Speaker Localization using Eenhanced Beamforming

Hussein Attiya Lafta, **Ali Yakoob Yousif**
University of Babylon College of Science - Computer Dept.
wsci.husein.attia@uobabylon.edu.iq wsci.ali.yakoob@uobabylon.edu.iq

Abstract

This paper deals with the speaker localization inside the room, the proposed system based on using distributed microphones arrays and benefit from its advantages over single microphone and enhances new method for speaker localization using direction of arrival DOA and enhanced Minimum variance distortionless response (MVDR) beamforming method. Each microphone array containing four microphones with uniform spaced on line is placed in a suitable location inside the room, the obtained results from the proposed system reveals the efficiency and robustness of proposed techniques comparing with other localization method. It is also improves SNR computed to evaluate the enhanced beamforming the results show that the output signal is stronger against the noise and interferences.

Keyword: Speaker localization, Microphone array, beamforming.

الخلاصة

هذا البحث يتناول موضوع تحديد مكان المتكلم داخل الغرفة، حيث ان الموضوع المقترح مبني على أساس استخدام مصفوفة الميكروفونات الموزعة للاستفادة من مميزاتها عن استخدام ميكرفون مفرد وكذلك العمل على تحسين طريقة جديدة لتحديد مكان المتكلم باستخدام طريقة الاتجاه الواصل وطريقة تكوين الشعاع نوع استجابة التباين الأقل المقاوم للتشويه (MVDR). كل مصفوفة ميكرفونات تحتوي على أربعة ميكروفونات موزعة بانتظام وموضوعة بأماكن مناسبة داخل الغرفة. النتائج المتحصلة من عملية المحاكاة للطريقة المقترحة تعكس الكفاءة والقوة للنظام المقترح مقارنة مع طرق أخرى، وكذلك تبين من استخدام عامر الإشارة–الضوضاء المحسن بين ان الإشارة الناتجة تكون ذات مقاومة اعلى للضوضاء والتداخل.

الكلمات المفتاحية: تحديد مكان المتكلم، مصفوفة الميكرفونات، تكوين الشعاع.

1 Introduction

Microphone arrays have become important technology in the field of speech processing. These systems may be electronically directed to promote the desired source signal while attenuating interfering talkers and ambient noise. An array of microphones can typically overtake a single, well-aimed, highly directional microphone without requiring local placement of transducers or inconveniencing the talker with a hand-held or head-mounted microphone. These criteria make it beneficial in settings comprising multiple or moving sources. In addition, they possess a features that a single microphone does not have; namely automatic detection, localization, and tracking of active speaker in its recipient part.

A fundamental requirement of these microphone array arrangements is the need for determining the site of speech source. For audio-based functions, a precise fix on the main speaker, in addition to acknowledging any interfering speakers or intelligible noise source, it is essential for the effectiveness of array conduction as well as the enhancement of a particular source concurrently while attenuating those regarded as undesirable.

An efficient beamformer must be able to include a continuous and exact location process within its algorithm. This method demands the usage of a location predictor capable of fine resolution at high update speed. Moreover, any estimator of a similar manner will be required to less demanding computationally, so it will be applicable for actual systems (Stoica and Moses, 2005).

Beamforming is a technique used for extraction of a signal in a noisy environment. It is a technique which has applications within many fields. In acoustic beamforming, microphone arrays are used to extract one or more sound sources, which are contaminated with noise, from a certain direction (Brandstein and Ward, 2001). Wave fields are sampled in both space and time by the microphone array, which is why beamforming can be called spatio-temporal filtering, as opposed to conventional temporal filtering, where only time samples are considered (Stoica and Moses, 2005). Beamforming applications can be made either as a narrowband source is placed near the mic array or far field, when the source is placed far from the mic array (Brandstein and Ward, 2001).

Beamforming techniques can be broadly divided into two categories conventional beamformers and adaptive beamformers.

1.1 Conventional beamformers

The conventional subtype uses a range of time-delays and weightings in order to collect the signals from the sensors (mic. Array), utilizing only data about the position of the sensors in space and the directions of the wave, whereas adaptive beamforming binds this data with features of the array's actual signals, in order to optimize the rejection of undesirable signals originating from other direction.

