Speech Enhancement Using Beamforming Techniques

A Project Report

submitted by

Namratha S (11EC49)

Priyesh Ghiya (11EC72)

Sandeep B V (11EC82)

Supreeth Prajwal S (11EC101)

under the guidance of

Dr. Deepu Vijayasenan

in partial fulfilment of the requirements

for the award of the degree of

BACHELOR OF TECHNOLOGY

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE - 575025

April 21, 2015

DECLARATION

by the B.Tech students

We hereby ***declare*** that the Project entitled **Speech Enhancement Using Beamforming Techniques** which is being submitted to the ***National Institute of Technology Karnataka, Surathkal*** in partial fulfillment of the requirements for the award of the Degree of ***Bachelor of Technology*** is a ***bonafide report of the research work carried out by us***. The material contained in this thesis has not been submitted to any University or Institution for the award of any degree.

Namratha S (11EC49)
Priyesh Ghiya (11EC72)
Sandeep B V (11EC82)
Supreeth Prajwal S (11EC101)

Department of Electronics and Communication Engineering

Place: NITK - Surathkal
Date: April 21, 2015

CERTIFICATE

This is to certify that the B.Tech. Project Work Report entitled **Speech Enhancement Using Beamforming Techniques** submitted by:
Sl.No. Register Number & Name of Student
(1) **11EC49, Namratha S**
(2) **11EC72, Priyesh Ghiya**
(3) **11EC82, Sandeep B V**
(4) **11EC101, Supreeth Prajwal S**
as the record of the work carried out by them, is accepted as the B.Tech. Project Work Report submission in partial fulfillment of the requirements for the award of degree of ***Bachelor of Technology in Electronics and Communication Engineering***

Guide: Dr. Deepu Vijayasenan
(Signature with Date)

Chairman- DUGC:
(Signature with Date and Seal)

ABSTRACT

In this project we have proposed a new algorithm for high-quality speech capture. The proposed speech enhancement algorithm uses a multi-channel beamforming technique. An advantage of employing beamforming is that the algorithm takes into account the spatial information coming from a target speaker to a microphone array so as to enhance speech. The algorithm eliminates the restriction on the geometry of the microphone arrays.

We have used *beamformIt* for implementing the beamforming technique. We have implemented a de-reverberation system via maximising the kurtosis. Additionally, a few extra modifications have been made on *beamformIt* which has further improved the output SNR values of the complete system.

Further, we have implemented a calibrated constrained sub-band beamforming algorithm where in which the filter weights are based on the Wiener solution. This algorithm is dependent on the microphone array geometry and works only in the case when the signal source remains stationary. In order to remove the dependency on the microphone array geometry, a kurtosis based blind beamforming algorithm is implemented. Various experiments have been performed to evaluate each of the above algorithms and the results have been tabulated in this report.

It was found that a frequency domain implementation of beamforming resulted in introduction of artefacts to the output, thus we have implemented an algorithm that performs beamforming in time domain. The algorithm performs a blind beamforming with the constraint of maximizing the energy/kurtosis. It was observed that the SNRs of the output from the above algorithms improved by around 3dB in comparison to the delay and sum beamformed outputs.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Problem definition

A prominent issue of recording speech data in realistic environments is the corruption of speech signal with background noise. In situations where the speaker is far from microphones, the microphone sensors not only capture speech signal from the speaker but also captures other background noise signals. The presence of such noise signals potentially degrades the quality of speech recordings. Reverberation effects that occur due to reflection of sound wave signals from hard surfaces such as tables and walls are another factor which lead to further degradation of such speech recordings.

Traditionally, speech enhancement techniques have been broadly classified into two categories, single channel and multi-channel processing techniques. Some of the single channel processing techniques involve spectral subtraction [1], Wiener filtering [2], Bayesian filtering [1], blind de-convolution [2], joint particle filter approach [3], etc. The above techniques rely upon noise spectral estimation which may be inaccurate in realistic environments. Additionally most of the above techniques either address the effects of noise or reverberation alone. On the other hand, multi-channel speech enhancement technique like beamforming [4] take into account the spatial information coming from a target speaker to a microphone array so as to enhance speech and suppress interference signals propagating from other positions.

Generally, geometric speech enhancement using beamforming rely on a specific microphone array geometry [5]. Even though such systems enhance Signal to Noise (SNR) ratio, they offer various constraints and restrict flexibility when employed in real world applications. Hence, in this work we plan to implement a speech enhancement system using beamforming that does not depend upon the microphone array geometry.

1.2 Previous work

Anguera *et. al.* used classic acoustic beamforming technique along with several algorithms to create a speaker diarization in meeting room domain [10]. The algorithms include blind reference channel selection, a two step time delay of arrival (TDOA) computation, Viterbi post-processing and a dynamic output signal weighting algorithm. Results showed a 25% relative improvement as compared to using a single microphone. Gillespie *et. al.* proposed a speech dereverberation technique by maximizing the kurtosis of the linear prediction (LP) residual [6]. They processed the microphone signals by a sub-band adaptive filtering structure using a Modulated Complex Lapped Transform (MCLT). Using these sub-band filters they maximized the kurtosis of the LP residual of the reconstructed speech. Yerneche *et. al.* implemented a DSP based sub-band beamforming algorithm for dual microphone speech enhancement [7]. This algorithm is known as Calibrated Weighted Recursive Least Squares (CWRLS). They used a recursive formulation method for updating the beamforming weight vector. Their results indicated 14dB SNR improvement, but the computational load of the DSP processor was up to 50% with 2 input channels. Jiang and Malvar proposed an adaptive speech noise removal algorithm

based on two stage wiener filtering [8]. The first wiener filter was used to estimate *a priori* signal to noise ratio aided by a classifier to separate speech from noise and the second filter generated the final output. This algorithm implemented MCLT. They achieved a 6dB SNR improvement at the output.

1.3 Motivation

In recent times, speech recognition tasks are being implemented in a majority of systems like Xbox, Siri (speech recognition software in Apple's Iphone) which require a lot of human-computer interaction. In order for such systems to have a high recognition rate, the input audio signals should be noise free and de-reverberated. The primary motivation of this project is to eliminate the requirement of a standard microphone for recording of speech. We plan to implement a speech enhancement system by processing audio signals that are recorded using a simple smart phone that is available to all of us. Another major drawback of the current speech enhancement systems is that they require a specific geometry of the microphone array. The requirement of a specific geometry may prove to be a hindrance in practical situations like recording of a classroom lecture or recording in auditoriums. In this project we hope to implement a speech enhancement system that eliminates the restriction on the geometry of the microphone arrays.

1.4 Overview

In this project, we have developed a software which eliminates noise and improves SNR values of audio signals. The report is organized as follows: In Chapter 2, an algorithm for speech dereverberation by maximizing the kurtosis is discussed. Chapters 3 discusses the theory of beamforming. In Chapter 4, the Calibrated Constrained Sub-band beamforming algorithm is discussed. The evaluation results of the above algorithm has also been discussed. In Chapter 5, the kurtosis based blind beamforming algorithm is discussed. Chapter 6 discusses the implementation of a blind beamformer in time domain. Various modifications to the said algorithm have been explained in this chapter. In Chapter 7, an integrated blind beamforming algorithm that is designed to maximize the kurtosis is explained. Chapter 8 deals with the analysis of the entire project.

CHAPTER 2

Speech dereverberation via maximising kurtosis

2.1 Introduction

In this chapter we discuss an efficient algorithm for speech de-reverberation using sub-band adaptive filtering. The main concept is to control the adaptive sub-band filters by a kurtosis metric on Linear Prediction (LP) residuals. By doing so we try to make efficient use of the *a priori* knowledge that the information to be recovered is speech.

2.2 Why Kurtosis?

For a clean voice speech, LP residuals generally have strong peaks corresponding to glottal pulses, whereas for reverberated speech such peaks are spread in time [12]. The amplitude spread of the LP residuals can be used as the reverberation metric. [6] have performed an experiment to test the above concept wherein they conclude that LP residual is a reasonable measure of reverberation.

In figure 2.1, the kurtosis of a) Noise signal is 0.3273, b) Clean speech signal is 0.9678, and c) Noisy speech signal is 0.7695. This re-iterates our assumption that maximising the kurtosis will lead to de-reverberation of speech signal.

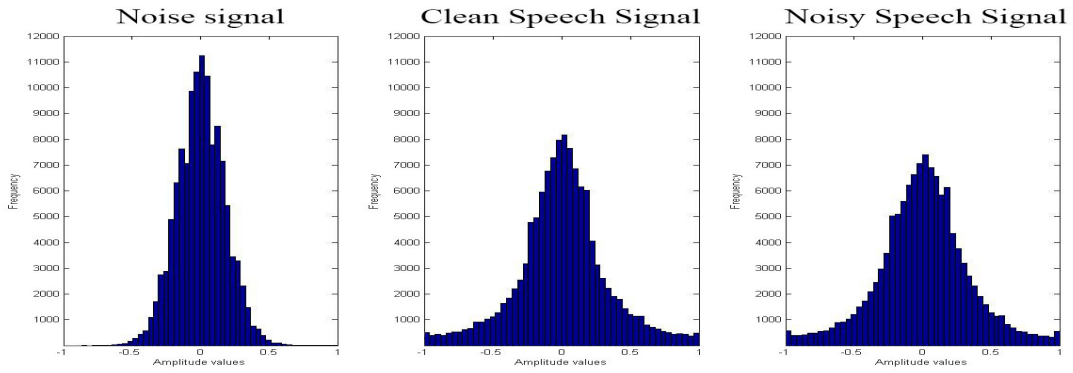


Figure 2.1: *Histogram Plots of various signals*

2.3 Design of adaptive filter h(n) in time domain

The general block diagram of the de-reverberation system is shown in figure 2.2. The received noisy reverberated speech signal is x(n) and its corresponding LP residual is $\tilde{x}(n)$. H(n) is the L-tap adaptive filter at time n. The output is $\tilde{y}(n) = h^T\tilde{x}(n)$, where $\tilde{x}(n) = [\tilde{x}(n\text{-}L+1)...\tilde{x}(n\text{-}1)\tilde{x}(n)]^T$. An LP synthesis filter yields y(n) the final processed signal. We used a feedback function, f(n), for the adaptation of h(n). An usual problem with figure 2.2 is LP reconstruction artefacts. This can be overcome by computing h(n) from $\tilde{x}(n)$ and applying it directly to x(n) as shown in figure 2.3.

