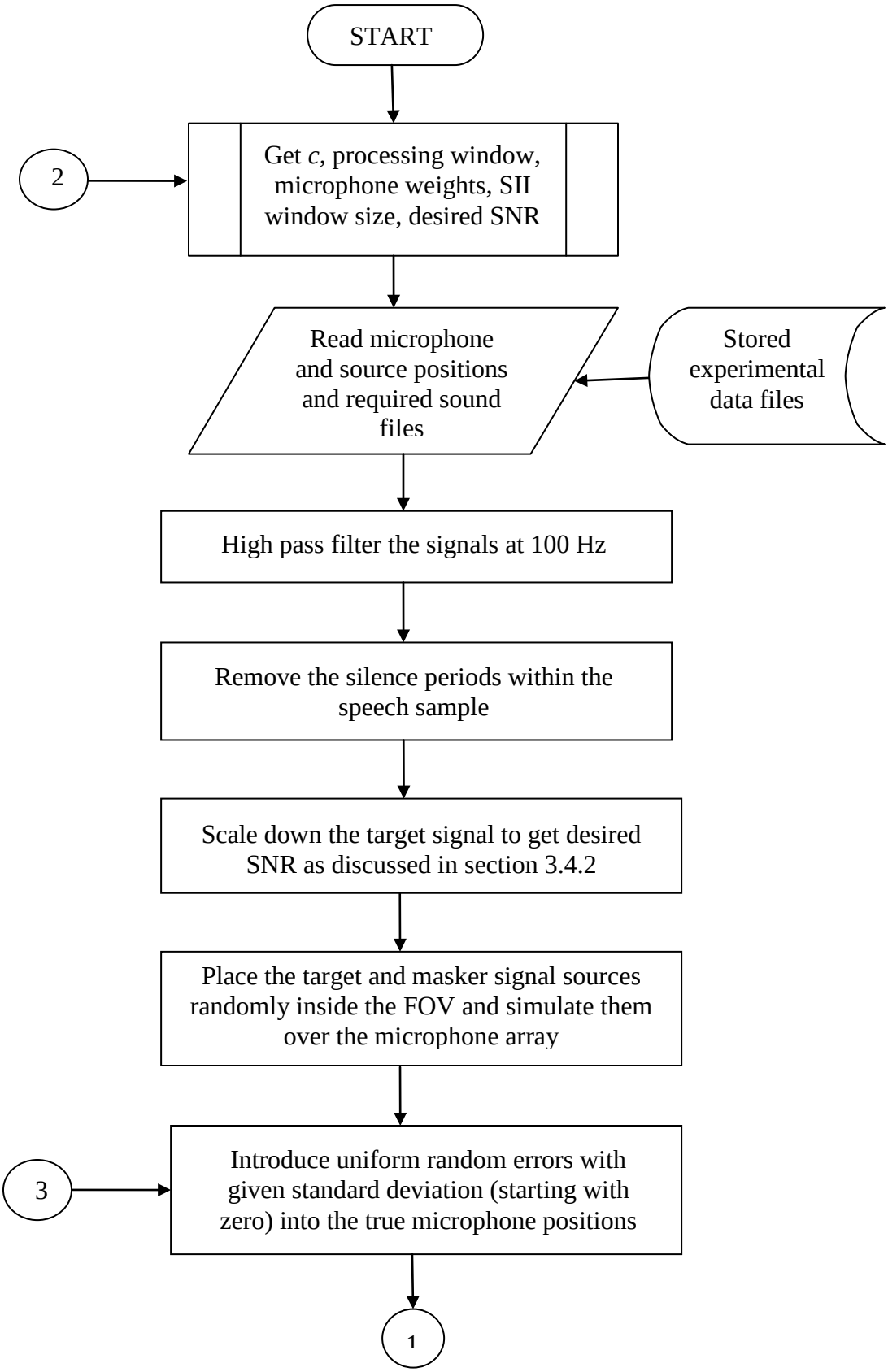
The linear array consists of 16 microphones along a plane parallel to the edge of the FOV and located 1.5m above the floor. Microphones were equally spaced at 0.21m. The perimeter array consisted of 16 microphones symmetrically distributed above the FOV forming the vertices of a square with an equal spacing of 0.9 m between the microphones. The planar array consisted of 16 microphones arranged in a plane parallel to the floor. The microphones are placed along the vertices of two concentric rectangles in a plane 1.99 m above the floor. The non-uniform 3-D array irregularly places the 16 microphones within the FOV in a random manner. The target and noise sources always exist within the FOV for all the arrays. The exact microphone positions were predefined during the simulations and thus the simulator provides the advantage of controlling the positions of the microphones in the array. Table 4 gives the statistics of the microphone distribution geometry such as maximum distance between any 2 microphones, centroid of the array, average spacing between closest pairs and dispersion of the microphone array.

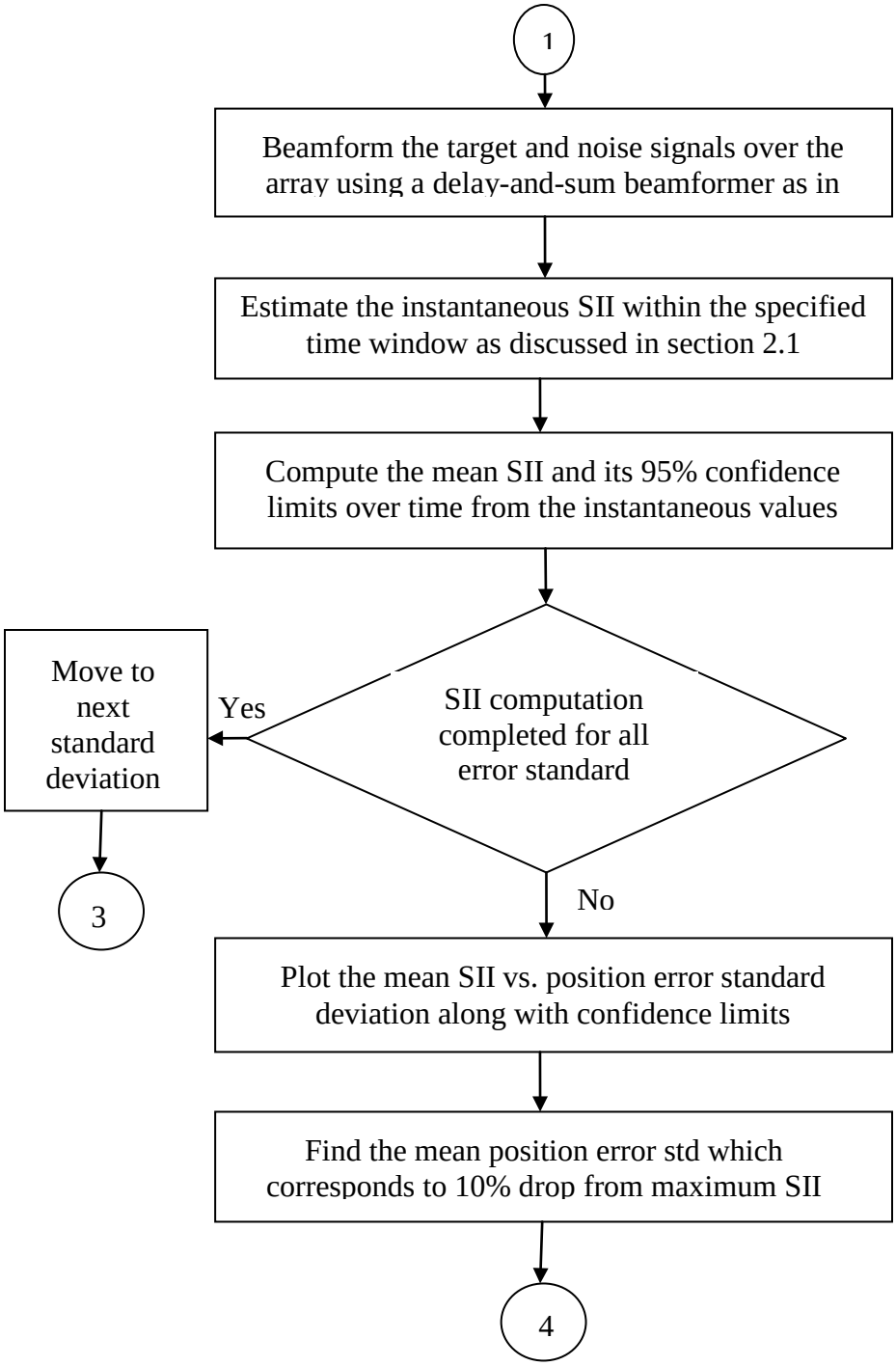
Table 4: Distributive Statistics of the Microphone arrays

	Linear Array	Planar Array	Perimeter Array	Non-uniform 3D Array
Maximum distance between any 2 mics	3.17 m	4.88 m	5.09 m	4.71 m
Centroid of the mic array	(1.75,3.48,1.50)	(1.79,1.80,1.99)	(1.80,1.80,1.99)	(1.66,1.86,1.40)
Average spacing between closest pairs	0.21 m	0.89 m	0.9 m	0.78 m
Dispersion of the mic array	(1.01,0.01,0)	(1.27,1.19,0)	(1.54,1.54,0)	(1.39,1.44,0.59)

3.3 Simulation flowchart

The simulation is described in the flowchart as shown below in Figure 8.


START
2
Get c, processing window, microphone weights, SII window size, desired SNR
Read microphone and source positions and required sound files
Stored experimental data files
High pass filter the signals at 100 Hz
Remove the silence periods within the speech sample
Scale down the target signal to get desired SNR as discussed in section 3.4.2
Place the target and masker signal sources randomly inside the FOV and simulate them over the microphone array
3
Introduce uniform random errors with given standard deviation (starting with zero) into the true microphone positions
1


1
Beamform the target and noise signals over the array using a delay-and-sum beamformer as in
Estimate the instantaneous SII within the specified time window as discussed in section 2.1
Compute the mean SII and its 95% confidence limits over time from the instantaneous values
SII computation completed for all error standard
Yes
Move to next standard deviation
3
No
Plot the mean SII vs. position error standard deviation along with confidence limits
Find the mean position error std which corresponds to 10% drop from maximum SII
4

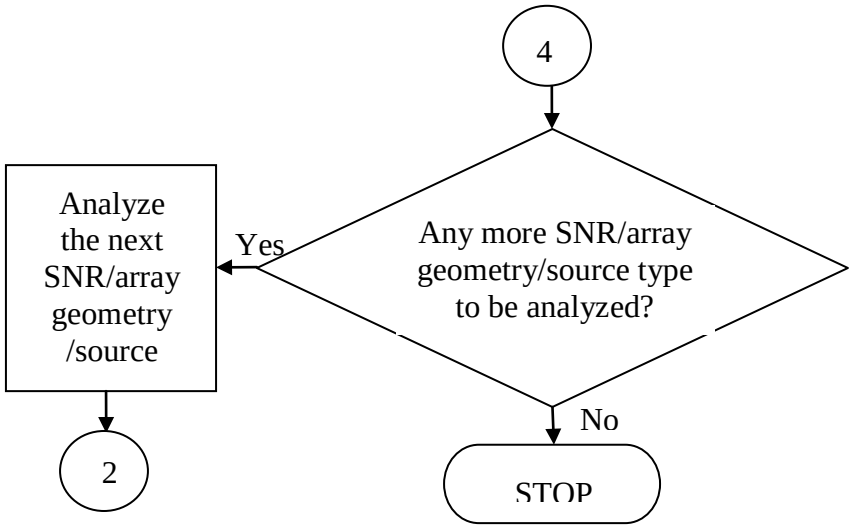


Figure 8: Flowchart implementation of the simulator

The simulations in this work consider 4 different microphone spatial distributions and make use of single-microphone speaker recordings as the test signals as discussed earlier. In order to calculate the SII for a given speech-in-noise condition, separate target and masker signals are needed [12]. Thus the simulator uses two source signals; one which is the focus of the beamformer and is the target signal and another one outside the focal point is the masker. The target signal is simulated under two masker conditions: interfering speech and white noise. The target signal is scaled and linearly added with the masker signal at desired SNR levels to get the total test data for the beamformer and SII estimation as discussed in the latter sections. Excess masking occurs when target and interferer are voices of same sex resulting in quite poor intelligibility. Thus, for the interfering speech signal, a voice from the opposite sex is always chosen. For example, a male speech signal is considered with a female interfering speech signal and vice versa. White Gaussian background noise was used as the masker for the second condition.