1.2 Adaptive beamformers:

The adaptive beamformers is capable of adapting its reaction to different conditions. Two types of adaptive beamformers including time domain beamformers and frequency domain beamformers, the first is based on time operations. The main type of beamformer is delay and sum beamformer. It defers the inward signals from each array component into time interval and links them together. Occasionally, multiplication is carried out with a window throughout the array to enhance the main lobe ratio, then furthermore to include zeroes in the characteristic. The frequency domain beamformer can be classified into two types. The first type splits the various frequency constituents which are existed in the received signal into distinct frequency bins or bands that either use filter or FFT bank. Whilst a variety of delay and sum beamformers are pertained to each frequency band or bin, it is feasible to aim the main lobe to multiple directions for several frequencies resulting in a more flexible approach. The second category of frequency domain beamformers utilizing the spatial frequency. meaning that an FFT is captured across different elements of the array, therefore, the output of the N point FFT is N channels evenly divided in space. (McCowan, 2001).

2 Localization in real environment:

When using multi microphone arrays inside the room, each one will estimate DOA for sound source, the combine all the computed angles to estimate the position of the sound source. For the environment without noise, all directions line produced by the DOAs form a single point of intersection, which is the position of the sound source, are shown in figure (2.1a). In real life world, the noise will decrease the accuracy of DOAs, therefore the lines drawn by the DOAs do not form a single point of intersection, as shown in figure (2.1b). For estimating the precise position, then each DOAs weight must be estimated to form an estimate of the location. Another option is to find the point from which the distance to the lines drawn from the microphone arrays is minimal. (Deza and Deza, 2013)

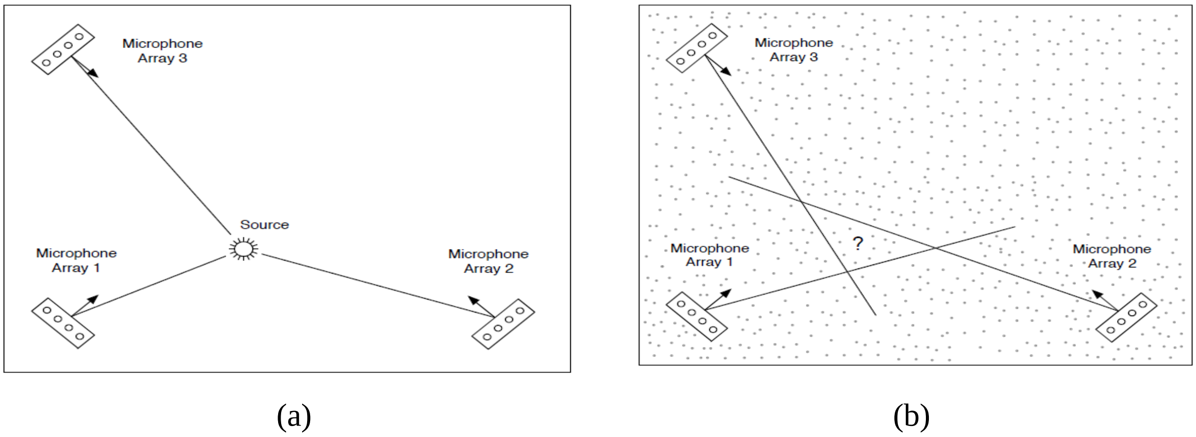

Figure (2.1) Correct estimation of DOAs of the source (a) without noise and (b) real environment (with noise)

Finding the point with minimal distance to the lines drawn by the microphone arrays can be solved using a least-squares approach (van der Heijden *et al*., 2004). The perpendicular distance from a point P = (x$_1$; y$_1$) to a line ***l*** with the equation ax + by + c = 0 is given by the following expression (Deza and Deza, 2013).

$$distance(P.l) = \frac{|ax_1+by_1+c|}{\sqrt{a^2+b^2}} \text{……………………..(2.1)}$$

The objective of finding the point with minimal distance to the *K* lines drawn by the arrays is stated as minimizing the cost function: (Deza and Deza, 2013).

$$J = \sum_{k=1}^{K} \left(\frac{|a_k x_1+b_k y_1+c|}{\sqrt{a_k^2+b_k^2}} \right)^2 \text{………………(2.2)}$$

This is done by partial differentiation and equating to zero, which results in the equations: (Deza and Deza, 2013).

$$\frac{\partial J}{\partial x_1} = \sum_{k=1}^{K} 2a_k^2\, x_1 + 2a_k b_k y_1 + 2a_k c_k = 0 \text{......................(2.3)}$$

and

$$\frac{\partial J}{\partial b_1} = \sum_{k=1}^{K} 2b_k^2\, y_1 + 2a_k^2 x_1 + 2a_k c_k = 0 \text{…………….......(2.4)}$$