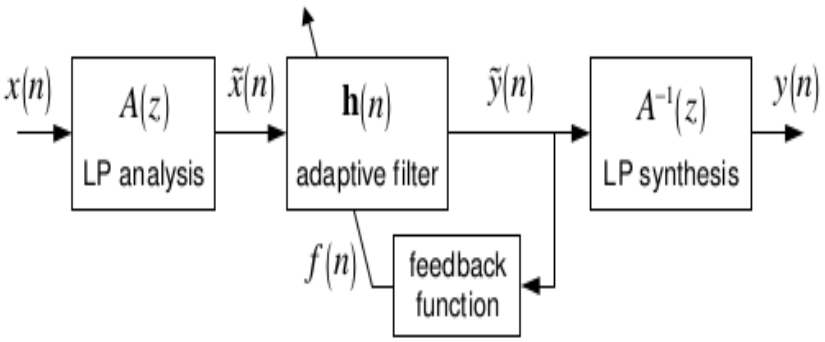


Figure 2.2: *Basic block diagram of the Adaptive system*

To derive the adaptive equations, we begin by stating *J(n)* which is given by

$$J(n) = E\{\tilde{y}^4(n)\}/E^2\{\tilde{y}^2(n)\} - 3 \tag{2.1}$$

Where the expectations E can be estimated from sample averages. In a method similar to [13] we can compute the gradient of *J(n)* with respect to the current filter by

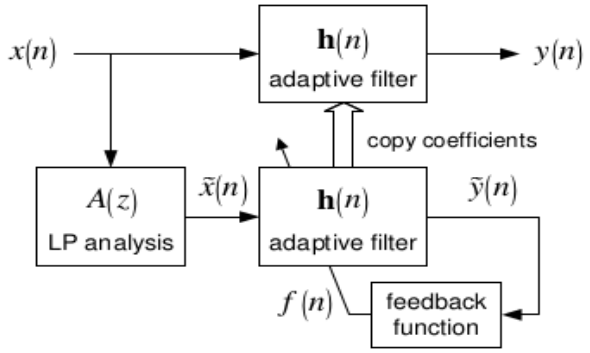


Figure 2.3: *De-reverberation system that avoids LP reconstruction artefacts*

$$\frac{\partial J}{\partial h} = \left(\frac{4((E\{\tilde{y}^2\}\tilde{y}^2 - E\{\tilde{y}^4\}\tilde{y}}{E^3\{\tilde{y}^2\}}\right)\tilde{x} = f(n)\tilde{x}(n) \tag{2.2}$$

Where f(n) is the feedback function. This function is used to control the filter updates. The final structure of the update equations for a filter that maximizes the kurtosis of the LP residual of the input waveform is given by

$$h(n+1) = h(n) + \mu f(n)\tilde{x}(n) \tag{2.3}$$

where

$$f(n) = \frac{4[E\{\tilde{y}^2(n)\}\tilde{y}^2(n) - E\{\tilde{y}^4(n)\}]\tilde{y}(n)}{E^3\{\tilde{y}^2(n)\}} \tag{2.4}$$

$$E\{\tilde{y}^2(n)\} = \beta E\{\tilde{y}^2(n-1)\} + (1-\beta)\tilde{y}^2(n) \tag{2.5}$$

Parameter μ controls the speed of adaptation and β controls the smoothness of the moment estimates.

2.4 Frequency domain implementation

The direct implementation of the time domain LMS-like adaptation equation 2.3 generally results in large variation in the eigen vectors of the autocorrelation matrices of the input signals which may lead to very slow convergence or no convergence at all under noisy conditions [14]. We implement a sub-band adaptive filtering structure based on the modulated complex lapped transform(MCLT) as proposed in [8]. A single channel MCLT based sub-band implementation of the structure of figure 2.3 is shown in figure 2.4.

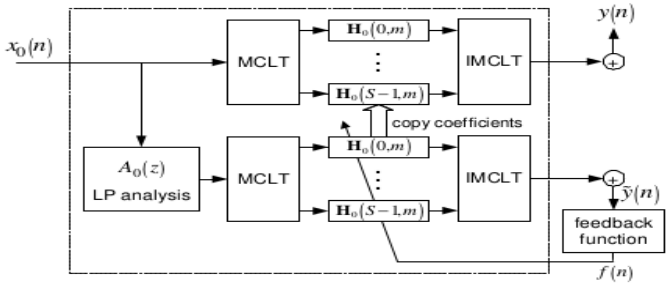


Figure 2.4: *Complete de-reverberation system*

The input speech signals are decomposed via MCLT into M complex sub-bands. Each sub-band s of each channel c is processed by a complex FIR adaptive filter with L taps, H_c(m,n), where m is the MCLT frame index. The new update equation in frequency domain is given by

$$H_c(s, m+1) = H_c(s, m) + \mu F(s, m)\tilde{X}_c^*(s, m) \tag{2.6}$$

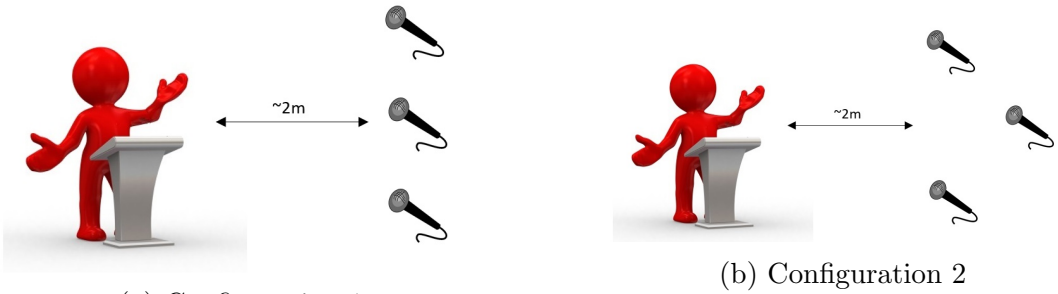
Where the superscript * denotes complex conjugation.

2.5 Evaluation of Algorithm

For all the experiments, we used $\mu = 0.0004$, β=0.9 and $\mathbf{H}_c$(s,0)=[1 0 0 0 .... 0]^T

2.5.1 Data collection

An array of three to four smart-phones having microphones were used for recording audio signals. All the smart phones used a common software for recording the audio. The speaker was positioned at about 2 meters away from the microphones and the configuration of the microphones is shown in figure 2.5. The recordings were in .wav format so as to not lose any data as a result of compression. A room environment was chosen for recording the speech samples, and speech from about 6 people were recorded. Furthermore, we intend to collect speech data by recording lectures in classrooms, corridors and auditoriums.



(a) Configuration 1

(b) Configuration 2

Figure 2.5: *Configuration of the microphones*

2.5.2 Computation of SNR

Speech samples from each person were passed as input through *BeamformIt* [10] for the purpose of speech enhancement. This enhancement can be visualised through SNR values of the speech signals. SNR values are calculated using [11], where SNR is defined as,

$$10 \log \left(\frac{peak\ speech\ power}{mean\ noise\ power} \right), \tag{2.7}$$

where power refers to the variance of a signal computed over a 20ms window. A signal energy histogram is generated by computing the root mean squared (RMS) power, in decibels, over a 20ms window and then updating the appropriate histogram bin. The window is then shifted by 10ms and the power is computed for the next window. SNR values were computed for all input and output speech samples.

2.5.3 Experiment 1

Original speech signals were passed as inputs to the de-reverberation system. β was set at 0.99. The SNR of the de-reverberated output signal improved by 4dB. However, the perceptual quality of speech deteriorated.

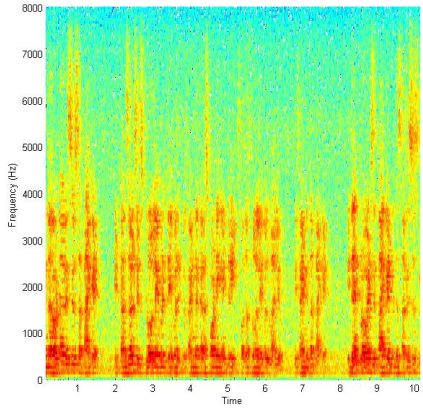
It was observed that the de-reverberation system artificially introduced additional frequency components (artefacts) at the system output. These artefacts were the main cause for deterioration in speech quality. Figure 2.6a shows the spectrogram of the original signal and figure 2.6b shows the spectrogram of the de-reverberated signal.

Figure 2.7 shows the spectrograms for a 3s window. It can be clearly observed that the de-reverberation system performs the filtering. But additional frequency components are created which affect the perceptual quality of the output signal.

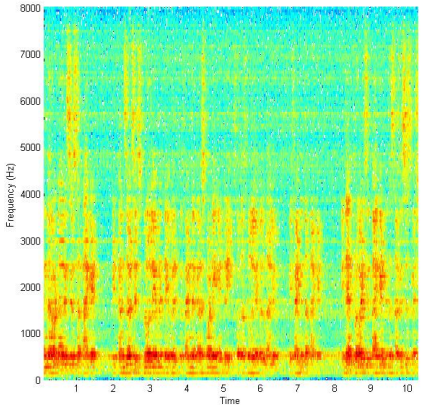
2.5.4 Experiment 2

The filter coefficients $h_c(n)$ are computed from the second order (Correlation measures) and fourth moment estimates as given in equation 2.5. The parameter β controls the smoothness of the moment estimates.