The simulation is performed with random placement of spatially separated source and masker signals inside the FOV. These signals are then simulated over the array of interest using the functions (simarraysig.m, delayt.m, roomimpres.m) from the Array Toolbox

[42]. The simulator details are given in [41] and the actual simulator is part of the Array toolbox [42]. The speed of sound was assumed to be 345 m/s with no reverberations and an air attenuation of 3.28e-5 dB per meter-Hz. A Monte Carlo simulation was performed with increasing random position errors. The position error was increased in the beamform algorithm by adding uniformly distributed random numbers to the *x*, *y*, and *z* coordinates of the measured positions. For each standard deviation, 25 independent position errors were simulated and beamforming computations were made. The target and masker signals were separately beamformed over the position of the target source, and are used to compute the SII for the given speech-in-noise condition. 95% confidence limits of the SII estimates were also estimated to represent the error bars in the results. The parameters/constraints involved during the simulation are explained below, which are also used during the experiments that are discussed in the next chapter.

3.4 Design parameters

(a) SII window

During the estimation of SII, the signals were partitioned into smaller time frames to account for the temporal variations in the target signal and background noise. The separate spectrum power levels (using Fast Fourier Transform, FFT) of the input speech and noise signals are computed within those time frames. This overlapping time window was slid over the signals and the instantaneous SII was estimated within each window. Therefore, the mean of these instantaneous SII values over time (with pauses and near silent periods of speech censored out) gives the estimate of SII for that particular speech-in-noise condition. The window length should be chosen small enough, on the order of several milliseconds (ms), to track the relevant variations of the signal over time. A longer time window results in a poorer grasp of temporal variations of the signal [14]. In the experiments and simulations given in this work, a 100 ms window with a 50% overlap was used during the estimation of SII.

(b) High pass filtering

The speech signals recorded using the microphones may include additive noise due to ambient conditions and low frequency room modes. So, the acquired signal is high-pass filtered at 100 Hz to eliminate the low frequency noises. The effect of this filtering is evident in Figure 9 which shows the filtered version of the raw signal along with the original recorded signal indicating the significant reduction in levels of background (room) noise. The room noise is specified as a steady state room noise, based on the statistics computed from the signal segment from the first few seconds of the signal as indicated in the Figure 9.

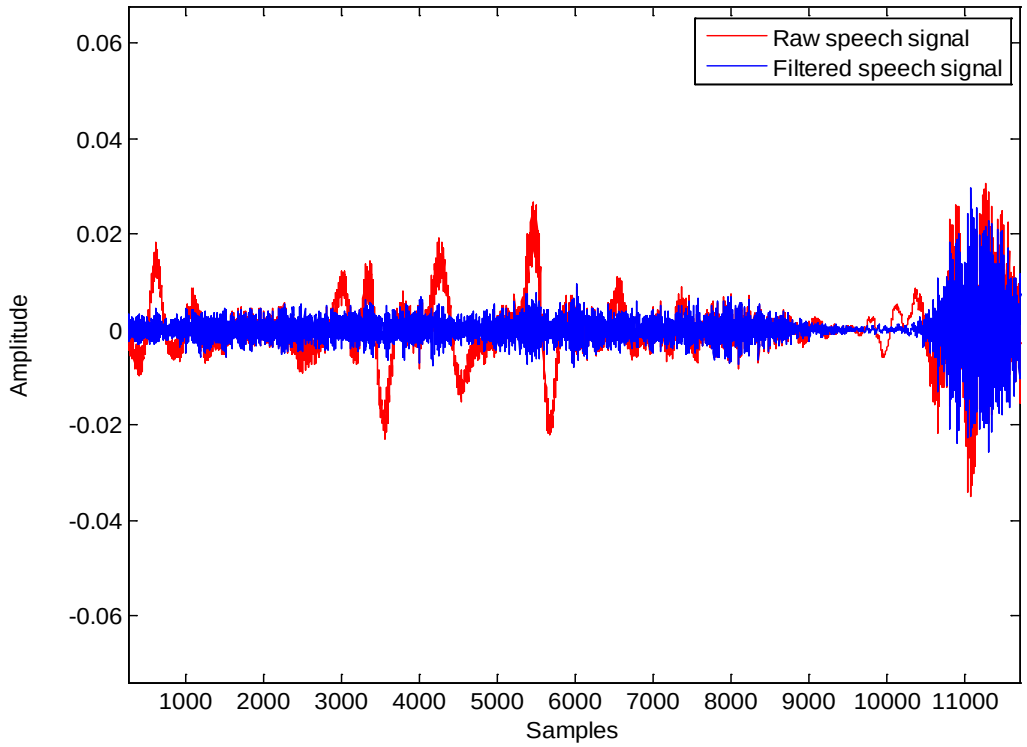


Figure 9: High pass filtered speech signal

(c) Microphone weights

A set of weights is determined for the microphones in the array based on their distances to the source. The computed weights are applied to the microphones before summing them together to produce the beamformer's output in both the experiments and

simulation. A weighting parameter γ is used to determine the set of weights, a_i, in order to deemphasize or emphasize the distant microphones as shown in Equation 26. The weights are given by:

$$a_i = \left(\frac{d_{min}}{d_i} \right)^{\gamma} \tag{26}$$

where d_i is the distance between the microphone and the source and d_{min} is the minimum of the distances (which corresponds to the closest microphone). All the microphones receive a uniform weight when γ is equal to zero. When γ is equal to 1, it results in an inverse distance weighting where the closest microphone is gets a weight of one. A larger positive γ gives more weight to the closer microphones whereas a negative γ gives more weight to the distant microphones. The weighting parameter γ is always considered to be 1 during the analyses in this work.

3.4.1 Random variable distributions

To investigate the errors in calibrating the microphone positions and its impact on beamformer's response and intelligibility, these positional errors are modeled using random variables. These random variables are generated such that they are uniformly distributed with zero mean and variance σ^2 and are centered on the true value. The probability density function (pdf) of the continuous uniform distribution is given as [44]

$$f(x) = \begin{cases} \dfrac{1}{x_1 - x_0} & \text{for } x_0 \le x \le x_1, \\ 0 & \text{for } x < x_0 \text{ or } x > x_1, \end{cases} \tag{27}$$

Suppose, if the random error with a limit *a* deviates on both sides of the true value uniformly in the range $\left[-\frac{a}{2}, \frac{a}{2} \right]$, then the standard deviation σ can be derived from pdf:

$$\sigma = \frac{a}{\sqrt{12}} \tag{28}$$

Uniform distribution is chosen over normal distribution for the random variables, as it generates a set of random errors for which all have an equal probability over the maximum deviations from the true position. Also, it has a constant probability density in the given limits of the error and zero probability density elsewhere. In a normal distribution, only 68% of the generated random errors are within one standard deviation away from the given mean (between mean minus 1 times standard deviation and the mean plus 1 times standard deviation) and a significant portion of the other errors are within two or three standard deviations of the mean. This may lead to discrepancies with actual positioning error, which is limited to the precision of the measurement technology used to place or locate the microphones. However, a Gaussian distribution can account for occasional large errors due to mis-measurement or typo. Since the focus of this work in on practical errors in the calibration or placement processing, a uniform distribution is used to model the positional errors.

A uniform distribution is more consistent with errors from lack of precision in the measurement process. For the sake of comparison, however, a normal distribution was considered to model the positional errors in microphone placements and the results were plotted along with uniformly distributed positional errors as shown in Fig. 10. A Monte Carlo simulation of 25 runs with random placement of signal and noise sources is carried out. The error bars correspond to 95% confidence limits of the SII estimates. The results using obtained from both the distributions are found to be close as shown in Fig. 10. There were minor differences between the graphs at certain errors but the difference is no larger than 2% drop in SII, and the uniform distribution falls above the normal distribution for most of the errors as expected.

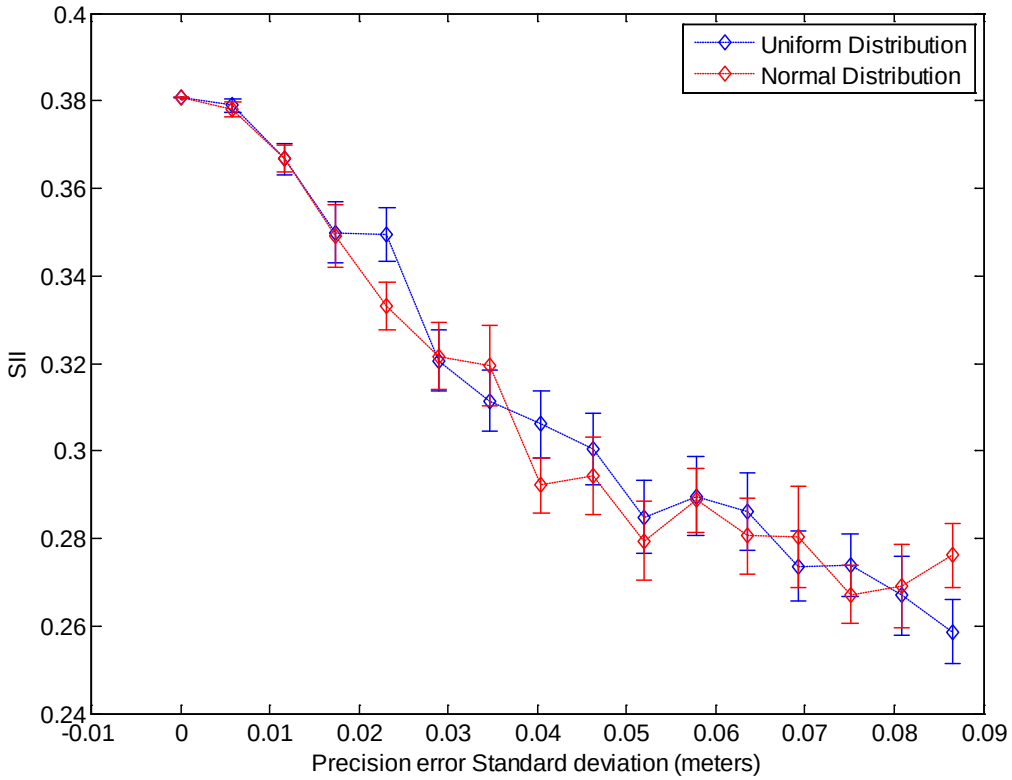


Figure 10: Microphone positional errors with Uniform distribution and Normal distribution

3.4.2 SNR calculations

Based on the Speech Reception Threshold (SRT) data [14] and various listening tests with the recorded data sets, an intelligibility rating of 0.3 or greater is required for a normal-hearing listener to recognize most words. Values below this result in a significant increase in the number of words rendered unintelligible. A critical feature for a beamforming application is to improve the SII for the barely intelligible speech, rather than improving the index for speech that is already quite intelligible. To examine the effect of precision errors under these conditions, signals from speaker of interest and masker signals were combined at an SNR to result in an unintelligible signal over all the microphone channels (SII < 0.3) and beamformed over the array to yield a better intelligible signal.