From (2.3) and (2.4) the following matrix equation is formed

$$\begin{bmatrix} \sum_{k=1}^{K} a_k^2 & \sum_{k=1}^{K} a_k b_k \\ \sum_{k=1}^{K} a_k b_k & \sum_{k=1}^{K} b_k^2 \end{bmatrix} * \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} -\sum_{k=1}^{K} a_k c_k \\ -\sum_{k=1}^{K} b_k c_k \end{bmatrix} \text{… … … ...(2.5)}$$

And finally, we can get the point with the minimal distance to the lines using the microphone arrays:

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{K} a_k^2 & \sum_{k=1}^{K} a_k b_k \\ \sum_{k=1}^{K} a_k b_k & \sum_{k=1}^{K} b_k^2 \end{bmatrix}^{-1} * \begin{bmatrix} -\sum_{k=1}^{K} a_k c_k \\ -\sum_{k=1}^{K} b_k c_k \end{bmatrix} \text{…………(2.6)}$$

3 The proposed method:

In the proposed method of using multiple microphone arrays in order to enhance the methods of a sound source localization, the proposed method enhances beamforming performance, compared with using a single microphone array, by using several microphone arrays and estimate the localization of the speaker based on the DOAs using microphone array beamformers. The SNRs for the resulting outputs of the microphone array beamformers are compared for all frames. The output with the best

SNR is saved in an output buffer. This way, the system uses the output frames that have the best SNRs. This should improve performance compared to a system that uses conventional beamforming.

The proposed system steps that apply for each frame of the input signal as the following:

1. Estimate DOAs.
2. Perform beamforming in the directions of the estimated DOAs.
3. Estimate location of sound source using the DOAs estimated in step 1.
4. Estimate SNR for each array, choose the array with the best SNR.
5. Determine an adjusted angle from the chosen array to the source.
6. Perform beamforming in the direction of the adjusted angle.

4 Localization using enhanced beamforming

Capon beamforming or sometimes called Minimum variance distortionless response (MVDR) beamforming (Capon, 1969, Cox *et al.,* 1987). This technique uses a frequency-domain MVDR (FMV) algorithm developed by (Lockwood *et al.,* 2004). The original algorithm uses two-microphone. In this project, we develop the algorithm by using several number of microphones. The input signals are transformed into the frequency domain every L = 16 samples, using a Hamming window. The F=32 most recent FFTs are stored in a buffer, and then correlation matrix R_k is computed for each frequency bin ***k*** using: (Lockwood *et al.,* 2004).

$$R_k = \begin{bmatrix} \frac{M}{F}\sum_{i=1}^{F} X_{1k.i}^{*}X_{1k.i} & \frac{1}{F}\sum_{i=1}^{F} X_{1k.i}^{*}X_{2k.i} \cdots & \frac{1}{F}\sum_{i=1}^{F} X_{1k.i}^{*}X_{4k.i} \\ \frac{1}{F}\sum_{i=1}^{F} X_{2k.i}^{*}X_{1k.i} & \frac{M}{F}\sum_{i=1}^{F} X_{2k.i}^{*}X_{2k.i} \cdots & \frac{1}{F}\sum_{i=1}^{F} X_{2k.i}^{*}X_{4k.i} \\ \vdots & \vdots & \vdots \\ \frac{1}{F}\sum_{i=1}^{F} X_{4k.i}^{*}X_{1k.i} & \frac{1}{F}\sum_{i=1}^{F} X_{4k.i}^{*}X_{2k.i} \cdots & \frac{M}{F}\sum_{i=1}^{F} X_{4k.i}^{*}X_{4k.i} \end{bmatrix} \ldots\ldots\ldots\ldots\ldots..(4.1)$$

where M = 1:3 is a regularization constant used to avoid matrix singularity. The values of **L, F** and **M** are the same as used by (Lockwood *et al.,* 2004). The matrices R_k are updated every ***L*** = 16 samples. The output of the beamformer is:

$$Y_k = h_k^H X_k \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots....(4.2)$$

where $\mathbf{h_k^H}$ is a conjugate transposed vector with frequency-domain filter coefficients. In the MVDR approach we want to pass the undistorted signals with a given DOA **θ**, and attenuate signals with all other DOAs as possible, the optimization goal is to find the minimize expectation output for each frequency band:

$$\min_{h_k} h_k^H R_k h_k \quad subject\ to \quad h_k^H a(\theta) = 1 \ldots\ldots\ldots\ldots\ldots\ldots(4.3)$$

where **a(θ)** is the direction vector. This method is called Capon (Stoica and Moses, 2005).