The spectrogram of de-reverberated output for β=0.99 is shown in figure 2.8a. Since

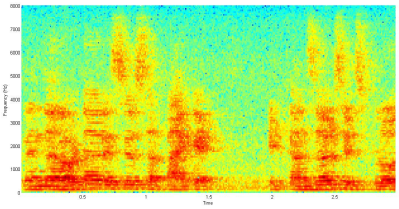


(a) Original signal

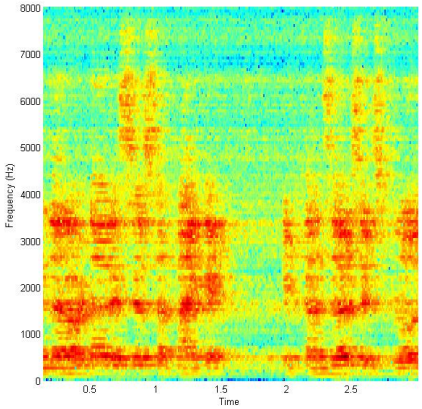


(b) De-reverberated signal

Figure 2.6: *Spectrograms of original signal and de-reverberated signal (10s window)*



(a) Original signal



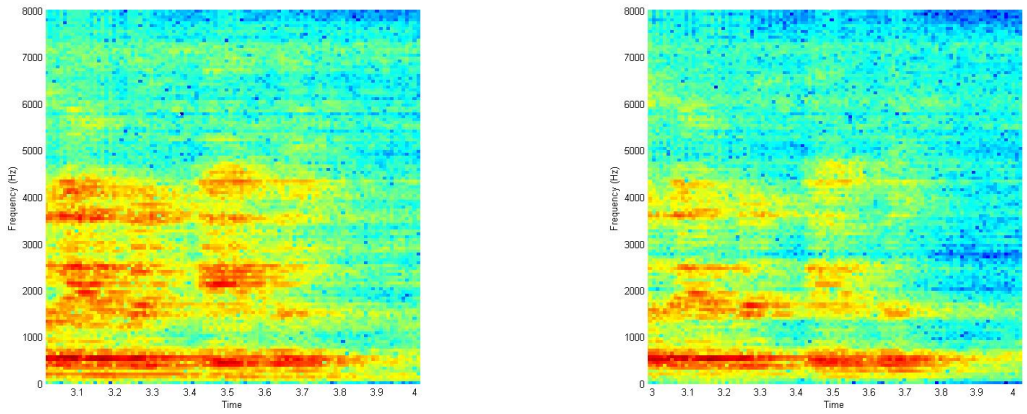
(b) De-reverberated signal

Figure 2.7: *Spectrograms of original signal and de-reverberated signal (3s window)*

the data collected did not have a high reverberation effect in them, the β value of 0.99 resulted in a very high smoothening of the moments as a result of which artefacts were introduced at the output. The experiment was conducted again for β values equal to 0.8, 0.85 and 0.9. There was a substantial decrease in the artefacts for the above three cases and the perceptual quality of the de-reverberated output was the best for β =0.9. Figure 2.8b shows the spectrogram of the de-reverberated output for $\beta = 0.9$.

2.6 De-reverberation system: Results

Figure 2.9 plots the change in SNR values between the de-reverberated output and the original signal. A total of 5 sets of data were collected with each set comprising of 3 channels. From the figure, we can observe that the SNR of the de-reverberated output is 4dB higher in comparison to the input signal.



(a) $\beta = 0.99$ (b) $\beta = 0.9$

Figure 2.8: *Spectrograms of de-reverberated signal for different β values*

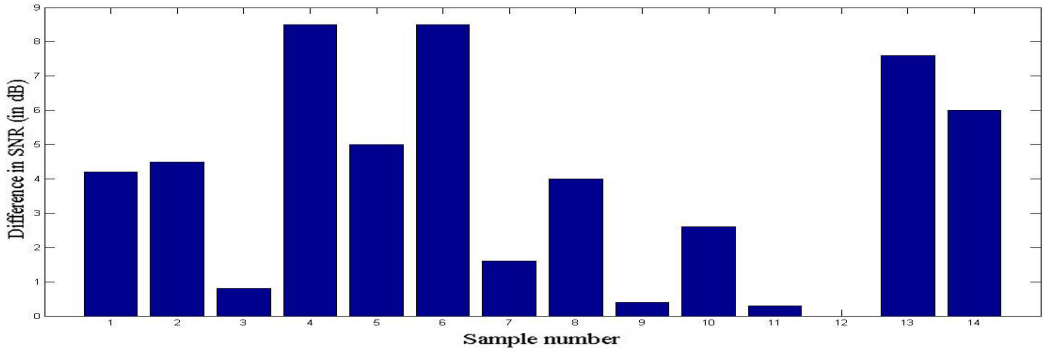


Figure 2.9: *Difference in SNR between the de-reverberated output and original signal*

CHAPTER 3

Beamforming

3.1 Introduction

Beamforming, also known as *spatial filtering*, is a signal processing technique used in sensor arrays for directional signal transmission or reception [4]. The main objective of the system is to separate out noise and interfering signals from speech signals arriving from a desired direction. The system also performs spatial filtering to separate overlapping frequency content that originate from different spatial locations. Beamforming techniques have been applied to many areas such as radar, sonar, seismology and speech enhancement.

Beamforming techniques can be broadly classified into fixed or adaptive beamforming. Fixed beamforming techniques initially fix their parameters and maintain it throughout the entire processing. While on the other hand, adaptive beamforming techniques dynamically update their parameters based on the input signal being received. In terms of implementation, fixed beamforming techniques are simpler than their adaptive counterparts but fail to eliminate highly directive (and sometimes changing) noise sources effectively. Delay and Sum Beamforming [9] technique is considered the simplest form of fixed beamforming.

3.2 Delay and Sum Beamforming

In Delay and Sum beamforming technique, the output signal y[n] is defined as

$$y[n] = \frac{1}{M} \sum_{m=1}^{N} x_m(n - \tau_m) \tag{3.1}$$

where M is the number of microphones and τ_m is the time delay for the m^{th} microphone. Here it is assumed that all the input channels are equally weighted to generate the output. A more general case of the delay and sum beamforming technique employs a sensor weight w_m for each of the input channels. The output is written as

$$y[n] = \sum_{m=1}^{N} w_m x_m(n - \tau_m) \tag{3.2}$$

The weight sensor w_m reduces the side lobe levels and enhances the beam shape. Figure 3.1 shows a block diagram of a delay and sum beamformer.

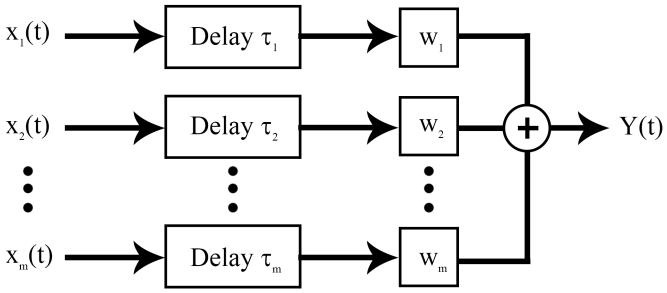


Figure 3.1: *Delay and Sum beamformer*

3.3 Experiments using *BeamformIt*

3.3.1 Experiment 1

Observations have shown that recording in different channels usually start at different time instances. As a result of which the search window for computing TDOA values in BeamformIt becomes very large. This makes computationally cumbersome and also leads to misprediction of TDOA values sometimes.

Therefore, we used the skew detector so as to align the input files across all the channels. In order to estimate the skew in different channels, an average cross-correlation metric is put in place in order to obtain the average (across time) delay between each channel and the reference channel for a set of long acoustic windows (around 5 seconds). Figure 3.2 shows the existence of skew amongst the three channels of a recording and the need for aligning them.

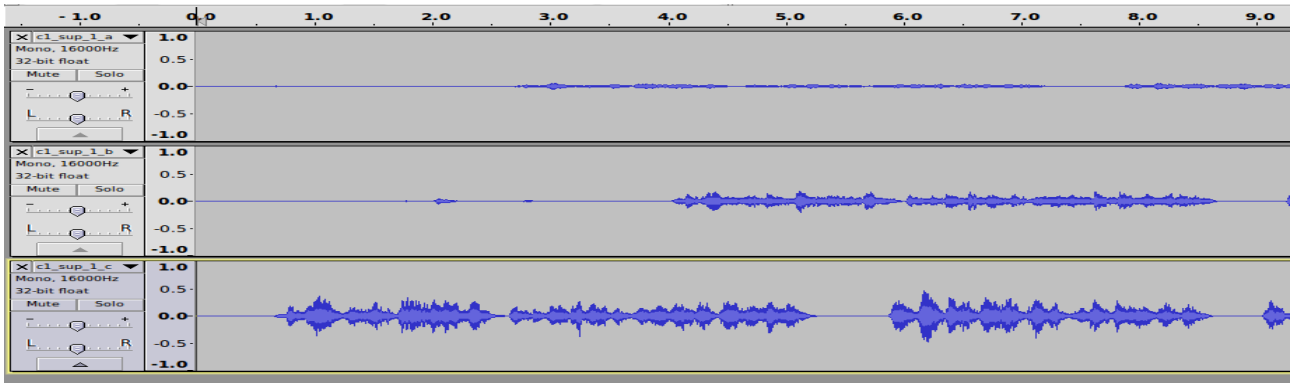


Figure 3.2: *Skew between input channels*

3.3.2 Experiment 2

In *BeamformIt*, the TDOA values are computed based on GCC-PHAT given by:

$$GCC - PHAT = \mathcal{F}^{-1}\left(\frac{X_i(f)[X_{ref}(f)]^*}{|X_i(f)[X_{ref}(f)]^*|}\right) \tag{3.3}$$