In order to achieve desired input SNR, the target signal is scaled up or down using a *scaling factor*(α). The *scaling factor* is derived using the desired SNR and RMS value of the target signal and noise. The RMS value is computed for the received signal at each microphone and is averaged over all channels, assuming that the DC component is removed. Consider $x_{m,i}[n]$ to be the target signal from a source located at $\vec{r}_i$, received by a microphone 'm' located at $\vec{r}_m$. Then, the RMS value of the signal computed over *N* samples is determined using the equation:

$$x_{rms} = \sqrt[2]{\frac{1}{N}\sum_{n=0}^{N-1}\left(\left|x_{m,i}[n]\right|\right)^2} \tag{29}$$

Similarly, for an interfering noise signal $y_{m,j}[n]$ located at $\vec{r}_j$, the RMS value of the noise over *N* samples can be computed:

$$y_{rms} = \sqrt[2]{\frac{1}{N}\sum_{n=0}^{N-1}\left(\left|y_{m,j}[n]\right|\right)^2} \tag{30}$$

Now, the *scaling factor* is derived using the equation

$$\alpha = \left[\frac{y_{rms}}{x_{rms}}\right]\left(10^{(snr/20\text{dB})}\right) \tag{31}$$

where α is the scaling factor, snr is the desired SNR (dB) and x_{rms} and y_{rms} are the mean RMS value of the target and noise signals averaged over all the microphones in the array. The initial test signal is a linear sum of the target signal (speaker of interest) and the noise signal. In order to achieve the desired SNR, the target signal in the total test signal is scaled using the *scaling factor* α as given in this equation:

$$s[n] = \alpha \cdot x_{m,i}[n] + y_{m,j}[n] \tag{32}$$

3.4.3 Periods of silence

The SII was designed so that it predicts the mean intelligibility of speech in noise rather than the intelligibility of individual words or phonemes. In any case, SII is badly defined in case of silent periods occurring within the normal speech. During these inherent pauses or near silent periods in the speech signal, the SII will always be zero regardless of the masking noise. As a result of this, the SII will never reach unity even when the target speech signal is presented at clear masking level. Moreover, the estimation of SII can be badly affected if one considers the silent periods occurring within normal speech signal. There might be large differences in the SII estimate from the actual intelligibility due to these silent periods between sentences which can vary between people [14]. Thus, for the results presented in this work, the periods of (near) silence occurring in the speech were removed for the SII computation for a more reliable and enhanced SII using the functions developed in Matlab.

The function removes the intervals of silence or pauses from a speech signal and filters it so that distortion (clicking from the concatenation of active speech segments) is reduced. Once these pauses are removed, the time length of the speech signal is reduced as shown in Fig. 11. The near silence periods are detected and removed using the envelope of the speech signal. The pauses are removed such that the speech and the noise signal remain synchronized in time. The envelope of the signal of interest is determined from the analytic signal computed using the Hilbert Transform. The discrete Hilbert Transform of an input signal $x_{m,i}[n]$ is given as [45, 46]

$$H_d\left\{x_{m,i}[n]\right\} = \frac{1}{\pi}\sum_{m=-\infty, m\neq n}^{\infty}\frac{x_{m,i}[m]}{n-m} \tag{33}$$

An analytic signal (complex time) can be constructed from a real-valued input signal $x_{m,i}[n]$ as its real part and it's Hilbert Transform $H_d[n]$ as its imaginary part:

$$x_a[n] = x_{m,i}[n] + jH_d[n] \tag{34}$$

The complex analytic signal can be expressed alternatively in terms of magnitude and phase as

$$x_a[n] = A[n]e^{j\phi[n]} \tag{35}$$

where is the $A[n]$ envelope of the signal and $\phi[n]$ is the phase of the signal. Then, the envelope of the signal $A[n]$ can be computed using the equation:

$$A[n] = \sqrt{x_{m,i}{}^2[n] + H_d{}^2[n]} \tag{36}$$

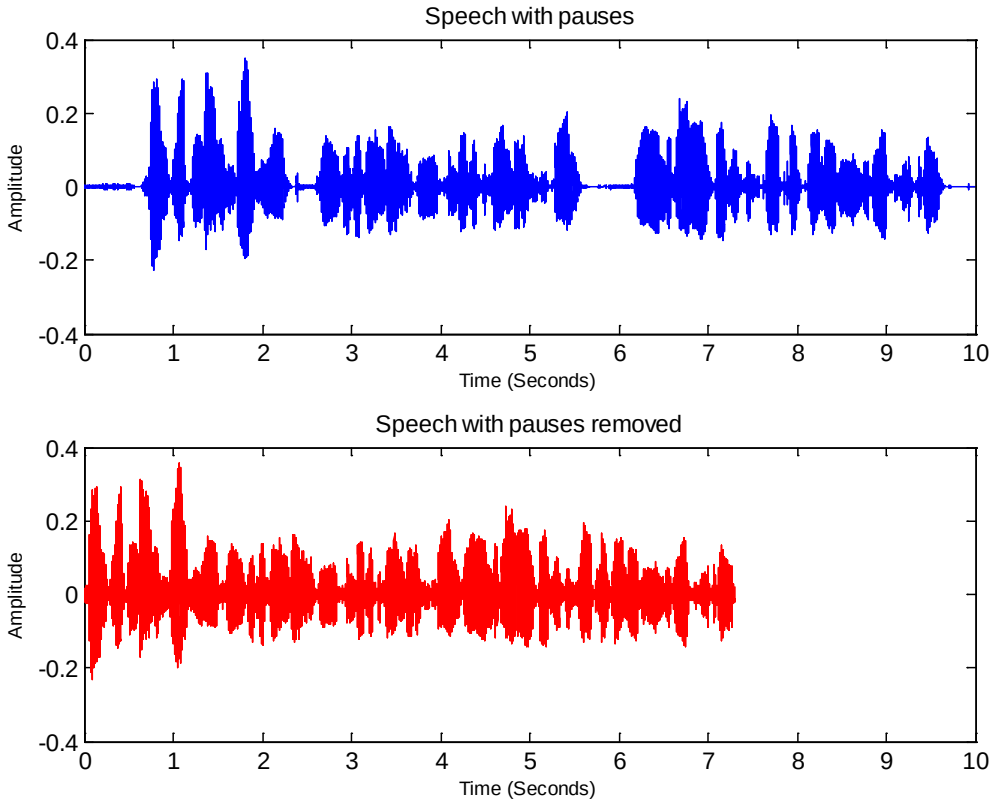


Figure 11: Original speech signal and speech signal (shortened in time) with periods of silence removed

The silence intervals from the speech are removed using an envelope threshold T which is a scaled function of the median of the envelope (usually one-fourth of a median). After computing the envelope, the speech samples of whose envelope magnitude values fall below the threshold are detected to be pauses and are removed. The resulting speech signal would be shortened in time due to the absence of silence intervals. Removing these pauses or near silence periods have improved the SII estimates and reduced its variations over time as shown in Fig. 12. The figure indicates that for the speech with pauses, the SII nears zero whenever a pause occurs in between the speech. In case of the speech with pauses removed, the SII reaches only a minimum of 0.25, and this makes the SII more consistent when averaged over time. From Fig. 12, the mean and standard deviation of SII for the original speech (with pauses) is 0.45 and 0.23. In case of the speech with pauses removed, the mean of SII is 0.57 and its standard deviation is 0.14. Therefore, removing the silence periods within the speech leads to a more reliable and enhanced SII estimate.

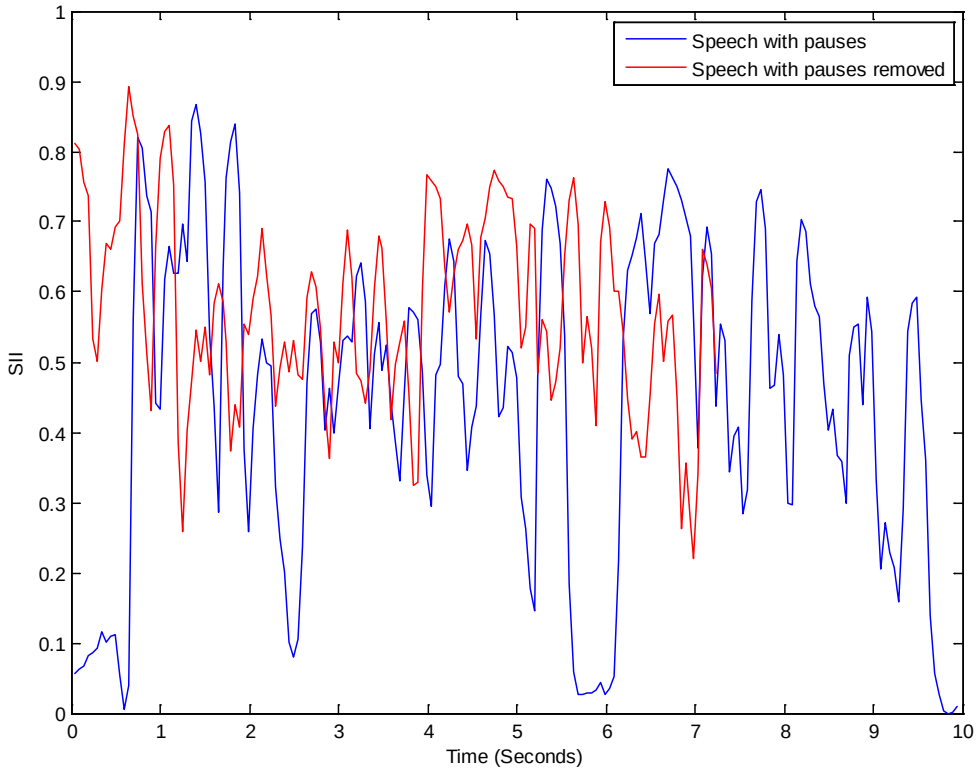


Figure 12: SII estimates for speech signal with pauses and speech signal with pauses or near silence periods removed

3.5 SII loss from positional errors

Equation 24 derives a sinc function relationship in the expected power loss from the standard deviation of the positional error as a function of wavelength. It predicts that the beamformer offers no enhancement to the target signal, for error standard deviations greater than one quarter wavelength due to the effective incoherent summation of the frequencies, beyond this quarter wavelength limit. The simulation analysis in Chapter 2 shows that the positional errors with standard deviations greater than 1.5 cm results in significant power losses, in the frequency range of 1500-5000Hz. These power losses are expected to result in SII degradation for the beamformed speech in noise. Hence, the impact of the positional errors on SII is examined for all the four different arrays using the simulator in this section.

The simulations are performed for different array configurations with a constant source and noise positions. A target speech signal (a male speaker single-microphone recording) is located centrally and 1.5m above the floor within the FOV. A female speaker single-microphone recording is used as the interfering noise and is placed at a height of 1.1m from the floor level and 1m diagonal to the left of target signal. These signals are simulated over the given array using the functions from the Array Toolbox. An input SNR of -12dB is maintained for all the four arrays such that a quite-intelligible speech (approx. SII = 0.35) is obtained. The speed of sound and air attenuation was assumed as discussed previously. Figure 13 plots the SII with increasing precision errors for all the four arrays. The error bars correspond to 95% confidence limits of the SII estimates.

From Fig. 13, the SII at zero error precision standard deviation for the different arrays can be compared. The maximum SII for the given arrays seem to be fairly close, with a maximum of roughly 5% difference in SII between the planar and perimeter arrays. The linear array also looks to be in par with the other complex arrays in terms of maximum SII. The different array configurations do not seem influence the maximum intelligibility at zero precision error in a great manner.