We can compute the spatial filter as the following:

$$h_k = \frac{R_k^{-1}a(\theta)}{a^H(\theta)R_k^{-1}a(\theta)+\sigma'} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.4)$$

where **σ** is a small constant. The FMV method of Lockwood et al. assumes the direction vector, is known. However, the Capon method can also be used to find the DOA of a source. This is performed by finding the largest peak of the function:

$$\frac{1}{a^H(\theta)R_k^{-1}a(\theta)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.5)$$

Location of the speaker is determined using the enhanced FMV beamforming approach. For each microphone array, the frequency-domain output of each microphone is stored in a buffer containing the **F** latest FFTs. From this data a correlation matrix $\boldsymbol{R_k}$ is formed using expression (4.1). The FFT buffer and the correlation matrix are updated every **L** samples. This approach is similar to that of (Lockwood *et al.,* 2004). A direction vector is formed for a number of DOA candidates $\boldsymbol{\theta} \in [-90°; 90°]$ at which we calculate the output power of the beamformer for each frequency bin $\boldsymbol{k}$. We do this to determine at which angle $\boldsymbol{\theta_{DOA}}$ the magnitude of the beamformed signal is at its maximum. This angle $\boldsymbol{\theta_{DOA}}$ is the DOA of the sound source relative to the microphone array. The spatial filter weights can be calculated after determining DOA to the speaker for the current frame. The weights are saved in a matrix in order to determine the SNR of the outputs after beamforming has been done. To obtain the time-domain output of the beamformer, the spatial frequency domain weights are applied to the buffered FFT data to obtain the Fourier transform of the output. The output of frequency-domain is converted into time-domain using an N-point inverse FFT every L samples. This approach is similar to that of (Lockwood *et al.,* 2004). In this project, choosing the central L samples and using them directly for the output resulted in artifacts, the reason is behind the use of the overlap-add method has been used instead.

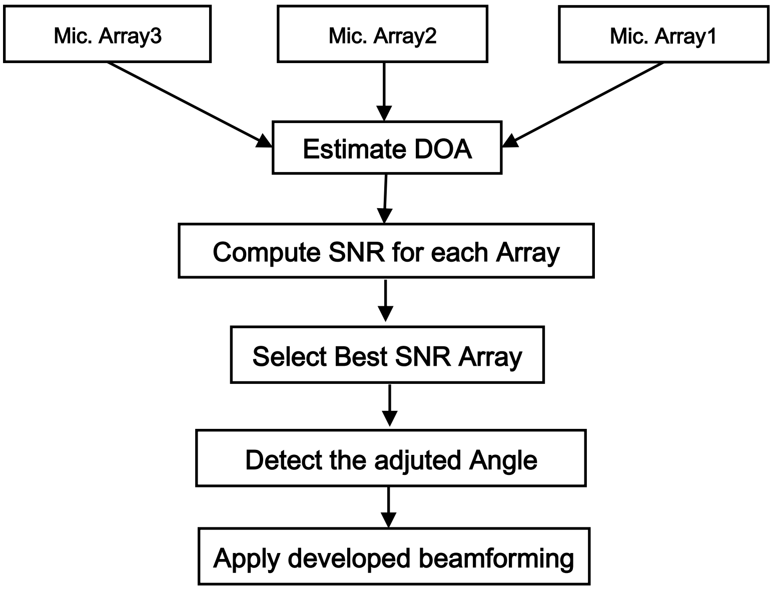

Figure (4.1) System Diagram

The position of the speaker is estimated using the DOAs estimated by the microphone array beamformers. The DOAs are combined to form an estimate of the location. This is done using a least-squares approach, where the goal is to find the point from which the distance to the lines drawn by the DOAs and the microphone array positions is minimal.