Since the data we collected did not contain reverberation to a large extent, the scaling factor $|X_i(f)[X_{ref}(f)]^*|$ lead to mis-prediction of delay values. The code of the *BeamformIt* was changed so as to modify the GCC-PHAT as given by

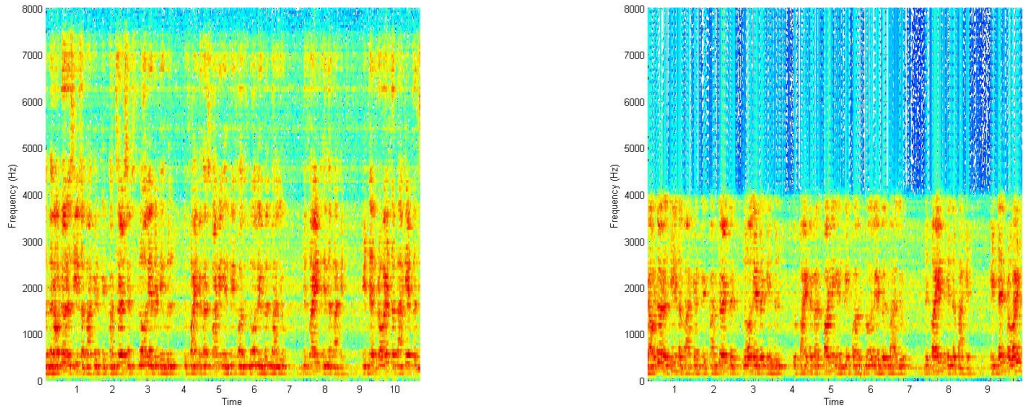
$$GCC - PHAT = \mathcal{F}^{-1}\left(X_i(f)[X_{ref}(f)]^*\right) \tag{3.4}$$

The beamforming when implemented with the above equation led to proper prediction of delay values and hence the SNR of the beamformed signal improved substantially.

3.3.3 Experiment 3

It was observed that the de-reverberation system at $\beta = 0.99$ introduced artefacts in the outputs. In order to minimize these artefacts the de-reverberated signals were passed through a lowpass filter with a pass-band frequency of 3.8kHz and a stop-band frequency of 4.2kHz. The filtered outputs were used as inputs for $beamformIt$ and the delay values obtained were used as the TDOA values in $beamformIt$ for the de-reverberated signals as inputs. Figure 3.3 shows the spectrograms of the de-reverberated signal and filtered signal respectively. The SNR of the outputs improved on an average by only 0.5dB (which was not a signficant improvement). Furthermore, upon setting $\beta = 0.9$ in the de-reverberation system, the effects of artefacts were reduced to an extent that there was no need to further filter the de-reverberated signal.

Thus, we did not obtain convincing results to justify beamforming by the above approach.



(a) Original Signal (b) Filtered Signal

Figure 3.3: *Spectrograms of the de-reverberated signal and filtered signal*

3.4 *BeamformIt*: Results

Figure 3.4 shows a comparison of the SNR values between the regular beamformed output and the de-reverberated beamformed output. A total of 5 sets of data were collected with each set comprising of 3 channels. From the figure, we can observe that the SNR of the de-reverberated beamformed output improved by around 4dB in comparison to the regular beamformed output in a majority of the cases.

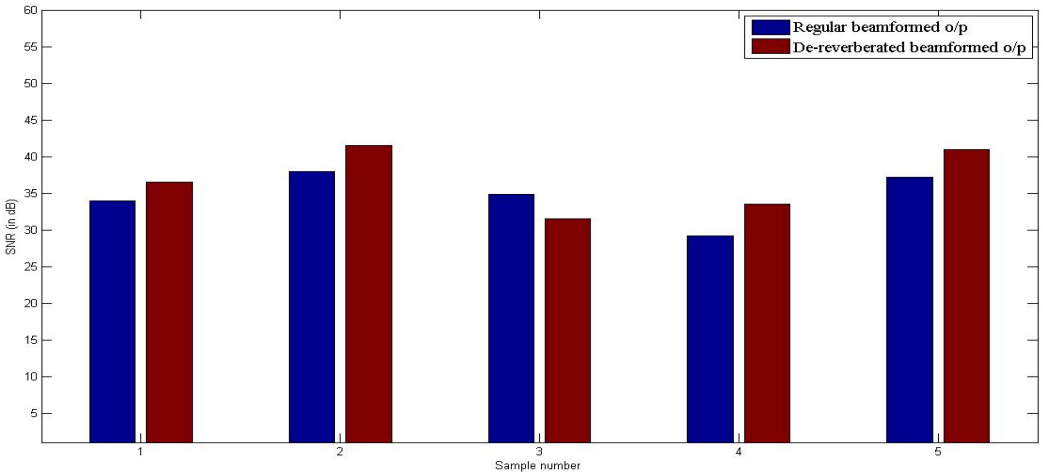


Figure 3.4: *SNR of the beamformed outputs for different samples*

CHAPTER 4

Calibrated Constrained Sub-band Beamforming

4.1 Introduction

In this chapter, we implement a constrained sub-band beamforming algorithm that is constructed around the principle of an array calibration to the real acoustical environment [15]. This algorithm performs background noise reduction while producing an undistorted, filtered version of the signal originating from a desired location.

The beamformer is based on the principle of a soft constraint formed from calibration data. The benefit of calibration is that the real room acoustical properties will be taken into account during beamforming. A sub-band beamforming implementation is chosen in order to allow the use of efficient, but computationally demanding, adaptive structures.

Information about the speech source location is put into the algorithm in an initial acquisition phase by computing the source covariance estimates for microphone observations when the source signal of interest is active alone. The objective is then formulated in the frequency domain as a weighted recursive least-squares (RLS) solution, which relies upon the precalculated covariance estimates. The adaptive beamformer continuously estimates the spatial information for each frequency band so as to track the variations in the surrounding noise environment. The algorithm then updates the beamforming weights recursively where the initial precalculated covariance estimates constitute a soft constraint.

The weight update equation for the adaptive beamformer makes use of a combined covariance matrix computed at each iteration. From the observation that the combined covariance matrix comprises of a precalculated fixed part and a recursively updated part, the beamforming problem can be divided into a fixed part and an adaptive part.

4.2 The Constrained RLS Beamformer

The adaptive array processing of the spatial and temporal microphone samples help to separate signals that have overlapping frequency content, but are originated from different spatial locations. The beamformer optimizes the array output by adjusting the weights of finite length digital filters so that the combined output contains minimal contribution from noise and interference. As a consequence, the angle of the spatial pass-band is adjusted for each frequency.

4.2.1 Signal Model

Figure 4.1 represents a typical acoustic environment where a speech signal coexists with directional interfering signals and ambient noise. Consequently, the array input vector when all the sources are active simultaneously at discrete time instant l, and which mainly contains frequency components around the central angular frequency Ω, is expressed as,

$$X(\Omega, l) = X_s(\Omega, l) + X_i(\Omega, l) + X_n(\Omega, l) \tag{4.1}$$

where $X_s(\Omega, l)$, $X_i(\Omega, l)$ and $X_n(\Omega, l)$ are the received microphone input vectors generated by the desired source, the interfering sources and the ambient noise, respectively.

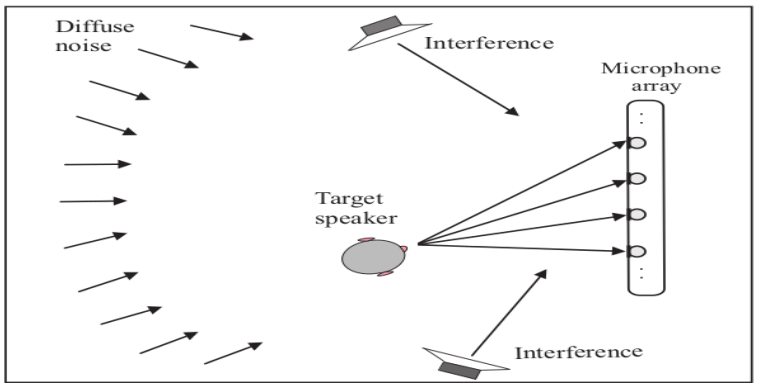


Figure 4.1: *Acoustic Model*

For the input vector $X(\Omega, l)$, the spatial covariance matrix is given by

$$R(\Omega) = E\{X(\Omega, l)X(\Omega, l)^H\} \tag{4.2}$$

The symbol $E\{.\}$ denotes the statistical expectation, and $(^H)$ denotes the Hermitian transpose. Assuming that the speech signal, the interference and ambient noise are uncorrelated, R(Ω) can be written as

$$R(\Omega) = R_s(\Omega) + R_i(\Omega) + R_n(\Omega) \tag{4.3}$$

where $R_s(\Omega)$ is the source covariance matrix, $R_i(\Omega)$ is the interference covariance matrix and $R_n(\Omega)$ is the noise covariance matrix for frequency Ω defined as,

$$R_s(\Omega) = E\{X_s(\Omega, l)X_s(\Omega, l)^H\} \tag{4.4}$$

$$R_i(\Omega) = E\{X_i(\Omega, l)X_i(\Omega, l)^H\} \tag{4.5}$$

$$R_n(\Omega) = E\{X_n(\Omega, l)X_n(\Omega, l)^H\} \tag{4.6}$$

4.2.2 Optimal Wiener Beamformer

The optimal filter weight vector based on the (narrow-band) Wiener solution [16] is given by

$$W_{opt} = [R(\Omega)]^{-1} r_s(\Omega) \tag{4.7}$$

where $r_s(\Omega)$ is the cross-covariance vector defined as

$$r_s(\Omega) = E\{X_s(\Omega, l)S(\Omega, l)^*\} \tag{4.8}$$

The signal $S(\Omega, l)$ is the desired source signal at time sample l and for frequency Ω, and $(.)^*$ stands for the conjugate operator. The output of the beamformer is given by

$$Y(\Omega, l) = W_{opt}(\Omega)^H X(\Omega, l) \tag{4.9}$$

4.2.3 Sub-band Beamforming

The broadband input signals are decomposed into sets of narrow-band signals using polyphase decomposition. If **K** denotes the total number of sub-bands, the sub-band signals each have a bandwidth that is approximately **K** times smaller in width than that of the full-band input signal. Figure 4.2 illustrated the overall architecture of the microphone array speech enhancement system, based on the sub-band beamformer.