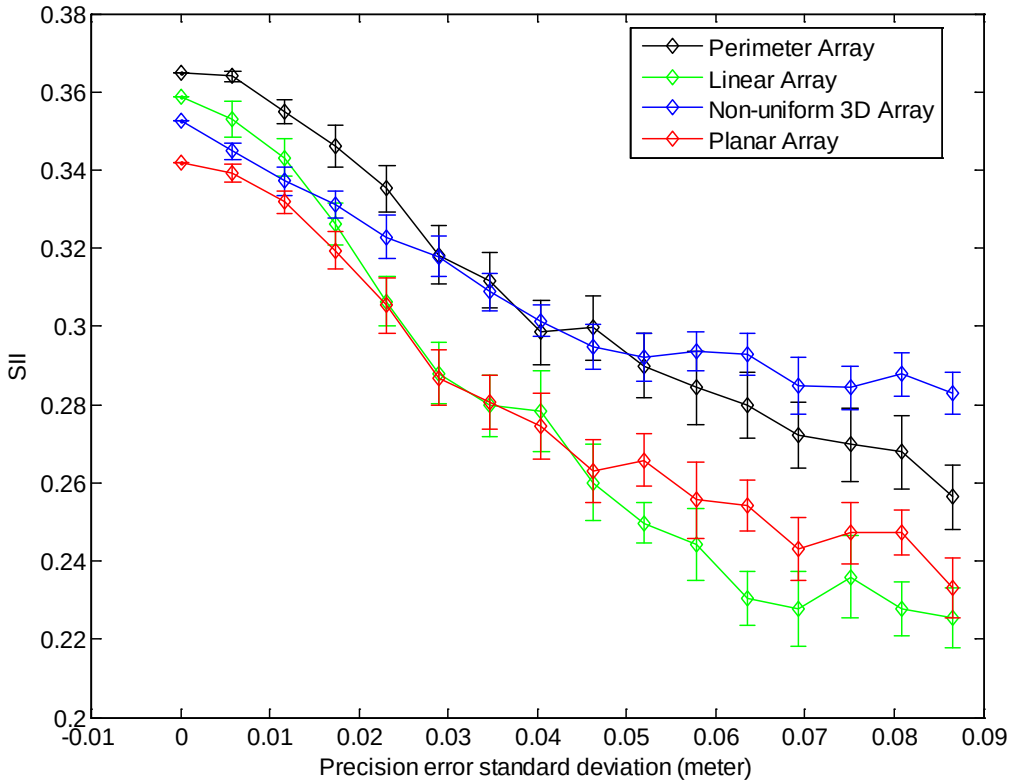


Figure 13: Impact of microphone positional errors on SII for different array distributions

Linear array drops rapidly when compared to other complex arrays. From Fig. 13, the drop in SII is more rapid for smaller errors for a linear and planar array in contrast to perimeter and non-uniform 3D arrays. An error standard deviation of around 3 cm yields a SII of 0.28 in case of linear and planar arrays whereas the perimeter and non-uniform 3D arrays has a SII of 0.32 (almost a difference of 10%). As the error increases, the SII continue to plunge for a linear array and fall behind the other arrays. This indicates that the linear arrays are more vulnerable to positional errors than other distributed arrays. The SII drop is slower for change in precision errors in case of a distributed 3D array. For a 3D array, only a 20% drop from maximum SII (zero precision error) is noted for a positional error standard deviation of 8 cm, which is considerably less compared to other arrays. Even at larger errors, the non-uniform 3D array maintains a SII of around 0.3, which results in a quite intelligible speech. Therefore, it suggests that the non-uniform 3D array is the more robust to precision errors than all the other arrays. From Fig. 13, it can

also be noted that for all the array designs, a similar trend can be seen where the SII drops steeply for initial errors and tend to settle down flat for error standard deviations greater than 4 cm, which corresponds to losing frequencies of 2000 Hz and above as predicted by Equation 24.

This work mainly focuses the improvements in SII for a barely intelligible speech (roughly one-third of the speech information is available for the listener) as discussed in section 3.4.2. An input SNR (with a help of a scaling factor) is used to scale the target signal such that the beamformed speech is quite intelligible. Figure 14 plots the SII as a function of input SNR, using different simulated arrays with a target male speaker and a female masker as discussed earlier in this section. No precision error is introduced to the microphone positions and it can be seen that SII approaches zero when the masking noise gets stronger and reaches close to unity as the target signal becomes stronger. Making the target signal stronger than the masker noise will increase its intelligibility and is well-predicted by the SII as shown in this figure. Various listening tests are also performed to see whether the actual intelligibility follows this behavior.

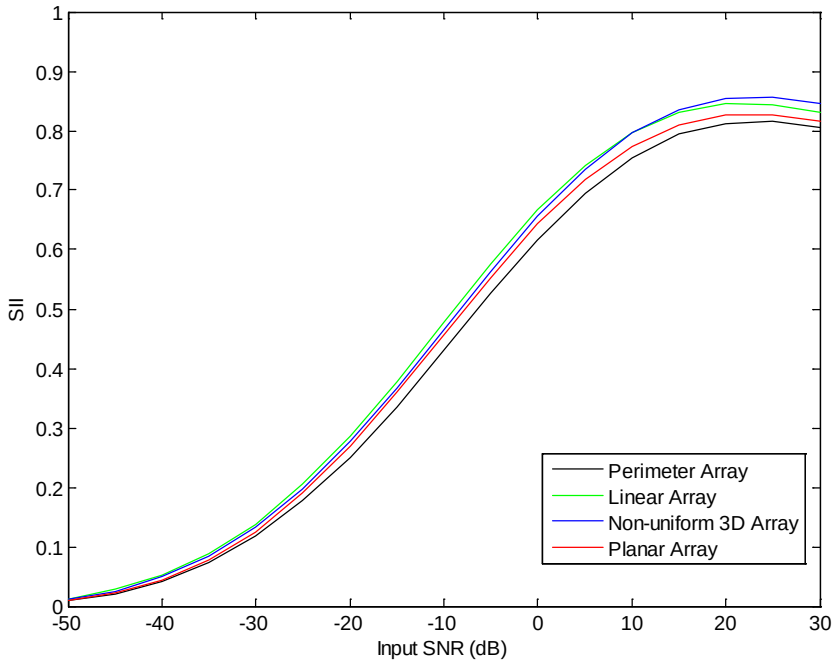


Figure 14: Input SNR vs. SII for different array distributions (interfering speech background)

For an interfering speech background, even at very low SNRs, there is still some speech information available to the listener and the SII exceeds zero. Increasing the SNR causes the SII to increase almost linearly until a 10 dB SNR is reached. The distortion and masking factors in SII restrict the SII from unity at higher speech levels. From Fig. 14, it can be noted that all the arrays demonstrate a similar behavior with respect to the changes in input SNR.

Figure 15 displays the SII as a function of input SNR for a male target speaker with a white noise background. With white noise background, no speech information is available at very low SNRs and the SII starts to deviate from zero as the SNR reaches a value of -30 dB. It increases almost linearly with SNR up to a value of 20 dB. Again at higher speech levels, the distortion factor causes the SII to level off, preventing it from reaching unity.

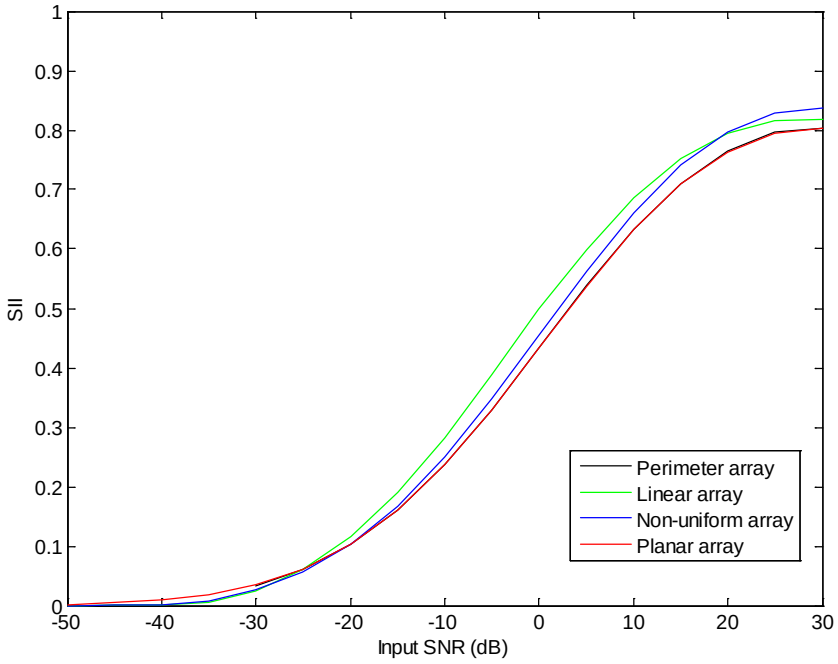


Figure 15: Input SNR vs. SII for different array distributions (white noise background)

From the figures 14 and 15, it can be observed that a quite intelligible speech (approximate SII value of 0.35) can be obtained at an input SNR of around -15 dB for an interfering speech background and -5 dB for a white noise background, for a variety of array distributions. These values seem to waver a little with the change in test sources and their positions inside the array. So, it would be more useful if a SNR range is specified to obtain a quite intelligible speech after beamforming. In order to achieve a quite intelligible speech with an interfering speech background, the input SNR has to be in the critical range of -20 dB to -10 dB. For white noise background, the critical range of SNR can be given as -10dB to -3dB.

In order to validate the simulation results, various experimental analyses were performed. The details of the experimental setup and data collection are described in the next chapter. It also discusses the simulation validation by comparing the results obtained from the simulator and experiments and presents tolerable limits on the positional errors.

CHAPTER 4

EXPERIMENT AND RESULTS

This chapter discusses the experimental setup, parameters involved during data acquisition and presents tolerable limits on the amount of positional errors for a variety of array configurations. The primary purpose of the experiments was to create conditions similar to the simulations and compare results to assess the validity of the simulation results. Sections 4.1 and 4.2 include the details about the test signals, test environment, hardware setup, and measurement of environmental parameters. Section 4.3 validates the simulator by comparing the experimental and simulation results. The latter section introduces tolerable limits on the positional errors using the results from the simulator.

In addition to the single-channel recordings made for the simulations, different multi-channel recordings using a single speaker were also made for the purpose of validating the simulator. More than one recording was made using the male and female speakers for the same array designs used during the simulations. The microphone configurations remained the same while recording the data for different speakers and the speaker was talking either standing or sitting at known position. Since a set of multi-channel single speech recordings were made separately with the same microphone geometries and different speakers, these recordings were linearly added with different power ratios to achieve desired SNR levels for the performance analyses as discussed in section 3.4.2. Note that both the simulator and the experiment use real speech data. The only difference is that multi-channel speech recordings were used during the experiments whereas single-channel speech recordings were simulated over multiple channels in the array in the simulator.

4.1 Test environment

The experimental room for collecting data sets was setup at the Audio Systems lab facility in CVVE. A 3.96 by 3.96 by 2.6 meters structure was constructed of aluminum

struts (80/20 Inc.-The Industrial Erector Set, Columbia City, Indiana) to mount the microphones in various geometries around the space of interest [42, 47]. This cage encloses the assumed FOV in which the speakers of interest were present as shown in Fig. 16. For all experiments described in this section the dimensions of the FOV were same as for the simulations: 3.6m for both length and width and 2.2m for the height.

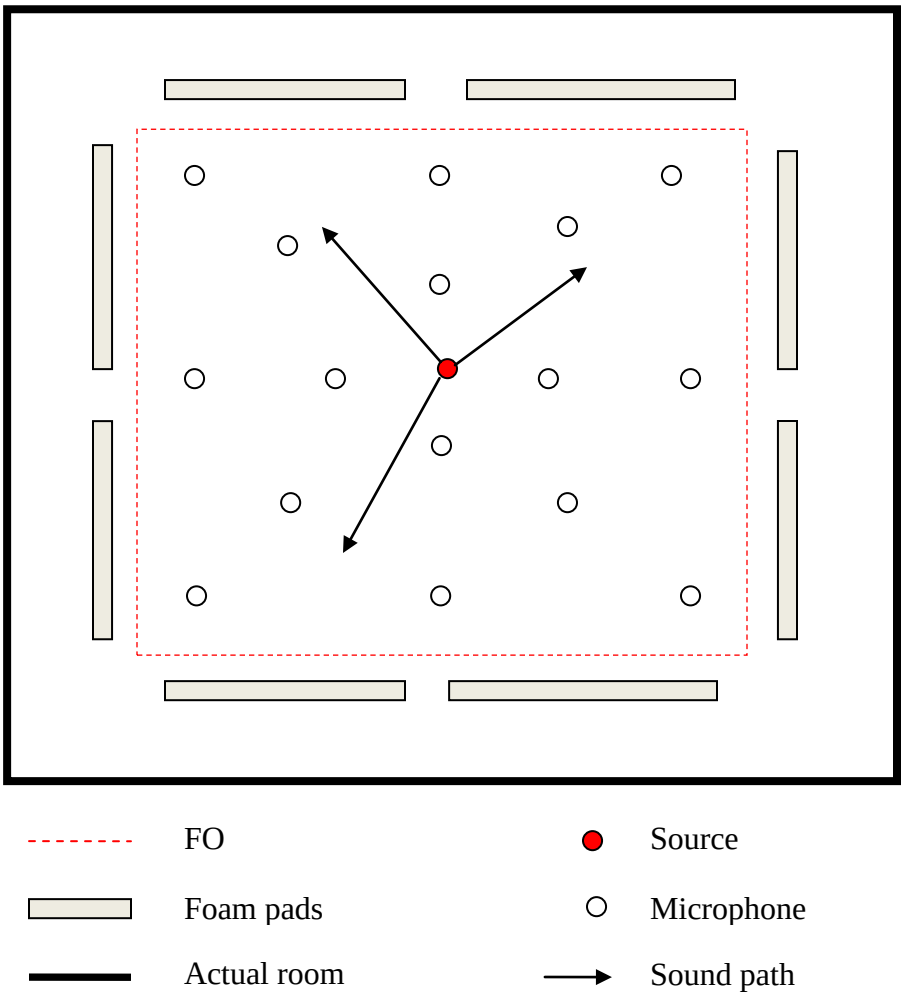


Figure 16: Test environment setup

The data collection and processing was driven by two AMD dual-core computers running Ubuntu Linux. A low latency audio server called Jack was installed in these machines and was used to record the data over multiple audio channels. A total of 16 omni-directional microphones (EMC8000, BEHRINGER International GmbH) were used

during the data collection. Each microphone was connected to an M-Audio Audio Buddy preamp, and digitized using two Delta 1010 cards by M-Audio which together support 16 analog input channels as shown in Fig. 17. Also acoustic treatments were placed behind the microphones to reduce noise, room modes and reverberation as shown in Fig. 18. In this case, three acoustic 0.03 meter foam pads (Auralex MAX-WALL 420) were set up to reduce room modes and ambient noises due to the computers, vents and traffic through the window. All the data recordings were done at a sampling rate of 22.05 kHz and were down sampled offline to 16 kHz in some cases for analysis.