The output of each of the microphone arrays is used to decide which microphone array to use for the current output frame. This is done by estimating the SNR of each of the outputs from the microphone arrays. The SNRs are estimated using the improved version of SNR_{dB}. The SNR of each of the outputs from the beamformers is calculated using the expression:

$SNR_{db} = 10log_{10}\left(\frac{\sigma^2 speech}{\sigma^2 noise}\right)$ ……………………..(4.6)

$SNR_{improved} = 10log_{10}\left(\frac{\sigma^2 speech_{out}}{\sigma^2 noise_{out}}\right) - 10log_{10}\left(\frac{\sigma^2 speech_{input}}{\sigma^2 noise_{input}}\right)$ …………….(4.7)

In time domain the SNR input is varied according to the equation (4.8) and (4.9) as follows:

The input signal at each microphone ***mic_i*** computed as the following:

$mic_i = s_i + \propto n_i$ *i*=1..4 ……………………….(4.8)

Where **α** computed by:

$\propto = \frac{1}{\sqrt{\frac{(SNR_{input} - SNR)}{10}}}$ ……………………………..(4.9)

Where SNR$_{input}$ is the desire to noise ratio and it varies as 0dB, 10dB, 20dB, 30dB.

The estimated location is used to adjust the detected direction, and then beamforming is performed in the direction of the new angle **θ**' using the microphone array with best SNR estimate. The output of the system consists of the output frames with the best SNRs from the microphone arrays.

5 Simulation Results:

The simulation of adaptive beamformer with microphone arrays inside room is performed using Matlab 2014a simulator. The speech signal is sampled at 22050 Hz. The simulation of signals for speaker and surrounding noise is implemented using image source model. The microphone array containing 4 mics. uniformed linear spaced and each one receives the sum of reverberated signals of speaker and noise, and the distance between each successive mic. equal 2 centimeters. The evaluation is implemented on algorithm with several SNR input using different noise power. The system contains three microphone array positions. Additive white noise is used in the simulation. Frequency domain analysis is performed at input snr of 10 db and 30 db respectively.

In order to achieve real room environment and generate the reverberations, room impulse response (RIR) is used and implemented using Matlab simulation tools as a filter and then this filter convolves with noise and speaker signal as the following:

*x(t)=s(t)*h(t)*…………….(5.1) , where ***s(t)*** is speaker signal and ***h(t)*** filter function.

In this paper, the room dimensions are [6, 4, 3] for coordinates x, y and z. Speaker location at [3, 2, 1], and Mic Arrays coordinates as MicArr1 [1, 1, 1], MicArr2 [1, 3, 1], MicArr3 [5, 3, 1]. Sampling frequency used is 32 kHz, and n represent no. of reflection needed.

Figure (5.1) shows RIR with reflection factor 0.95 and 5 % is absorbed. The amplitude of impulse is represented by Y, and X represents the time.

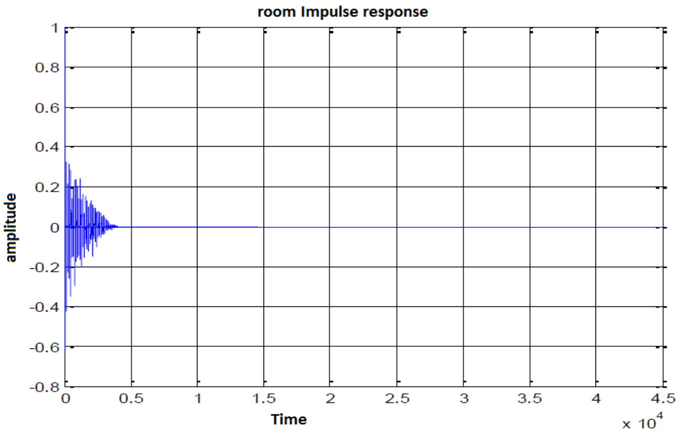

Figure (5.1) RIR with reflection parameter 0.95

Table (5.1) Evaluation of adaptive beamformer for three microphone arrays positions

Mic. Array Position	SNR input	SNR output	SNR Improved
Mic.Array 1 [1,3,1]	0	17.91	17.91
	10	22.714	16.492
	20	31.695	13.187
	30	39.721	10.368
Mic.Array 2 [5,3,1]	0	16.121	16.73
	10	22.714	16.492
	20	31.695	13.187
	30	39.721	10.368
Mic.Array 3 [1,1,1]	0	18.327	18.421
	10	23.212	15.56
	20	30.252	13.116
	30	41.063	12.723