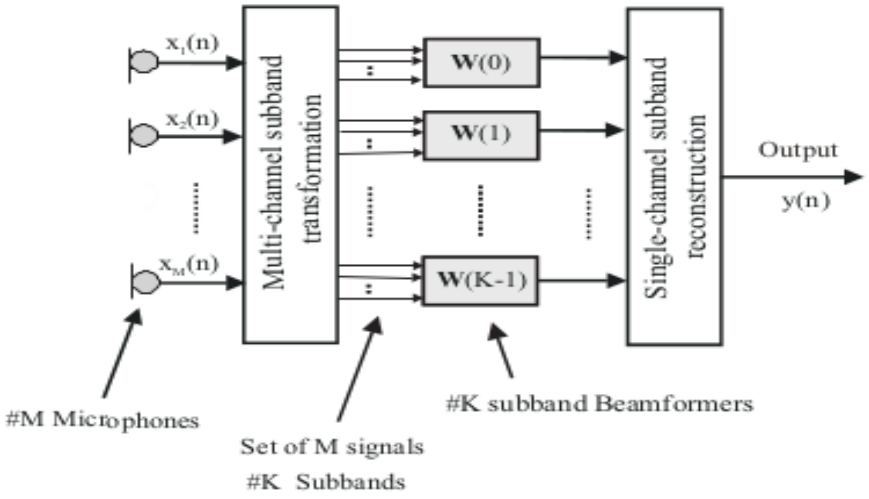


Figure 4.2: *Structure of Sub-band beamformer*

4.3 Algorithm Implementation

The array sensor input signals are samples with a frequency F_s and decomposed each into a set of K corresponding sub-band signals. These narrow-band signals constitute the inputs to a set of K sub-band beamformers.

Calibration Phase:

- Calculate the estimated source covariance matrix, $\tilde{R}_s(k)$, and the estimated cross-covariance vector, $\tilde{r}_s(k)$, according to equations 4.10 and 4.11 when the source of interest is active alone

$$\tilde{R}_s(k) = \sum_{l=0}^{N-1} X_s(k, l) X_s(k, l)^H \tag{4.10}$$

$$\tilde{r}_s(k) = \sum_{l=0}^{N-1} X_s(k, l) X_{s,r}(k, l)^* \tag{4.11}$$

- Calculate the observed data covariance matrix when known disturbing sources are active alone, i.e. $\tilde{R}_i(k)$, according to equation 4.12

$$\tilde{R}_i(k) = \sum_{l=0}^{N-1} X_i(k,l) X_i(k,l)^H \tag{4.12}$$

- The covariance matrices are saved in memory in a diagonalised form:

$$Q(k)^H \tau(k) Q(k) = (\alpha \tilde{R}_s(k) + \beta \tilde{R}_i(k))$$

  The eigenvectors are denoted:

$$Q(k) = [q_1(k), q_2(k), ..., q_M(k)]$$

- Initialize the weight vector $W_{ls}(k,l)$, as a zero vector
- Initialize the inverse of the total correlation matrix, $\tilde{R}(k,l)^{-1}$, denoted as

$$P(k,0) = \sum_{p=1}^{P} \gamma_p(k)^{-1} q_p(k) q_p(k)^H,$$

  and define the same size dummy variable matrix, **D**
- Choose a forgetting factor, $0< \lambda <1$, and a weight smoothing factor, $0< \eta <1$.

Operation Phase:

For l=1, 2, 3, ...

- When any of the covariance sources are active simultaneously, update the inverse covariance matrix as

$$D = \lambda^{-1} P(k,l-1) - \frac{\lambda^{-2} P(k,l-1) X(k,l) X(k,l)^H P(k,l-1)}{1 + \lambda^{-1} X(k,l)^H P(k,l-1) X(k,l)} \tag{4.13}$$

$$P(k,l) = D - \frac{(1-\lambda)\gamma_p(k) D q_p(k) q_p^H D}{1 + (1-\lambda)\gamma_p(k) q_p(k)^H D q_p(k)} \tag{4.14}$$

  where the index of the eigenvalues and eigenvectors is *p=l(modM)+1*
- Calculate the weights for each sample instant as

$$W_{ls}(k,l) = \eta W_{ls}(k,l-1) + (1-\eta) P(k,l) \tilde{r}_s(k)$$

- Calculate the output for the sub-band k as

$$Y(k,l) = W_{ls}(k,l)^H X(k,l)$$

The output from all sub-band beamformers are used in the reconstruction filter bank to create the time-domain output.

4.4 Evaluation of the Calibrated Beamformer

The algorithm is run for a one tap sub-band filter length, with the following parameter settings: $K = 64$, $L = 256$, $\lambda = \eta = 0.99$, $\alpha = \beta = 1$. The recordings are performed in a closed room with 3 microphones with a sampling rate of F_s=16 kHz.

The SNR of beamformed outputs for each of the recordings are shown in figure 4.3. The SNR of CRLS beamformed outputs are compared with the Delay & Sum beamformed outputs.

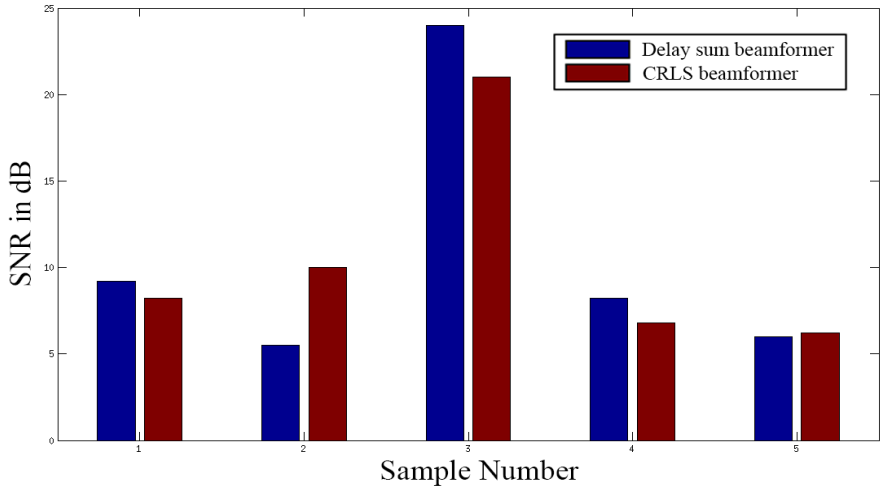


Figure 4.3: *SNR of beamformed outputs for different samples*

4.5 Conclusion

We have been able to implement a speech enhancement system in which the beamformer weights are updated recursively with a soft constraint imposed by covariance estimates. Although, the SNR of the beamformed outputs from the CRLS based algorithm are lower by 1dB (on an average) in comparison to the Delay & Sum beamforming algorithm, we have been able to show that updating the beamforming weights rather than aligning the input signals does results in an improved SNR. In the next chapter, we have implemented a kurtosis based Blind-beamforming algorithm that follows an LMS update algorithm to estimate the beamformer weights.

CHAPTER 5

Kurtosis based blind beamforming

5.1 Introduction

In this chapter, we present a beamforming algorithm for speech enhancement via kurtosis maximization [17]. The algorithm is first described for a single frequency bin and operates on instantaneous mixtures. Then, a convergence improvement is presented for use with real-time implementations. Finally, the algorithm is extended to all frequency bins, for use with convolutive mixtures. This approach is designed for cases where the speech and noise sources may be non-stationary.

Speech recorded in real acoustic environments can be modeled as the desired speech source $s(n)$ convolutively mixed with interference $v_m(n), m = 1, ..., M$, recorded at M microphones. The signals appear at the m^{th} microphone in the array as

$$x_m(n) = \sum_{p=0}^{P-1} h_m(p)s(n-p) + v_m(n) \tag{5.1}$$

To attempt to recover the speech signal, a $Q-tap$ FIR filter is applied to each microphone channel, and all channels are summed to form the output $y(n)$:

$$y(n) = \sum_{m=1}^{M} \sum_{q=0}^{Q-1} w_m(q)x_m(n-q) \tag{5.2}$$

This problem can be restated in the frequency domain as:

$$Y_k[r] = W_k^H X_k[r] \tag{5.3}$$

where $k = \{0...K-1\}$ is the frequency bin index, $r = \{0...R-1\}$ is the frame index, $X_k[r] = [X_{1,k}[r], ..., X_{M,k}[r]]^T$, and $W_k = [W_{1,k}, ..., W_{M,k}]^T$

A maximizing kurtosis technique is applied to every bin in the frequency domain. This will provide the reconstruction filters to maximize the kurtosis in each bin, since each bin is a complex instantaneous mixture.
The kurtosis of a complex random variable y is defined as

$$k(y) = \frac{E[|y|^4]}{(E[|y|^2])^2} - 2 \tag{5.4}$$

where $y = w^H x$.