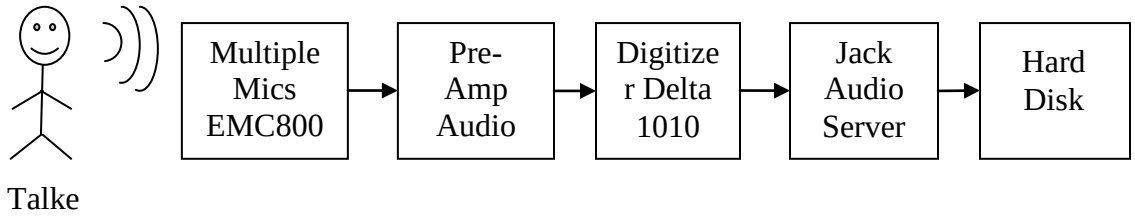


Figure 17: Data collection setup

4.2 Measurement of Environmental and Speaker parameters

(a) Microphone positions

The microphones were arranged in fixed geometries such as linear, planar etc., around the audio cage for each data capture experiment. The positions and configurations of the microphones remain the same for both the simulations and experiment. The microphone positions in the array were measured and verified using a laser measuring device (Leica DISTO A6). The microphone locations in 3D space were estimated by the triangulation method. Three reference points R1, R2 and R3 whose coordinates are fixed, at the corners of the audio cage. Then, the distance of the microphone from these reference points were measured using a measuring tape and laser beam. The actual positions of each microphone (*x, y* and *z* coordinates) were computed from these data. Care was taken to calibrate the microphones and pre-amplifiers for a constant gain over the array channels.

(b) Speed of sound

The speed of sound was estimated every time prior to the each data capture using the measured delay of arrival between 2 microphones for the sound from a predetermined source location with a source located co-linearly. A white Gaussian noise burst was used as the source for 25 seconds to enhance the correlation statistics between the 2 microphone signals. Once the cross-correlation peak corresponding to the time delay between the 2 microphones was estimated, it was used in the following equation to compute the velocity c_k for k = 1, 2, …, 25 time windows of 1 second duration:

$$c_k = \frac{\tau_k}{d_k} \tag{37}$$

where d_k is the microphone pair distance in meters and τ_k is the time delay between the microphones estimated through the cross correlation of the signals received at the microphones in seconds. The c_k values corresponding to a correlation magnitude of less than 0.4 were not used in the estimation. Of the remaining, the most repeated value of c_k is selected as the velocity of sound.

(c) Reverberation time

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. A white Gaussian noise burst was used to measure the RT_{60} time for the experimental environments. To get accurate RT_{60} value, the room was excited with the white noise and is played long enough so that the diffused sound reaches a steady state in the room. The source (loud speaker) is placed greater than 2 meters away from the microphones so that the direct path does not dominate the recording. Then the white noise source was abruptly stopped and the recording was continued for few more seconds capture the reverberating sounds as they fell below the noise floor. The signal

power plotted on the log scale falls linearly and the slope of this roll off was estimated with a censored least-squares line fit from the time right after the source was stopped to the time that it fell below the noise floor. The roll-off of sound from the room reverberation is found based on these two estimates. The slope of the roll-off is estimate in dB per second and used to extrapolate the amount of time it would take for the sound to fall 60dB from its maximum. This time was used as the RT_{60} time.

(d) Sound source location

During the experiments, the position of the target speaker inside the array was estimated using the SRP PHAT-β algorithm [24, 41]. An approximate position of the speaker's mouth was initially measured using the laser device just as the microphone positions. A Steered Response Coherent Power (SRCP) algorithm [24] was then applied in a 0.4 m neighborhood around that measured point to estimate the sequence of positions of the speaker for every 20 milliseconds. The SRCP was then computed over a 3-D spatial grid of spatial points every 0.04 meters. A whitening parameter β which determines the level of spectral whitening of the signal in the phase transformation (PHAT-β) was set to 0.6 for preprocessing before the position estimates. For every 20 ms window in time, the maximum SRCP value in space was chosen to be the location of the speaker. Let P_{ijk} be the detected peaks with *i, j, k* being the x, y and z co-ordinates that corresponds to the source location. A secondary threshold was applied to reduce the effect of noise during the periods of silence. Any detection with coherent power less than 5% of the maximum value of the peaks was considered as absence of the sound source and is represented by 'NaN' (Not A Number). The magnitude of the peaks P_{ijk} might still be large due to the presence of background noise or error in source location. Hence, to smooth out the SRCP peaks, they are passed through a sliding median filter over time with a window length of 21 samples to improve the estimate of the sound source location.

4.3 Simulator validation

For the sake of validating the simulator, experiments were performed in which two or three individual speech recordings were made using actual 16 channel microphone distributions (same as simulations), with talkers (opposite sex) at different positions but recorded separately. The test data for the beamformer and SII estimation was created at the desired SNR by scaling and adding these recordings as discussed earlier. One of the speakers was the focus of the beamformer and another outside the focal point was the interfering noise. The locations of the speaker in the experiment were then estimated using SRP PHAT-β algorithm for every 20ms as discussed in section 4.2. The microphone positions were measured using a laser meter and tape measure as described in 4.2.

While this measurement had an inherent precision limit (error), the position error was further increased in the beamform algorithm by adding uniformly distributed random numbers to the x, y, and z coordinates of the measured positions. This was done to see the impact of increasing position error and compare to a simulated array recording. Similar to the simulations, for each standard deviation, 25 independent position errors were simulated and beamforming computations were made (position of the speakers could not be varied in the experiment) from which SII was computed. To compare the experimental recording to the simulations, the closest microphone signals from the experiment were used in a simulator (where position error of both speaker and microphones was completely controlled). The array design and the positions of the target and noise sources remain the same for both the simulations and experiment. An approximate position was used for the target and noise sources during the simulations as the experiment estimates the positions for every 20 ms. The speed of sound and RT60 time were measured for each experiment and the values were unchanged during the corresponding simulations. Table 5 lists the various parameters that were maintained during the simulations and experiments for linear and non-uniform 3D arrays.

Table 5: Simulation and Experimental Parameters

Parameters	**Simulations**	**Experiments**
FOV	3.6x3.6x2.2 m	3.6x3.6x2.2 m
Speed of sound	347.5 m/s	347.5 m/s
RT 60	0.232	0.232
High-pass Filter Cutoff	100Hz	100Hz
Time window to estimate SII	100 ms	100 ms
Number of Microphones	16	16
Monte Carlo runs	25	25

The Monte Carlo runs were performed with increasing random precision errors and the results obtained from the simulator are compared to that of the experiments using a linear array in Fig. 18. Figure 18 indicates a similar SII drop as the precision error standard deviation increases for both the experiment and simulation. For the zero standard deviation error, the maximum SII from the experimental is 0.35, which is almost 10% less than that of simulation. A comparison of the SII value where the experimental data starts to the corresponding point on the simulation, suggests that an inherent precision error with standard deviation of around 2 cm was likely involved in the experiment, which was reasonable for the measurement system used.

The differences between the graphs are due to the accumulation of errors from the measurement, speed of sound estimate, and ambient room noise and reverberation present in the real data. Thus, for an interfering speech background, the 10% drop from maximum SII occurred at an error standard deviation of 0.6 cm for the experimental data, whereas for simulation it drops at 1.8 cm. However, both the graphs in Fig. 18 seem to follow a similar trend with a dramatic decrease during initial errors and settling to an almost flat SII variation for error standard deviations greater than 4cm, which corresponds to losing frequencies of 2000 Hz and above.

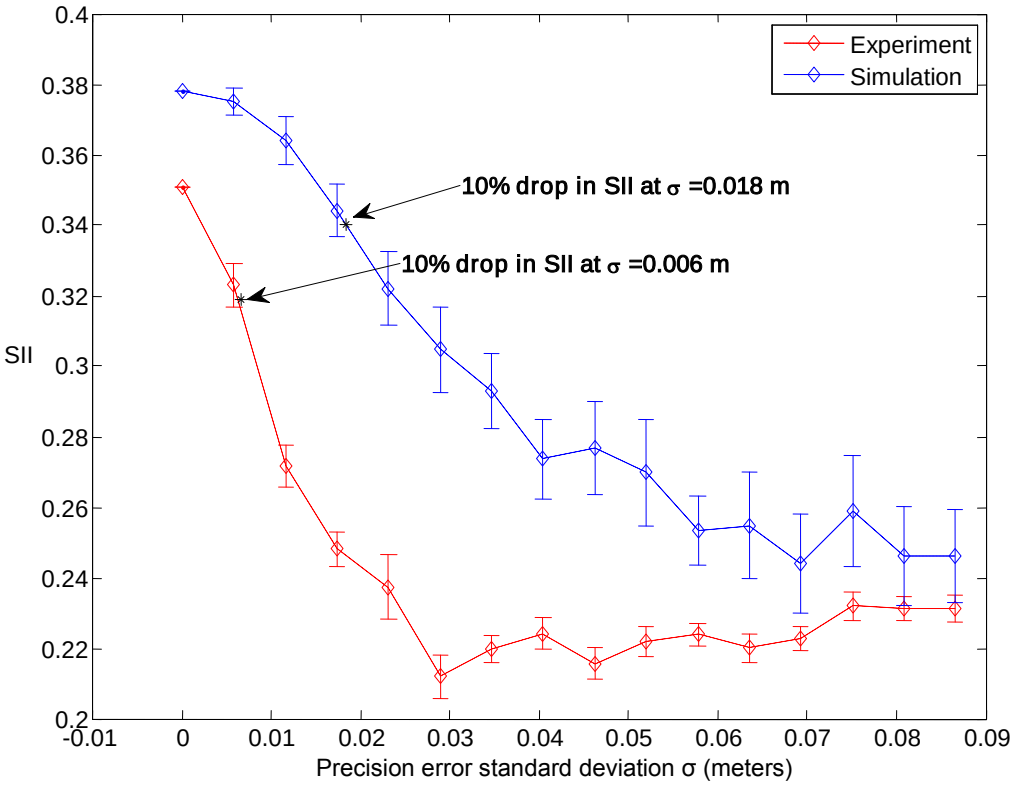


Figure 18: Comparison of experimental and simulation results for SII measures on beamformed signals with an interfering speech background as a function of precision error in microphone placement (linear array).

Similarly, the simulation and experimental results were compared for a non-uniform 3D grid array as shown in Fig. 19. The maximum SII for a zero standard deviation error is 0.25 during the experiment which is almost 30% less than that of simulations. The percentage drop in SII is much more compared to that of linear array which implies that the errors accumulated in simulation are larger than real data for the 3D array. Comparison of the graphs in Fig. 21 shows an inherent precision error standard deviation of around 5 cm is involved during the experiment. Similar to linear arrays, the trend followed by the two graphs in Fig. 19 are alike. This trend supports the case that as the error increases, the drop in SII settles down as the higher frequencies tend to merge as a result of coherence, making them insignificant as discussed in chapter 2. For a 3D grid array, a 10% drop from maximum SII in the experiments and simulations occurs at an error standard deviation of 2.4 cm and 1.83 cm respectively.