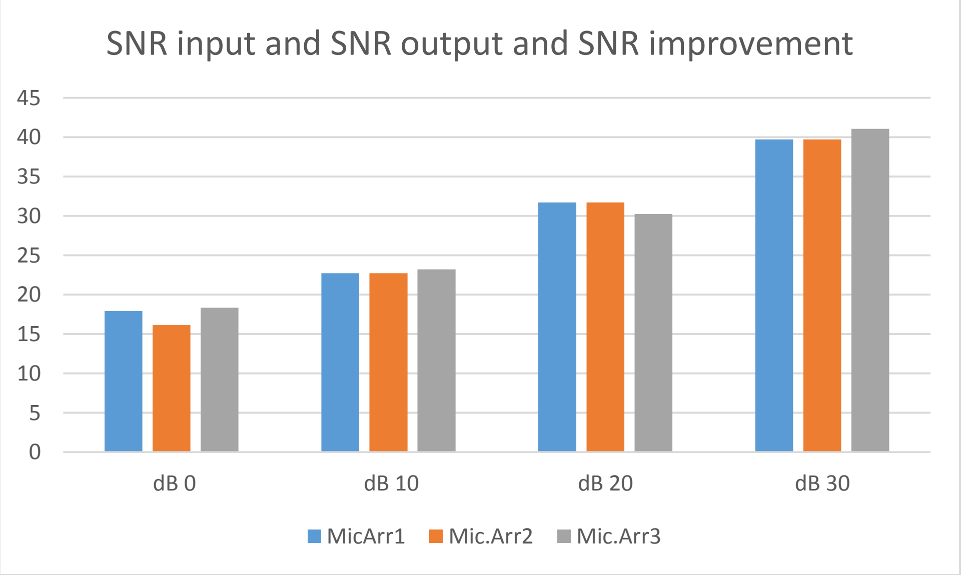

Figure (5.2) values of SNR$_{input}$, SNR$_{output}$ and SNR$_{improvement}$ for the positions of 3 microphone array

Different values of input SNR are used (0dB, 10dB, 20dB and 30dB) corresponding output SNR and applied on each position, and the improved SNR is computed using eqs. (4.6) and (4.7). Also α value computed using eq. (4.9). Table (5.1) shows values of input SNR and improves SNR for all mic. arrays positions for the output speech signals after implementing enhanced beamformer, and the results reveal better SNR values when SNR input is 30dB.

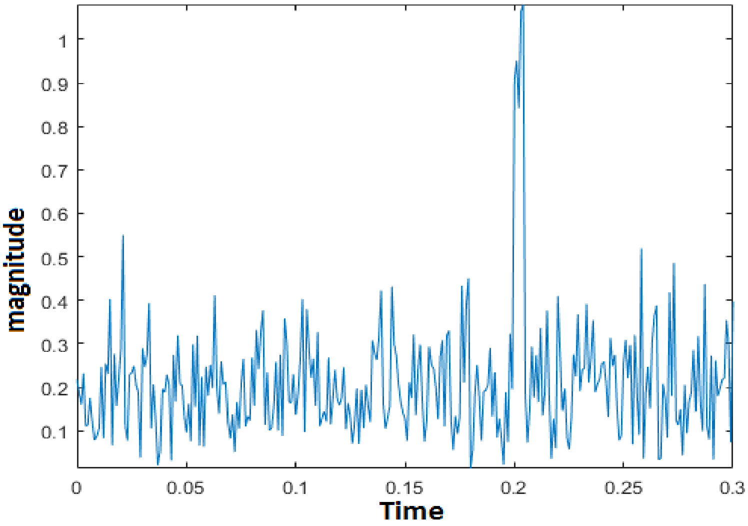

Figure (5.3) Adaptive beamformer output (without interference)

From the figure, we can see that the signal becomes much stronger compared to the noise. The output SNR is obviously stronger than that of the received signal, To see the beam pattern of the beamformer, we plot the array response along 0 degrees elevation with the weights applied. Since the array is a ULA with isotropic elements, it has ambiguity in front and back of the array. Therefore, we only plot between -90 and 90 degrees in azimuth.