5.2 Approach

Let the vector x represent a single frequency bins Fourier coefficient for each of the microphones in the array. The beamformer weights in a single frequency bin will be represented by w.

The objective is to find the reconstruction weights w_{opt} that maximize $k(w^H x)$, the narrowband kurtosis. Since the kurtosis surface is circularly symmetric, βw_{opt} is also a maximizer, for any $\beta = 0$. Therefore, we constrain $||w||_2^2 = 1$ to ensure $w^H x$ does not grow without bound. Since the problem is non-convex, a gradient ascent method can be employed to numerically find a local maxima. An LMS algorithm is employed, with modifications to project the normalized gradient back onto the unit sphere.

First, the gradient of the kurtosis of the output signal with respect to the reconstruction weights is estimated and normalized. The kurtosis surface is circularly symmetric, and the normalized gradient lies tangent to the unit sphere. The purpose of normalizing the gradient is to ensure a fixed step size in all frequency bins. While this may not be the optimal update strategy, it provides a simple way to ensure all bins converge at a similar rate, with stable convergence properties. The normalized gradient is scaled and added to the current w, as in a standard LMS update. This gives the intermediate vector a.

$$a = w_{n-1} + \mu \frac{\bigtriangledown_w k}{|| \bigtriangledown_w k||_2} \tag{5.5}$$

Since a no longer lies on the unit-norm constraint space, it is projected back onto the unit-sphere.

$$w_n = \frac{a}{||a||_2} \tag{5.6}$$

The kurtosis measure as stated in equation 5.4 can be rewritten as

$$k(y) = \frac{E[|w^H x|^4]}{(E[|w^H x|^2])^2} - 2 \tag{5.7}$$

and upon expanding in order to remove the absolute value operation, we get

$$k(y) = \frac{E[(w^H x x^H w)^2]}{(E[w^H x x^H w])^2} - 2 \tag{5.8}$$

The gradient of kurtosis with respect to the reconstruction weights is $(\bigtriangledown_w k)$

$$\frac{\partial k(y)}{\partial w} = \frac{\partial}{\partial w} \frac{E[(w^H x x^H w)^2]}{(E[w^H x x^H w])^2} \tag{5.9}$$

The gradient is expanded using the product rule and we finally get,

$$\frac{\partial k(y)}{\partial w} = \frac{4E[w^H x x^H w w^H x x^H]}{(E[w^H x x^H w])^2} - \frac{4E[(w^H x x^H w)^2]E[w^H x x^H]}{(E[w^H x x^H w])^3} \tag{5.10}$$

Factorizing each of the $E_n(f(w,x))$ into $a(w_n)E_n(b(x))c(w_n)$, for some functions a, b and c and using an autoregressive moving average technique for each of the $E_n(b(x))$, we reduce the computational complexity of this approach.

$$E_n(b(x)) = \alpha E_{n-1}(b(x)) + (1-\alpha)b(x_n) \tag{5.11}$$

The final gradient expression then becomes:

$$\frac{\partial k(y)}{\partial w} = \frac{4vec(w(vec(ww^H))^H)^H E[vec(x(vec(xx^H))^H)x^H]}{(w^H E[xx^H]w)^2} - \frac{4vec(ww^H)^H E[vec(xx^H)vec(xx^H)^H]vec(ww^H)w^H E[xx^H]}{(w^H E[xx^H]w)^3}$$

where the *vec* operator stacks the columns of an m x n matrix to form a mn x 1 column vector.

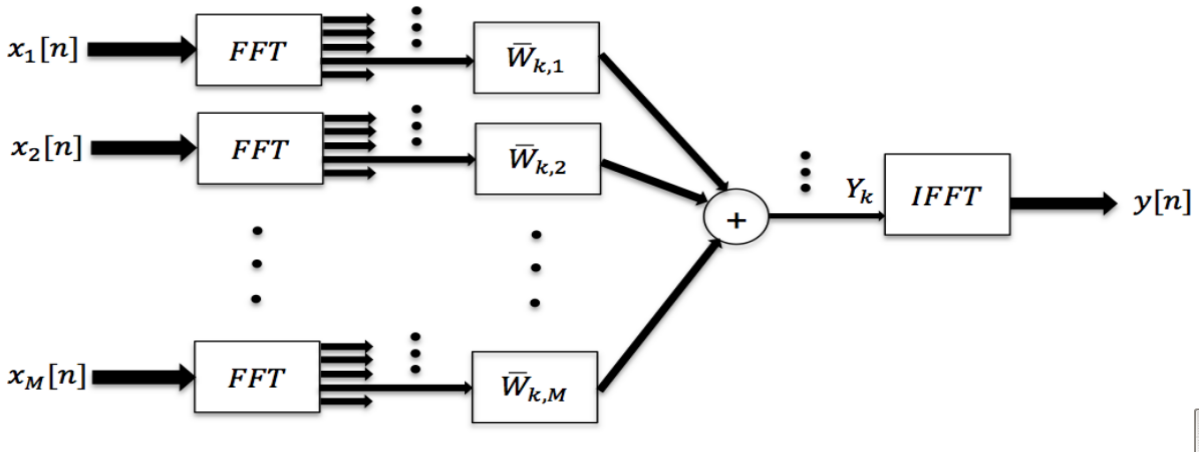


Figure 5.1: *Kurtosis based blind beamformer*

The above algorithm can be extended to N frequency bins by passing each of input channels through an N-point FFT and performing the above process over each of the frequency bins. We take an IFFT of the $y_k's$ obtained at the N subbands to get the beamformed output. 5.1 represents a general structure of a kurtosis based blind beamformer.

5.3 Evaluation of algorithm

The input samples taken for testing are the same samples used in delay sum beamformer with kurtosis maximization and without kurtosis maximization. Each of the input samples were split into 512 frequency bins using a 512-point FFT. Upon blind beamforming we see that artificial artefacts have been introduced and that the SNR values have not shown much of an improvement. These artefacts are produced because each subband tries to maximize the kurtosis at every instant, and this occurs even in those subbands which do not have any speech content. The SNR of the blind beamformer outputs have been plotted in 5.2.

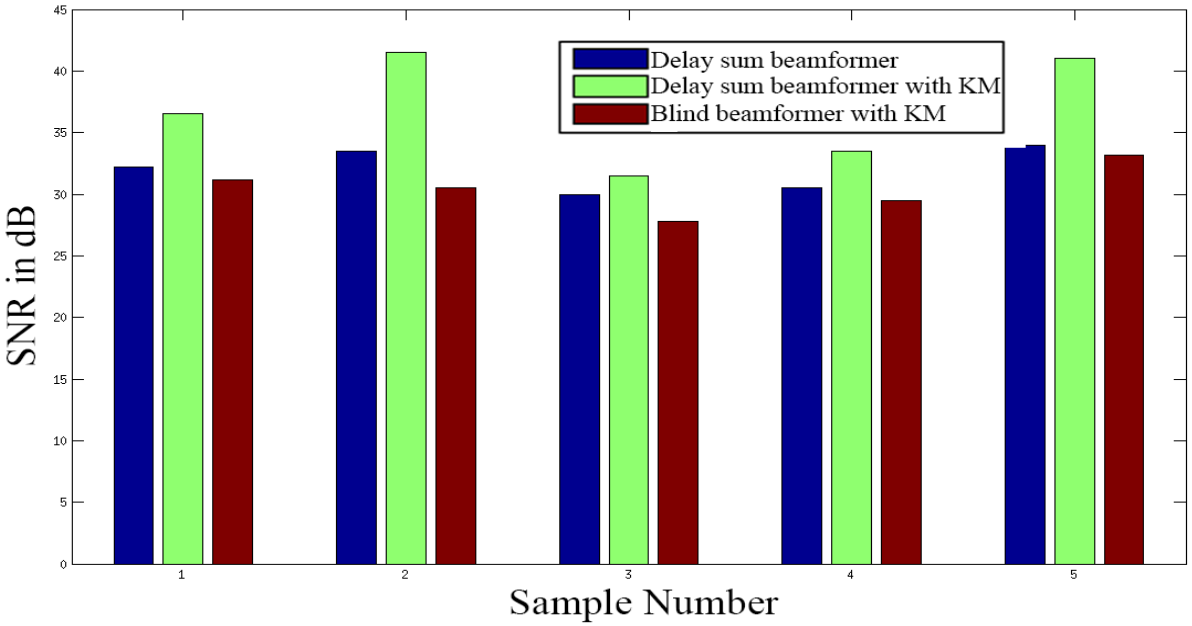


Figure 5.2: *SNR values of various beamformer systems*

CHAPTER 6
Integrated Blind Beamforming (with Energy Maximization)

6.1 Introduction

The Blind Beamforming algorithm as mentioned in Chapter 5 was implemented in frequency domain. The major drawback with this algorithm was that it introduced artefacts which deteriorated the intelligibility of the beamformed output. One of the reason behind the introduction of such artefacts was because of maximization of kurtosis even in those sub-bands where no speech content was present. In this chapter, we propose an algorithm overcomes the above said issue by performing blind beamforming in time domain.

6.2 N-tap filter Implementation

A generic structure of a multi-channel beamforming system is shown in figure 6.1 where a single weight is associated with each of the channels.