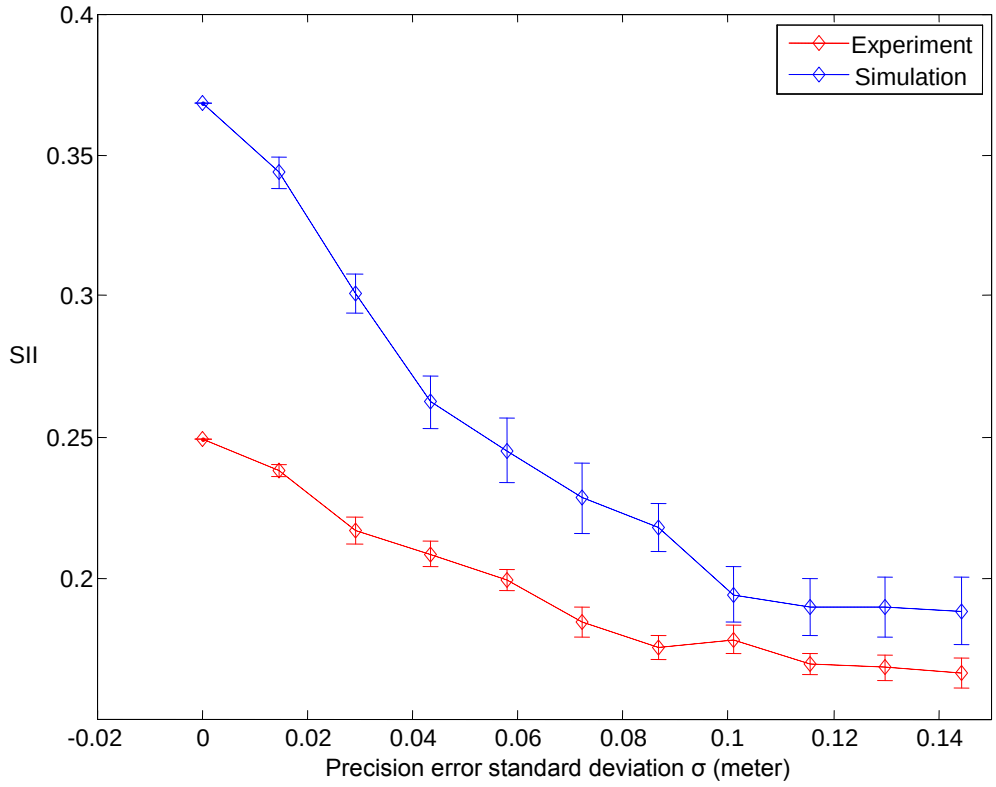


Figure 19: Comparison of experimental and simulation results for SII measures on beamformed signals with an interfering speech background as a function of precision error in microphone placement for a non-uniform 3D array

In addition, Figure 20 compares the simulation and experimental results for SII as a function of positional errors for a planar array. The trends of the graphs look similar to that of other arrays. By comparing the graphs in Fig. 20, an inherent precision error standard deviation of around 6 cm is expected to be involved during the experiment. The various errors accumulated during the experiment decrease the maximum experimental SII to 0.25 from a maximum SII of 0.38 that occurred during the simulations.

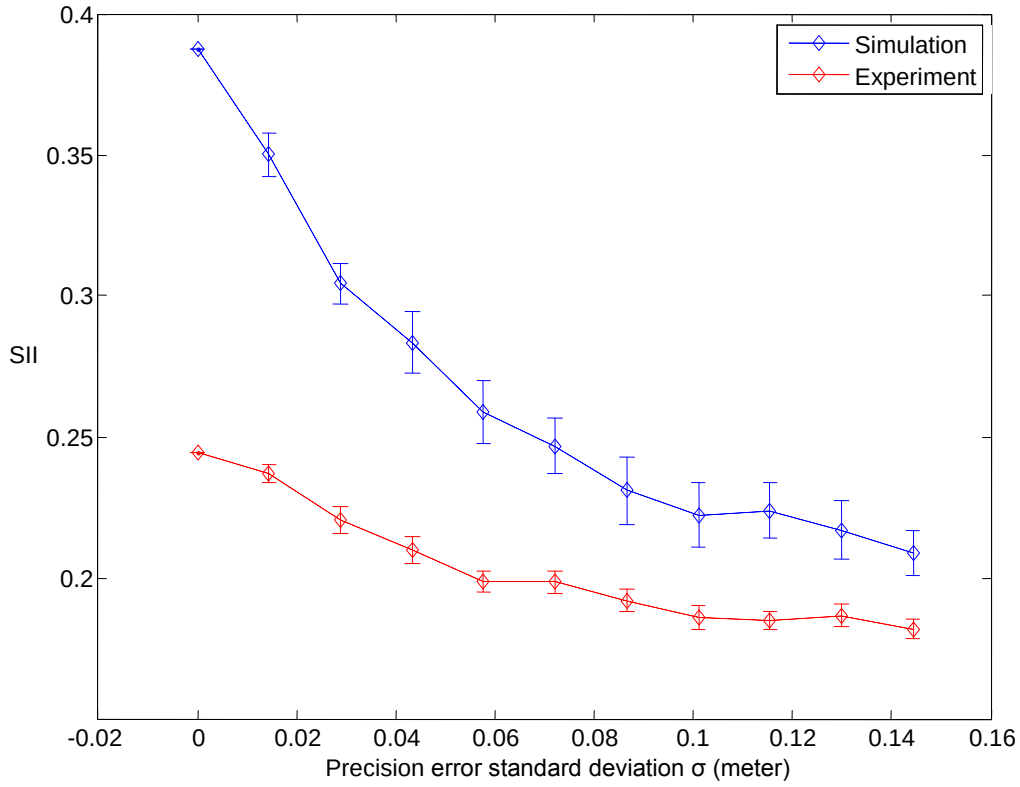
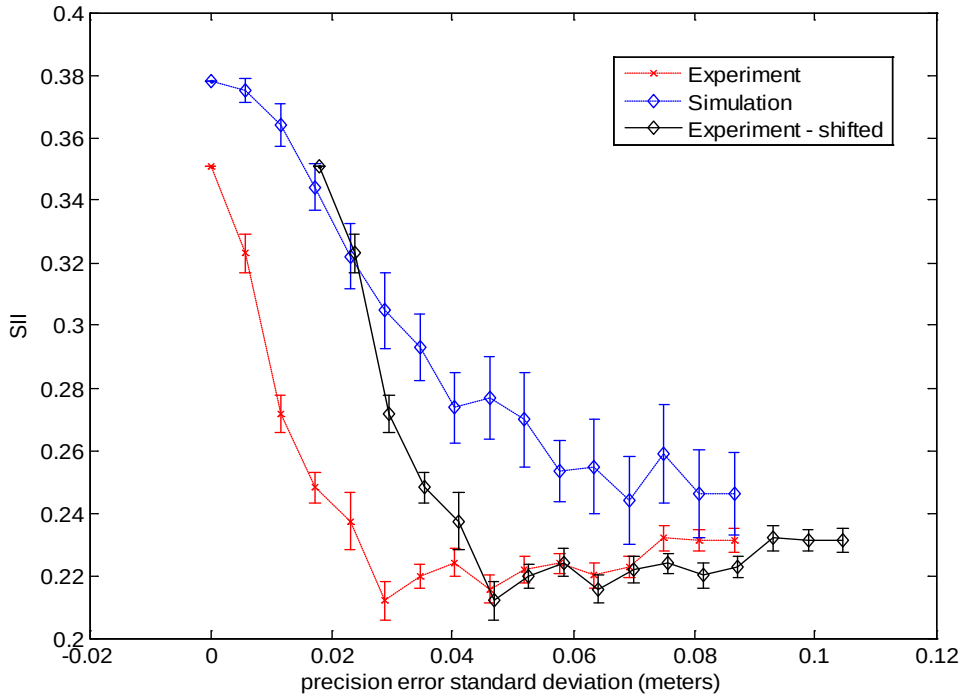
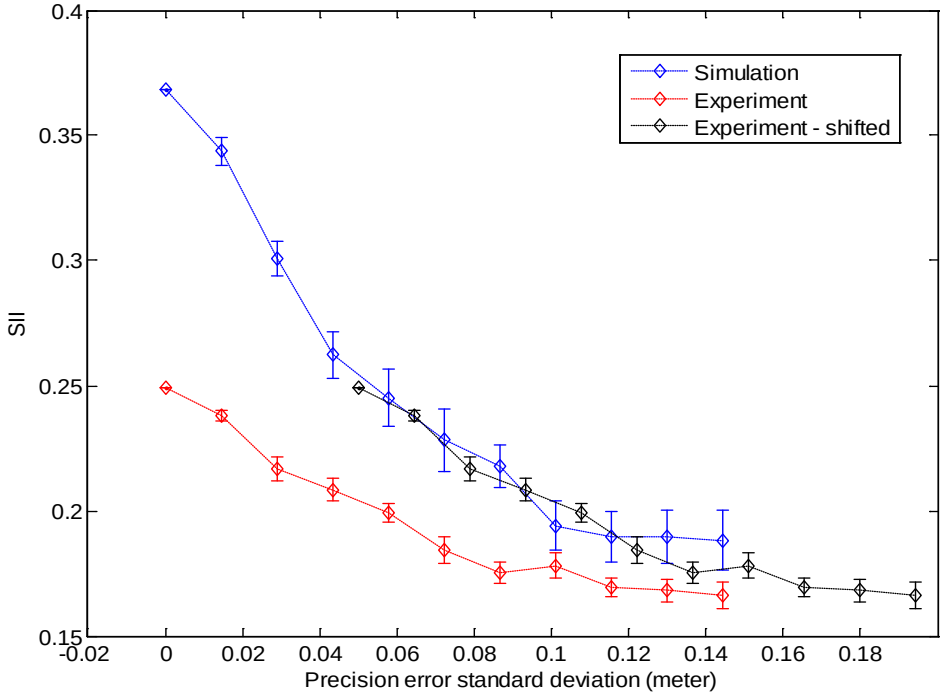


Figure 20: Comparison of experimental and simulation results for SII on beamformed signals with an interfering speech background as a function of precision error in microphone placement for a planar array

It can be observed from the above graphs that the experimental curves look very close in their performance to that of simulation curves, once the experimental precision is accounted, for various arrays. Figure 21(a) shifts the experimental data of the linear array horizontally to accommodate the inherent precision error of 2 cm and plots it along with the simulation data. It can be seen that the experimental and simulation curves tend to follow each other closely which supports the case of validating the simulator. The difference in the simulation and experimental graphs look much narrower in their performance, once the inherent precision error is accounted in the experimental data for a non-uniform 3D array as shown in Fig. 21(b).