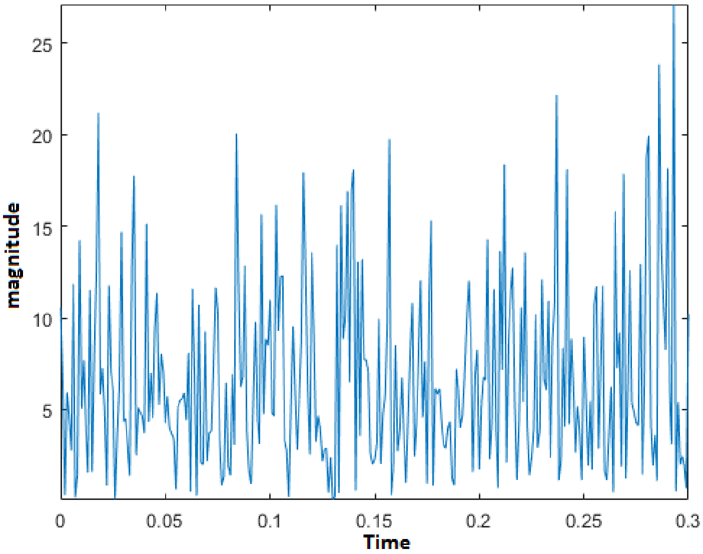

Figure (5.4) Adaptive beamformer output (real environments)

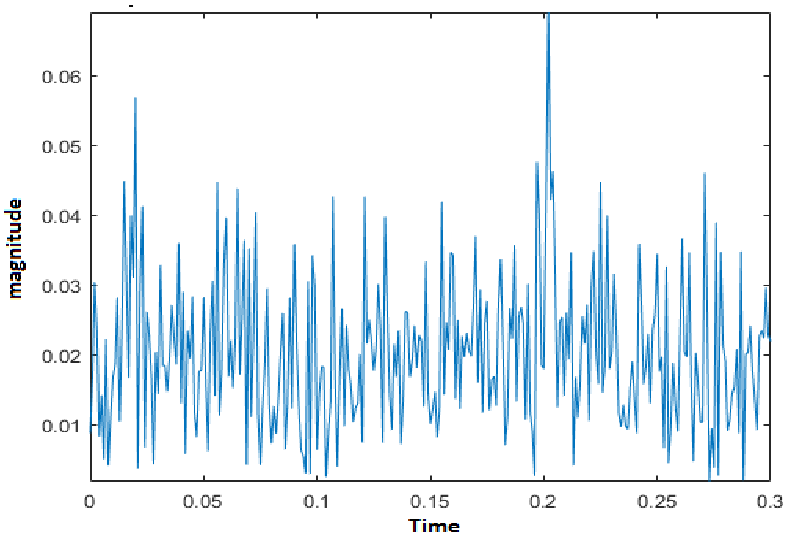

Figure (5.5) Adaptive beamformer output with signal direction mismatch

When we look at the beamformer response pattern, we see that the enhanced beamformer tries to suppress the signal arriving along 50 degrees because it is treated like an interference signal. The enhanced beamformer is very sensitive to signal-steering vector mismatch, especially when we cannot provide interference information. The beamformer also has a gain of 0 dB along the target direction of 50 degrees. Thus, the enhanced beamformer preserves the target signal and suppresses the interference signals. Figure (5.6) shows the response pattern compared with conventional beamformer.