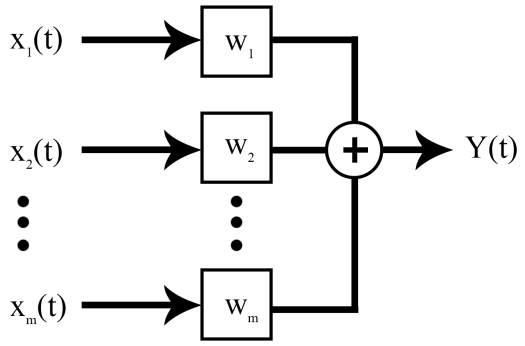


Figure 6.1: *1-tap filter implementation*

In order to re-create a system which is similar in function to a TDOA structure, we have extended a 1-tap filter to an N-tap filter as shown in figure 6.2

The motivation behind the N-tap filter implementation can be understood better with the following example: For a 3 channel system, at time instant 'n', suppose the delay values in equation 3.1 are 0, 2 and 3 for each of the channels. This will correspond to a higher value of weight to the 0^{th} tap, 2^{nd} tap and 3^{rd} tap for each of the channels respectively.

6.3 Beamforming with Energy Maximization

The objective of this approach is to find w_{opt} such that it maximizes the energy J at the output. We also constrain $||w||_2^2$=1 so as to make sure that $w^H x$ does not grow out of

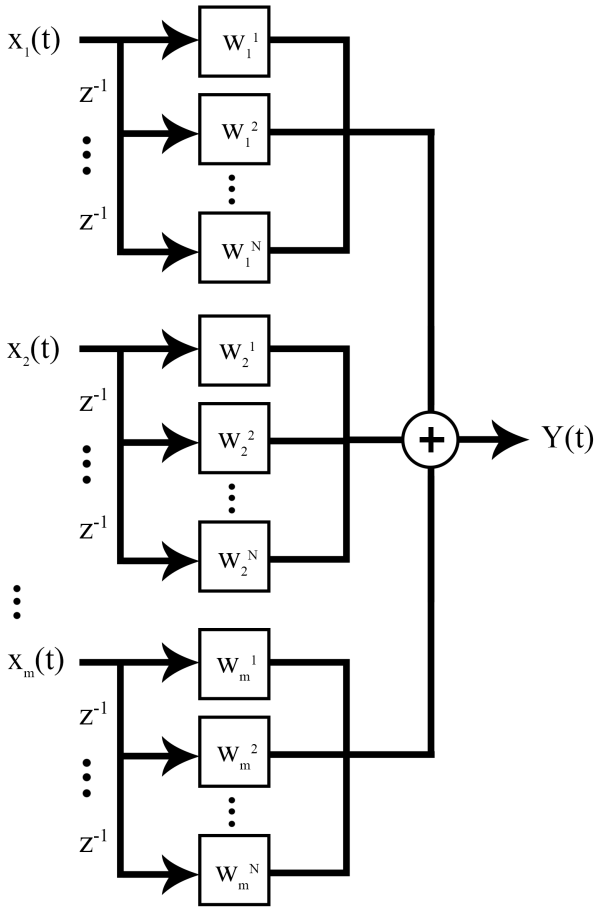


Figure 6.2: *N-tap filter implementation*

bounds. A gradient ascent method is then employed to numerically find a local maxima.

Let w represent the weight vector and let x represent the input vector. Then the output of the beamformer is given by

$$y = w^H x \tag{6.1}$$

The energy of the output is then given by

$$J = w^H x x^H w = y y^H \tag{6.2}$$

The gradient of J is then computed using equation 6.3.

$$\frac{\partial J}{\partial \mathbf{w}} = x x^H w \tag{6.3}$$

The gradient thus computed is then added to the current w to obtain the new weights

vector w_{opt} as shown in equation 6.4

$$w_{opt} = w_{old} + \mu \frac{\partial J}{\partial \mathbf{w}} \tag{6.4}$$

where μ is the learning factor.

6.4 Sparse representation of weights vector

Since an N-tap filter is basically an N^{th} order FIR filter, it was observed that employing all the N coefficients (non-zero) for beamforming resulted in an low-pass filtered output. In order to overcome the above effect, we looked at a sparse representation of the weights vector.

In order to get a sparse representation, we modified the energy equation as given in equation 6.5

$$J = w^H x x^H w - \lambda |w|_1 \tag{6.5}$$

The above equation maximizes $w^H x x^H w$ while minimizing $|w|_1$ Considering the differential of equation 6.5 does not exist at $|w|_1$=0, we have looked at the sub-gradient method which overcomes this issue.

The gradient of equation 6.5 is given by

$$\frac{\partial J}{\partial \mathbf{w}} = x x^H w - \lambda \frac{\partial |w|_1}{\partial \mathbf{w}} \tag{6.6}$$

The gradient value thus obtained is used to update the weights vector w_{opt} as given by equation 6.4. The weights vector thus computed is retained only if $J_{new} > J_{previous}$ where J_{new} is evaluated using w_{opt}. Figure 6.3 shows the variation in weights vector obtained using the above stated method.

As observed from figure 6.3, the above method was sparsifying the weights vector to a very large extent such that it chose only 1 tap across all the channels at a given instant, which defeated the purpose of performing beamforming. The output from the above method did not provide any improvements in SNR in comparison to the inputs. Thus, we have proposed an optimal method for selection of weights vector.

6.5 Optimal selection of weights vector

In all our previous experiments, all the channels had same number of taps which led to multiple solutions depending on the initialization of the weights vector. For example, suppose the first channel had a non-zero value at the 5^{th} tap and the other two channels had a non-zero value at 15^{th} tap and 10^{th} tap respectively, we could have another solution where in the first channel had a non-zero value at the 1^{st} tap and the other two channels had a non-zero value at 11^{th} tap and 6^{th} tap respectively. In order to have a unique solution, we set one of the channels as a reference channel by restricting the number of taps in that channel to 1. This means that the reference channels will have 1 tap and the

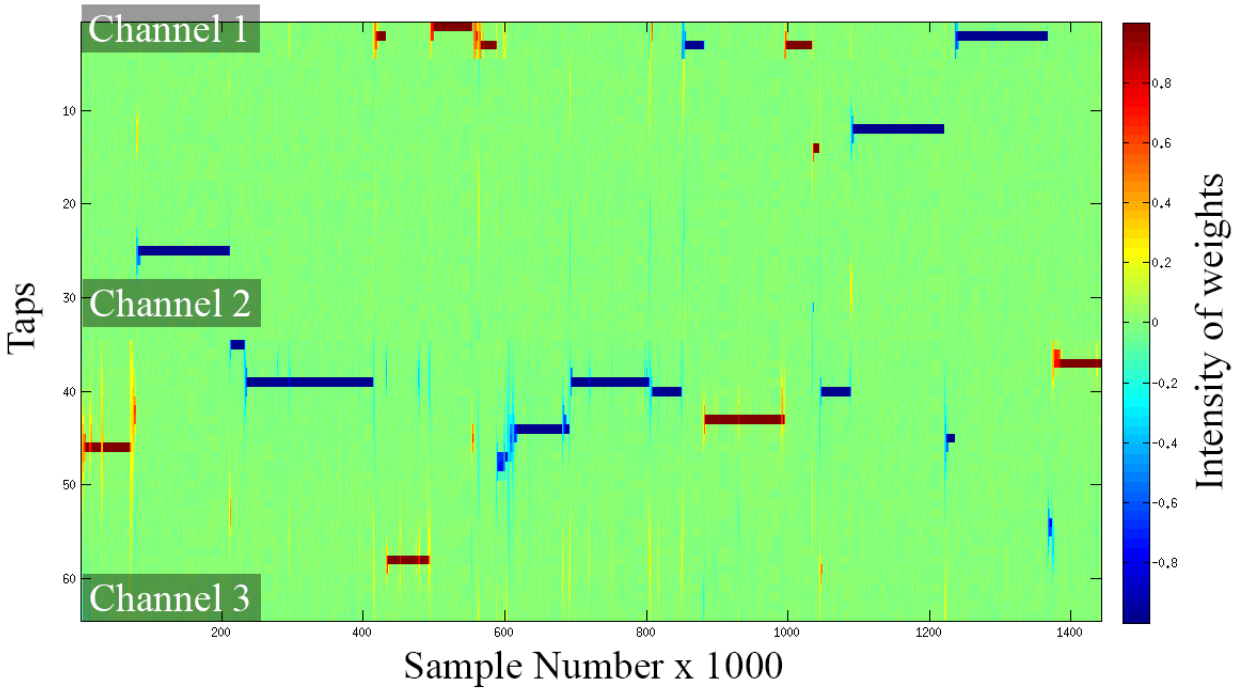


Figure 6.3: *Variations in weights vector (Sub-gradient method)*

rest M-1 channels will have N-tap structure respectively.

Since the main objective was to reduce the number of non-zero coefficients in the weights vector but not equal to single non-zero value across all channels. In the proposed method, we choose 1 peak from the reference channel and maximum 3 peaks from the each of the other M-1 channels respectively.

In an ideal situation for a stationary speaker, we expect the maximum 3 taps to remain constant throughout. But as can be seen from figure 6.4, there are sudden jumps in weights vector from one tap to another. In order to prevent the sudden jumps from one tap to another, we made a modification to smoothen both the weights vector and gradient as given in equations 6.7 and 6.8. Figure 6.5 shows the variation in weights with the smoothening constraint on the weights vector and gradient vector, as expected, we observe that the weights vector has now become more or less constant and is restricted to the same taps throughout.

$$w_{new} = (1 - \beta)w_{old} + \beta w_{current} \tag{6.7}$$

$$\frac{\partial J(n)}{\partial w} = (1 - \alpha)\frac{\partial J(n-1)}{\partial w} + (\alpha)(xx^H w) \tag{6.8}$$

In the above method, there was no restriction on the manner in which the peaks were chosen which meant that the weights vector could be updated to a different energy maxima at every instant. In order to update the weights in the direction of a single local maxima, we consider taking 3 continuous peaks for the M-1 channels and 1 peak for the reference channel.