(a)

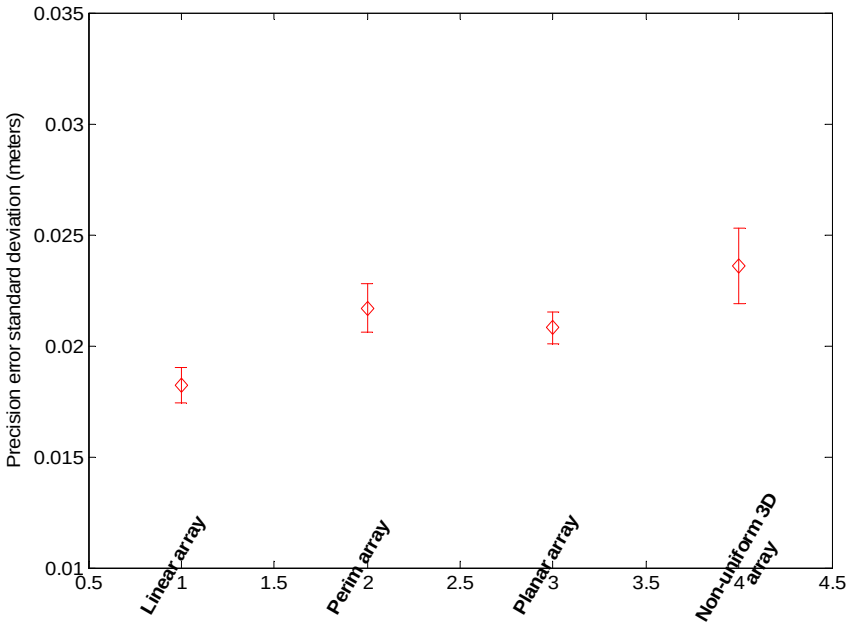


(b)

Figure 21: Comparison of simulation results with the shifted experimental results. For (a) Linear array (b) Non-uniform 3D array

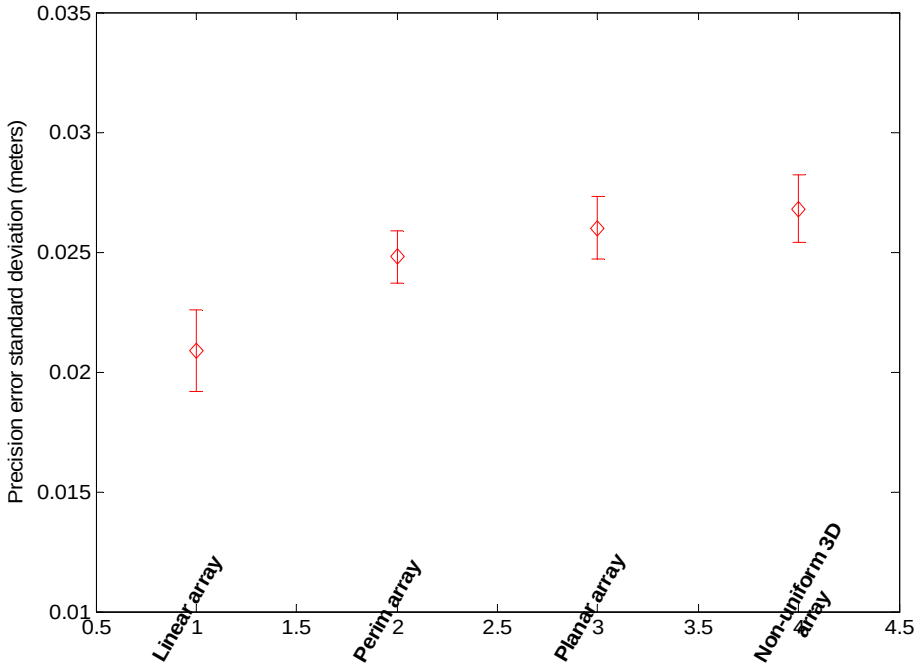
4.4 Tolerable limits on precision errors

To propose the tolerable limits on the errors, a Monte Carlo simulation of 50 runs was performed with random placement of target and masker within the FOV. Each run, in turn, included 25 independent position errors for each standard deviation as discussed previously. The precision error standard deviation which corresponds to a 10% drop from maximum SII was computed for each Monte Carlo run and is averaged over the total runs. The mean position error standard deviation at which a 10% drop in SII occurs was computed for all the four different arrays and shown in Fig. 22. The error bars correspond to 95% confidence limits. From Fig. 22a, for a male speaker, it can be seen that for an interfering speech background, a 10% drop in SII occurs somewhere between a standard deviation of 1.5 - 2.5cm, for different array distributions. For a white noise background, it occurs within a standard deviation of 2 – 3cm as shown in Fig. 22b. Figure 22 also suggests that the linear array seems to be more vulnerable to the precision errors than that of other more distributed and complex arrays.



(a)

Figure 22: Precision error standard deviation for which a 10% drop from maximum SII occurs under given masking conditions for a male speaker. For (a) Interfering speech (female) background (b) White noise background

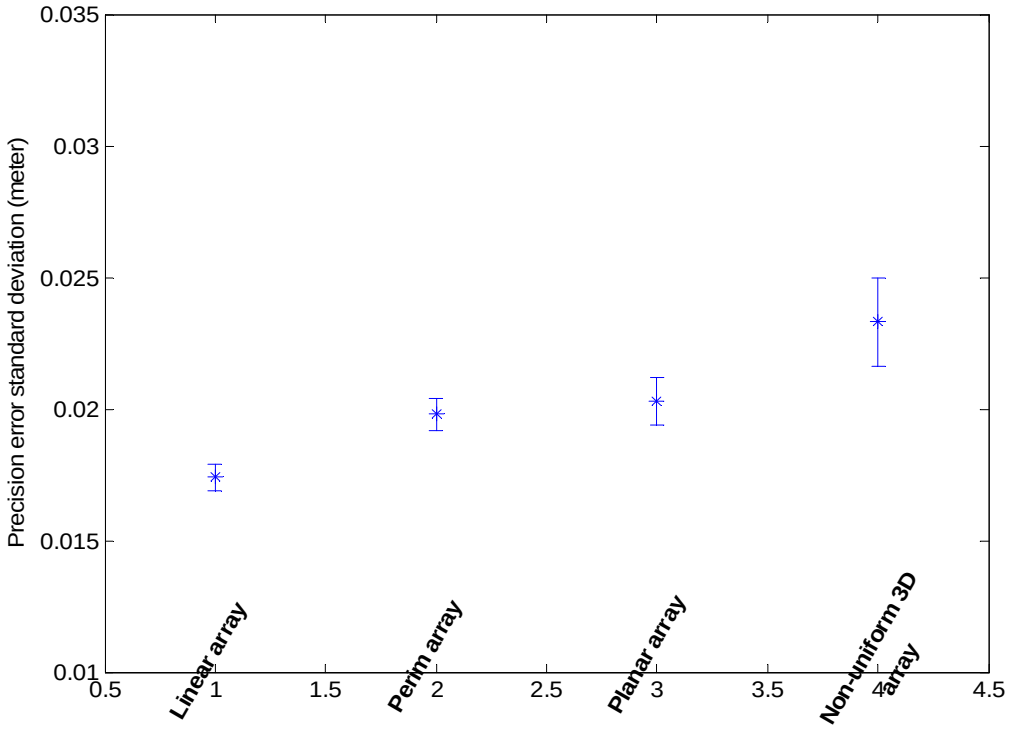


(b)

Figure 22, continued

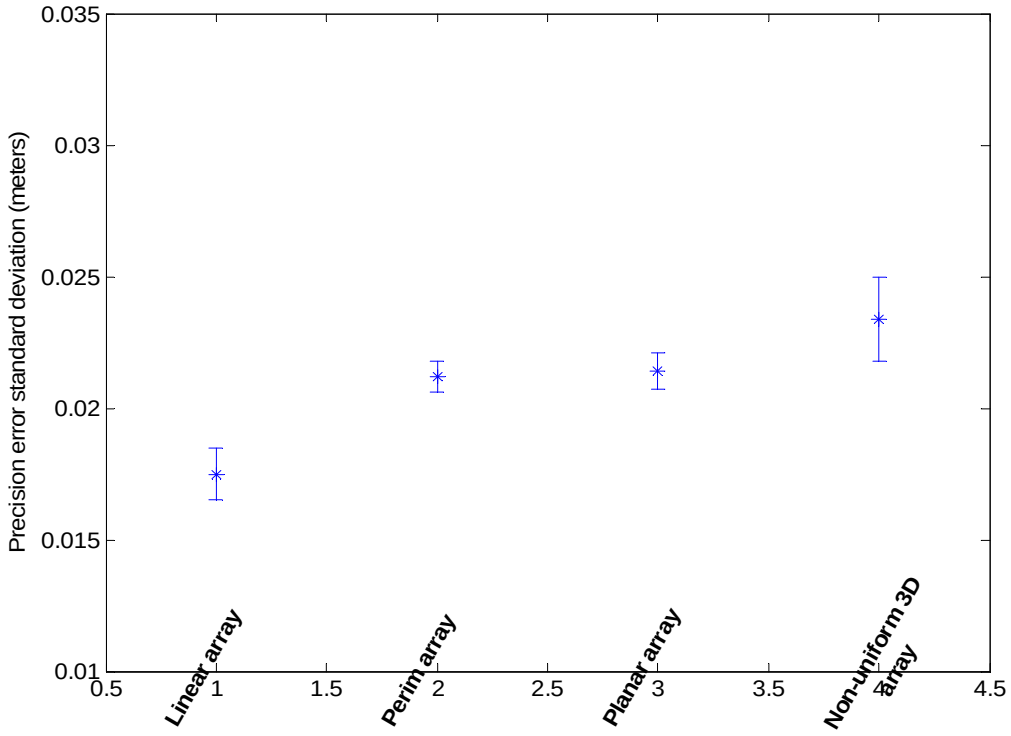
It also looks like that more distributed the microphones around the field of view, the better robustness to precision errors. The linear array which is simple in design tends to be more vulnerable to precision errors than other complex arrays. In case of an interfering speech background from Fig. 22(a), 10% drop from maximum SII occurs at a mean precision error standard deviation of 1.82cm for a linear array whereas it occurs at 2.17cm and 2.08cm for perimeter and planar arrays. The perimeter and planar arrays spread the microphones only over the ceiling or along a wall whereas the non-uniform 3D array places the microphones randomly around all the three dimensions of the FOV. Thus, the non-uniform 3D array is more distributed than the planar and perimeter arrays and hence more robust to precision errors. For a non-uniform 3D array, the mean precision error standard deviation for which a 10% drop in maximum SII occurs at 2.37cm, larger than that of planar and perimeter arrays.

Different target signals (man and woman) were used to illustrate the case when the spectral content of the target signal and masker had less overlap but yet were well within the range of the significant BIF values. Thus, a female target speaker was considered along with a male interfering speech and white noise background. Figure 23(a) indicates that, for a female speaker with interfering speech background, a 10% drop in SII occurs for an error standard deviation of 1.5cm to 2.5 cm similar to Fig. 22(a). However, for a white noise background, it occurs within an error standard deviation of 1.5 – 2.5 cm as shown in Fig. 23(b), a difference of 0.5cm with that of a male target speaker.



(a)

Figure 23: Precision error standard deviation for which a 10% drop from maximum SII occurs under given masking conditions for a female speaker. For (a) Interfering speech (male) background (b) White noise background



(b)

Figure 23, continued

As a summative statistic for performance, the mean percentage degradation in SII is examined for single location error standard deviation of 2 cm for all the microphone geometries considered in this work. The percentages were averaged across the male and female target speakers. The results are presented in Fig. 24. For the perimeter, planar, and distributed 3-D array, a location error standard deviation of 2 cm results in around 10% drop in SII as shown in Fig. 24. But, in the case of a linear array, the SII degrades by about 10-12%. So, this indicates that precision errors with a standard deviation of 1.5 cm can limit the losses in SII to less than 10% of the maximum beamformer performance for different array distributions.