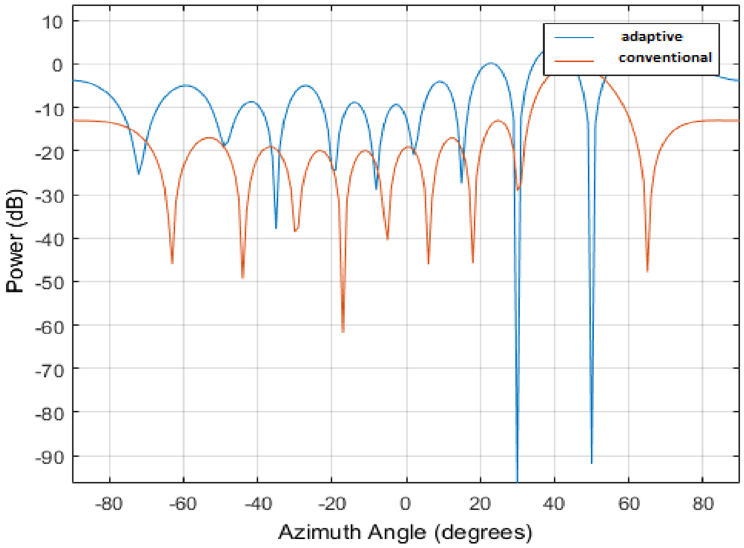

Figure (5.6) Azimuth angle for adaptive beamformer and conventional beamformer

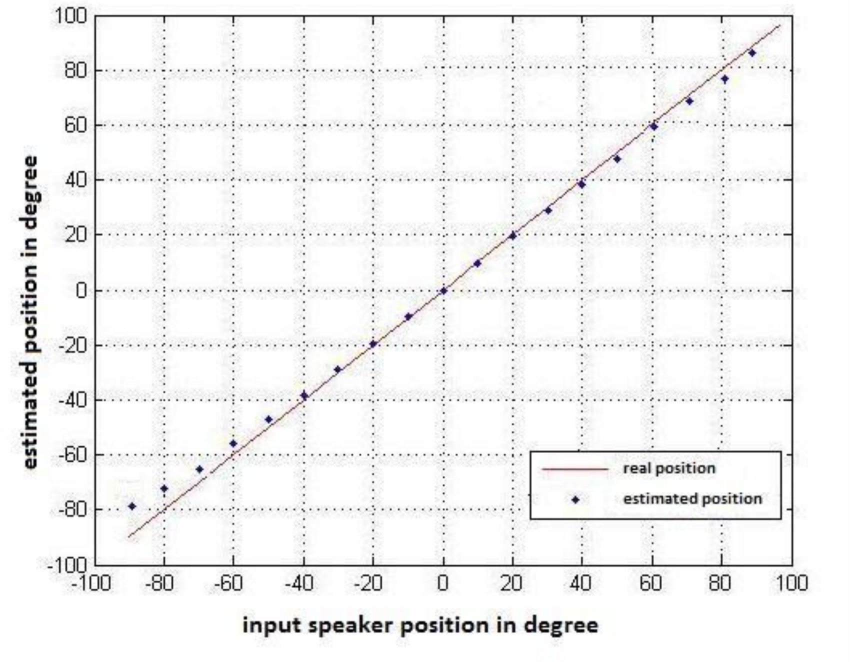

Figure (5.7) estimated and real position

In figure (5.7) the X represent the input speaker position in degree and y represent the estimated position using enhanced beamformer in degree. It is clearly from the above fig. the precision of estimated position with minimum error range (0.01 to 1.6) for the input source positions (-50° to 50°) when applying the enhanced beamforming.

Conclusion

The results of the project show that, the distributed microphone array outputs based on SNR estimates of the outputs give stronger signal against noise and reverberation. Further experiments would have to be conducted in order to determine whether the proposed method is capable of improving the SNR of the output signal**.** The position of mic. arrays inside the room is important and is affected by the surrounding noise attenuation and consequently affects the speaker signal. The enhanced beamformer shows good result using improved SNR of 10-30 dB according to room environment, least error value is acquired when the speaker stands against mic. array and high error occurs when the speaker is on same line with mic. array.

References

Brandstein, M. and D. Ward (Eds.) 2001, "Microphone Arrays – Signal Processing Techniques and Applications". Springer,.

Capon, J. , 1969 ,"High-resolution frequency- wavenumber spectrum analysis", Proceedings of the IEEE 57 (8), , 1408-1418.

Christensen, M. and A. Jakobsson, 2009, "Multi-Pitch Estimation. Synthesis Lectures on Speech and Audio Processing", Morgan & Claypool Publishers,. Toolbox available online at: http://www.morganclaypool.com/page/multi-pitch

Cox, H., and et al., 1987, "Robust adaptive beamforming. Acoustics", Speech and Signal Processing, IEEE Transactions on 35 (10), , pp.1365-1376.

Deza, M. M. and E. Deza, 2013 "Encyclopedia of Distances" (Second ed.), Springer,.

Ellis, D., "Aurora noise database", 2002. Technical report,. The noise database can be downloaded at:http://www.ee.columbia.edu/~dpwe/sounds/noise/.

Habets, E. , 2010 "Room impulse response generator for matlab",. http://home.tiscali.nl/ehabets/rir_generator.html.

Habets, E., 2011"Signal generator for matlab",. http://home.tiscali.nl/ehabets/ signal_generator.html.

Lockwod, M. E., and et. al., 2004 "Performance of time and frequency-domain binaural beamformers based on recorded signals from real rooms", Journal of the Acoustical Society of America 115 (1),.

McCowan. A. 2001"Robust Speech Recognition using Microphone Arrays," PhD thesis, Queensland University of Technology, Australia,.

Stoica P. and R. Moses, 2005 "Spectral Analysis of Signals", Prentice Hall,.

van der Heijden, F., and et al., 2004, "Classification, Parameter Estimation and State Estimation", Wiley.