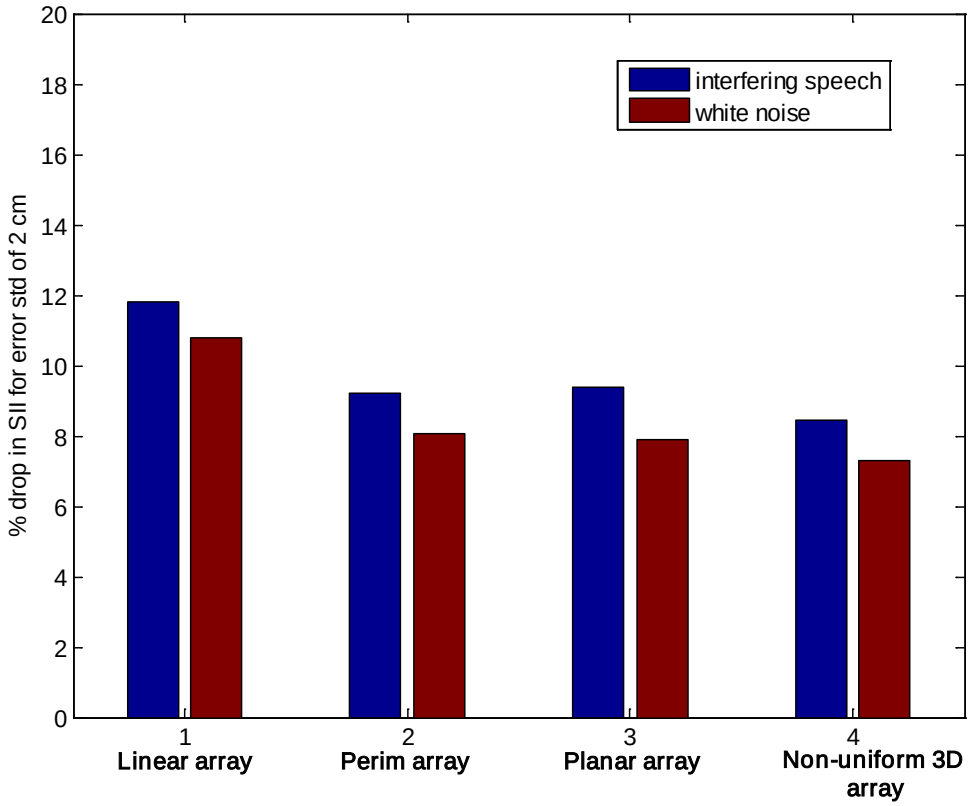


Figure 24: Mean Percentage drop in SII for an error standard deviation of 2 cm averaged across male and female speakers under given masking conditions

There might also be other factors involved that cause the difference in the array performance. A barely intelligible speech-in-noise condition was maintained before beamforming while studying the performance for different arrays. The initial intelligibility measures for each case might have been different, which differs with the source locations and test signals used. Suppose if the non-uniform array had a lower initial SII (at zero error), then the sensitivity may also be a function of it on the SII curve. Placing the sources at different positions changes the initial SII and thus a Monte Carlo simulation was performed with various random target and noise placements and the numbers were averaged out. The signals are scaled such that the mean maximum SII differ narrowly for various arrays. For example, the mean initial SII for linear, perimeter, planar and non-uniform arrays were maintained around 0.31, 0.32, 0.28 and 0.30 respectively in case of a male interfering speech condition.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

This chapter summarizes the impact of microphone positional errors on speech intelligibility based on the inferences derived from the simulations and experiments. The initial section describes the conclusions made on the acceptable limits in precision errors during the calibration process. Section 5.1 discusses the possible future directions that could be followed to extend the research on the enhancement of speech intelligibility.

This thesis examines the influence of spatial errors in microphone positions during the calibration process on speech intelligibility. These spatial errors get translated into time-delay errors between the microphone signals thus degrading the beamformer's output. These errors have a frequency-dependent impact on the enhancement from beamforming algorithms. Uniformly distributed random numbers were used to model the microphone positional errors. Analytical expressions were derived to show a sinc functional relationship in the expected power loss from the standard deviation of the positional error as a function of wavelength.

It is indicated from the derivations that a standard deviation of one-quarter wavelength would result in an effective incoherent summation and no enhancement from the beamformer. These results were then incorporated into the SII intelligibility metric and simulations and experiments were used to investigate the intelligibility loss for a variety of array geometries with different distracting sources. As this work mainly focused on improving the intelligibility for a barely-intelligible signal, the target signal was scaled to achieve a SII of about 0.3 after beamforming. Based on the recordings used during the simulations with 16 microphones, it is suggested that the input SNR has to be in the critical range of -20 dB to -10 dB to make the speech barely intelligible and to achieve a quite intelligible speech after beamforming, in case of an interfering speech background. For white noise background, this critical range of SNR is given to be -10 dB to -3 dB.

Moreover, results show that a microphone positional error with a standard deviation of less than 1.5cm, limits the losses in intelligibility metric to less than 10% of the maximum beamformer performance for different array configurations. It has also been shown that the more distributed the microphones are around the field of view, the better the robustness to precision errors. Of different array distributions experimented, the linear array tends to be more vulnerable whereas the non-uniform 3D array showed a robust performance to positional errors.

5.1 Future work

For a more comprehensive evaluation of the impact of microphone positional errors on speech intelligibility, following cases can be taken into consideration and analyses can be performed.

- Multi-talker cocktail party recordings can be included as another case of distracting source for the target speaker and intelligibility analyses can be performed.
- Experimental setups with different number of microphones and array configurations can be examined. Also, SNR analysis of each microphone can be carried out to analyze the impact on beamforming and relationships can be found which can guide the design of an array.
- Different partition bands procedures (critical band, equally-contributing critical band) and speech levels (shout, loud, raised) can be included in the SII estimation to inspect the intelligibility changes.
- SII computations can be modified to incorporate the adaptive beamforming techniques to improve the intelligibility in adverse speech-in-noise conditions.
- Extensive analysis can be performed with more test data that investigates different speech-in-noise conditions like changes in reverberation levels and speaker orientations, effects of monaural/binaural hearing and visual cues.

REFERENCES

1. http://www.kemt.fei.tuke.sk/Predmety/KEMT320_EA/_web/Online_Course_on_Acoustics/intelligibility.html
2. Speech Intelligibility, NEMA Supplement, Fire Protection Engineering Society of Fire Protection Engineers, 2002.
3. http://www.meyersound.com/support/papers/speech/intro.htm.
4. N.R. French, J.C. Steinberg, Factors Governing the Intelligibility of Speech Sounds, The journal of the Acoustical society of America, 1947.
5. http://en.wikipedia.org/wiki/Intelligibility_(communication).
6. Herman J.M. Steeneken (TNO Human Factors), The Measurement of Speech Intelligibility, Proceedings-Institute of Acoustics, 2001.
7. Adelbert W.Bronkhorst, The Cocktail Party Phenomenon: A review of Research on Speech Intelligibility in Multiple-Talker Conditions, Acta Acustica, 2000, p. 117-128.
8. Jeff Rodman, The effect of bandwidth on Speech Intelligibility, Polycom inc., White paper, 2003.
9. Kalyan S. Kasturi, Intelligibility of Filtered Speech and Estimation of Frequency Importance Functions, Master's Thesis, University of Texas at Dallas, 2002.
10. Storm, A., Speech Quality Investigation using PESQ in a simulated Climax System for ATM, Master's Thesis, Lulea University of Technology, 2007.
11. ANSI S3.2 "Method of measuring the Intelligibility of Speech over Communication Systems", American National Standards Institute, New York, 1989.
12. ANSI S3.5-1997 "American National Standard Methods for Calculation of the Speech Intelligibility Index", American National Standards Institute, New York, 1997.
13. Benjamin W.Y. Hornsby, The Speech Intelligibility Index: What is it and what's it good for, The Hearing Journal, 2004.
14. Koenraad S. Rhebergen, Niek J. Versfeld, A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners, Acoustical Society of America, 2004.
15. Monica L. Hawley, Ruth Y. Litovsky, and H. Steven Colburn, Speech intelligibility and localization in a multi-source environment, Acoustical Society of America, 1999.

16. Nima Mesgarani, Shihah Shamma, Ken Grant, Ramani Duraiswami, Augmented Intelligibility in simultaneous Multi-talker Environment, Proc. International Conference on Auditory Display (ICAD'03), 2003.
17. Tobias, J.V., Auditory Processing for Speech Intelligibility Improvement, Department of Transportation, Federal Aviation Administration, 1970.
18. Vincent Colotte and Yves Laprie, Automatic enhancement of speech intelligibility, IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '00 Proceedings, 2000.
19. Jan Rademacher, Joerg Bitzer, and Joerg Houpert, Increasing Speech Intelligibility by Denoising: What can be achieved, Proceedings: International Association for Forensic Phonetics and Acoustics, 2007.
20. Ephraim, Y.; Malah, D, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator, IEEE Transactions on Acoustics, Speech and Signal Processing, 1984.
21. Boll, S.F., Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Transactions on Acoustics, Speech and Signal Processing, 1979: p. 113-120.
22. Michael L. Seltzer, Bhiksha Raj, Calibration of microphone arrays for improved speech recognition, Proceedings of EUROSPEECH, Aalborg, Denmark, 2001.
23. Saunders, G., and Kates, M., Speech intelligibility enhancement using hearing aid array processing, Journal of the Acoustic Society of America 1997.
24. Kevin D. Donohue, Kevin S. McReynolds, Sound source detection threshold estimation using negative coherent power, IEEE Southeastcon, 2008: p. 575-580.
25. A. Wang, K. Yao, R.E. Hudson, D. Korompis, F. Lorenzelli, S. Soli, S. Gao, Microphone Array for Hearing Aid and Speech Enhancement Applications, IEEE International Conference on Application-Specific Systems, Architectures and Processors (ASAP'96), 1996: p. 231.
26. J.M.Sachar, H.F.Silverman and W.R.Patterson III, Microphone Position and Gain Calibration for a Large-Aperture Microphone Array, IEEE Transactions of Speech and Audio Processing, 2005: p. 42-52.

27. J.M.Sachar, H.F.Silverman and W.R.Patterson III, Position calibration of large-aperture microphone arrays, IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings, 2002.
28. http://www.kemt.fei.tuke.sk/Predmety/KEMT320_EA/_web/Online_Course_on_Acoustics/intelligibility.html
29. Barry D Van Veen, K.M.B., Beamforming: A Versatile Approach to Spatial Filtering. IEEE Signal Processing Magazine, 1988. 5(2): p. 4-24.
30. Metla, S.R., Microphone Array Processing for Speech Recognition in Noisy and Conferencing Environments University of Kentucky, Lexington, 2005.
31. Ramamurthy, A., Experimental Evaluation of Modified Phase Transform for sound source detection. University of Kentucky, Lexington, 2007.
32. M. Brandstein, D. Ward (Eds.), Microphone Arrays - Signal Processing Techiniques and Applications. Springer, 2001.
33. Raykar, V. C.,"A Study of a various Beamforming Techniques and Implementation of the Constrained Least Mean Squares (LMS) algorithm for Beamforming", 2001.
34. http://xenia.media.mit.edu/~mkc/micArray/node7.html
35. Betlehem, T., Williamson, R.C., "Acoustic beamforming exploiting directionality of human speech sources", IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings. (ICASSP '03), 2003.
36. http://www.cse.yorku.ca/~vgrlab/projects/eyesEarsDesc.html
37. Adcock, J.E., Optimal Filtering and Speech Recognition with Microphone Arrays, Brown University, Providence, RI, 2001.
38. Dibiase, J.H., A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays, PhD Thesis, Brown University, Providence, RI, 2001.
39. Sungjoo Ahn , H.K., Background noise reduction via dual-channel scheme for speech recognition in vehicular environment. IEEE Transactions on Consumer Electronics, 2005.
40. Vikas Raykar, Igor Kozintsev, Rainer Lienhart, Position Calibration of Microphones and Loudspeakers in Distributed Computing Platforms. IEEE Transactions on Speech and Audio Processing, 2005.

41. Kevin D. Donohue, J. Hannemann, Henry G. Dietz, Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments, Signal Processing 87(7): 1677-1691, 2007.
42. Kevin D. Donohue, Audio Array Toolbox, Sep 2008. http://www.engr.uky.edu/donohue/audio/Arrays/MAToolbox.htm
43. http://www.goldwave.com/
44. Weisstein, Eric W., "Uniform Distribution" From MathWorld--A Wolfram Web Resource; http://mathworld.wolfram.com/UniformDistribution.html
45. David G. Long, Comments on Hilbert Transform Based Signal Analysis, Brigham Young University, Utah, 2004.
46. Oppenheim A.V. and Shafer R.W., Discrete-Time Signal Processing, Prentice--Hall, Inc., 1989.
47. Kevin D. Donohue, http://www.engr.uky.edu/~donohue/audio/Examples/Examples.htm http://www.engr.uky.edu/~donohue/audio/Data/audioexpdata.htm [Audio lab setup information]

VITA

Arulkumaran Muthukumarasamy was born in Chidambaram, Tamil Nadu, India on July 3, 1985. He received his Bachelor's Degree in Electronics and Communication Engineering in 2006 from Anna University, Chennai, India. In pursuit of his higher education, he attended the Graduate School at University of Kentucky, Lexington, KY. He received the Kentucky Graduate Scholarship (KGS) based on outstanding academic achievements. He also worked as a Graduate research assistant for Dr. Kevin D. Donohue at the University of Kentucky, Lexington, KY.