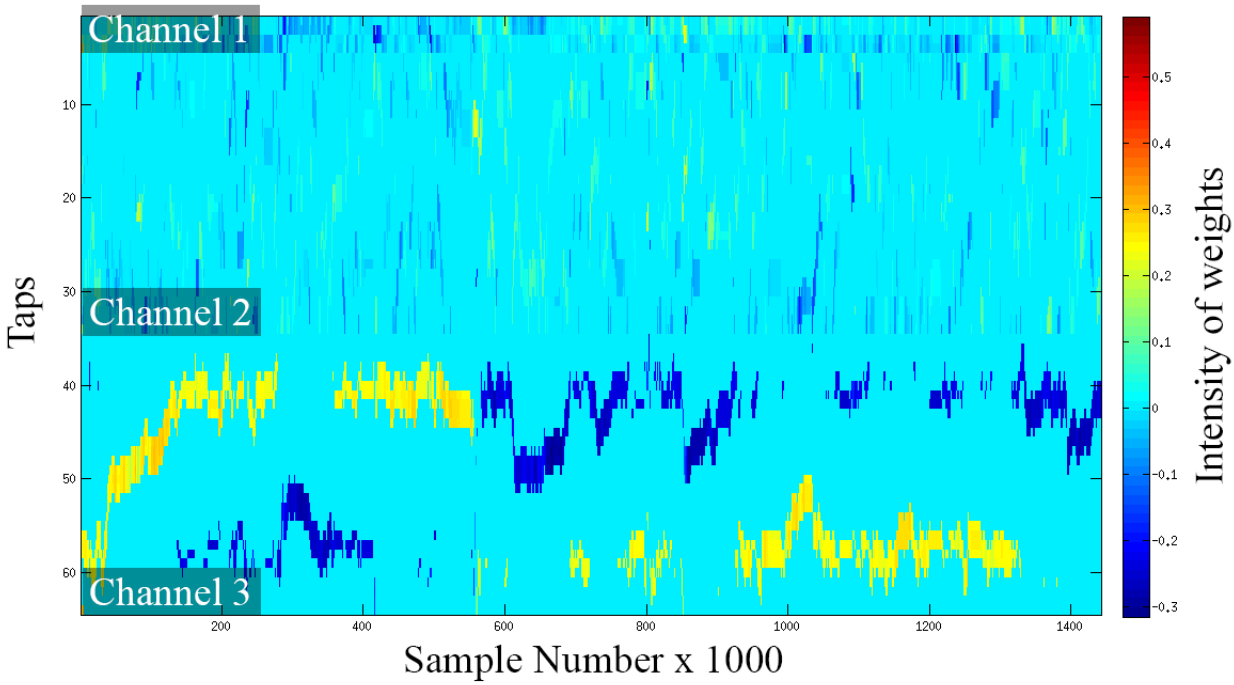


Figure 6.4: *Variations in weights vector (non-smoothed)*

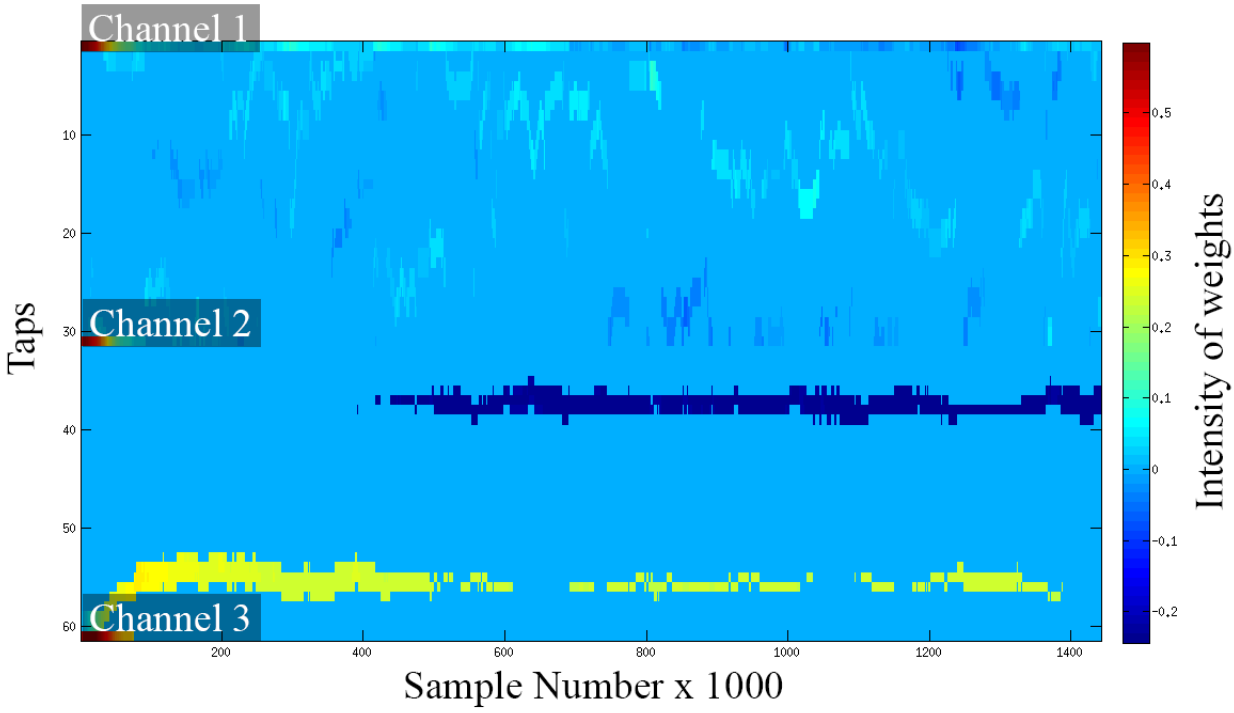


Figure 6.5: *Variations in weights vector (smoothed)*

6.6 Block Processing of weights vector

Since the TDOA method employs a frame based processing, where the frame size is 500ms and the frame shift is 250ms, we have thus made changes to the algorithm so as to implement a frame by frame based processing of the inputs.

In this block processing method instead of updating the weights vector at every time instant, for each block of size 500ms we iteratively find a weights vector that converges and set this weights vector for the whole block. The frame was then shifted by 250ms and a triangular window function was used to add the beamformed outputs across adjacent blocks. Considering block processing smoothens the weights vector inherently, we do not perform further smoothening of weights vector across the different blocks. The figure 6.6 represents the weights vector updated at every instant as mentioned in section 6.3 and we can clearly see that there are random jumps from one tap to another. The figure 6.7 represents the weights vector using block processing method as mentioned above where in there is a very smooth transition of values.

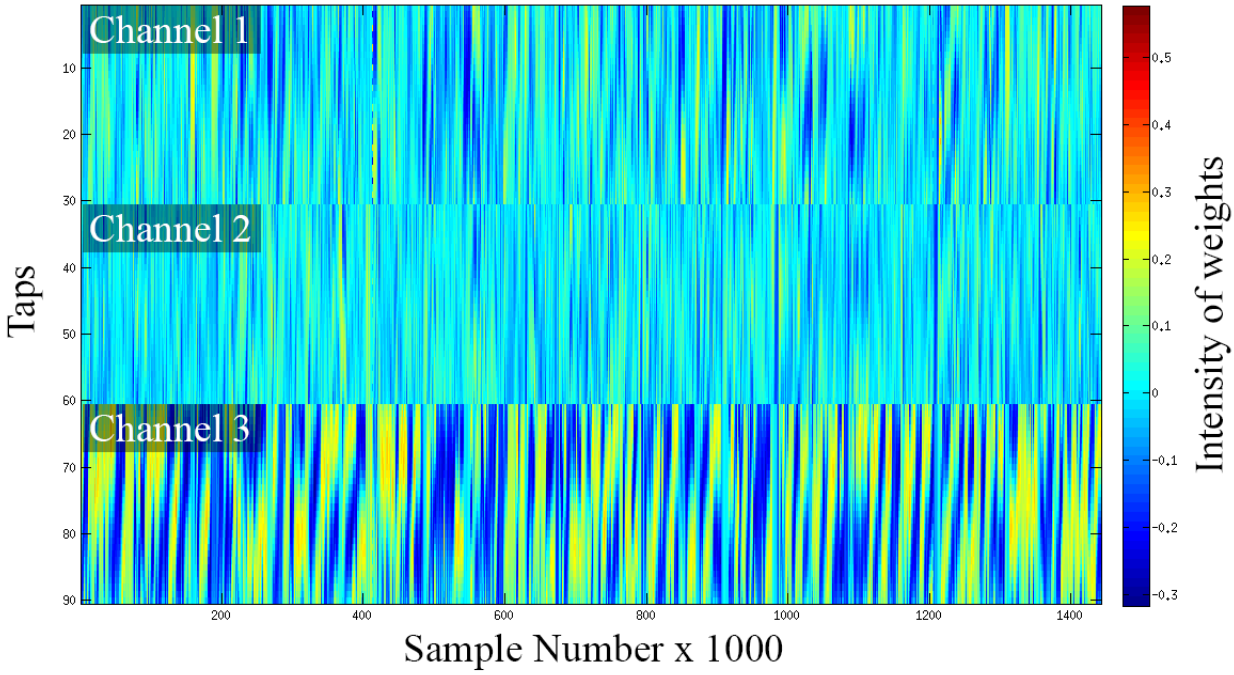


Figure 6.6: *Variations in weights vector (Simple N-Tap filter)*

Finally, under this method we integrated the block processing technique as mentioned above along with the optimal selection of weights vector technique as mentioned in section 6.5. The figure 6.8 represents the weights vector of such a system. It was observed that the perceptual quality of the output obtained by this method was significantly better that the other techniques.

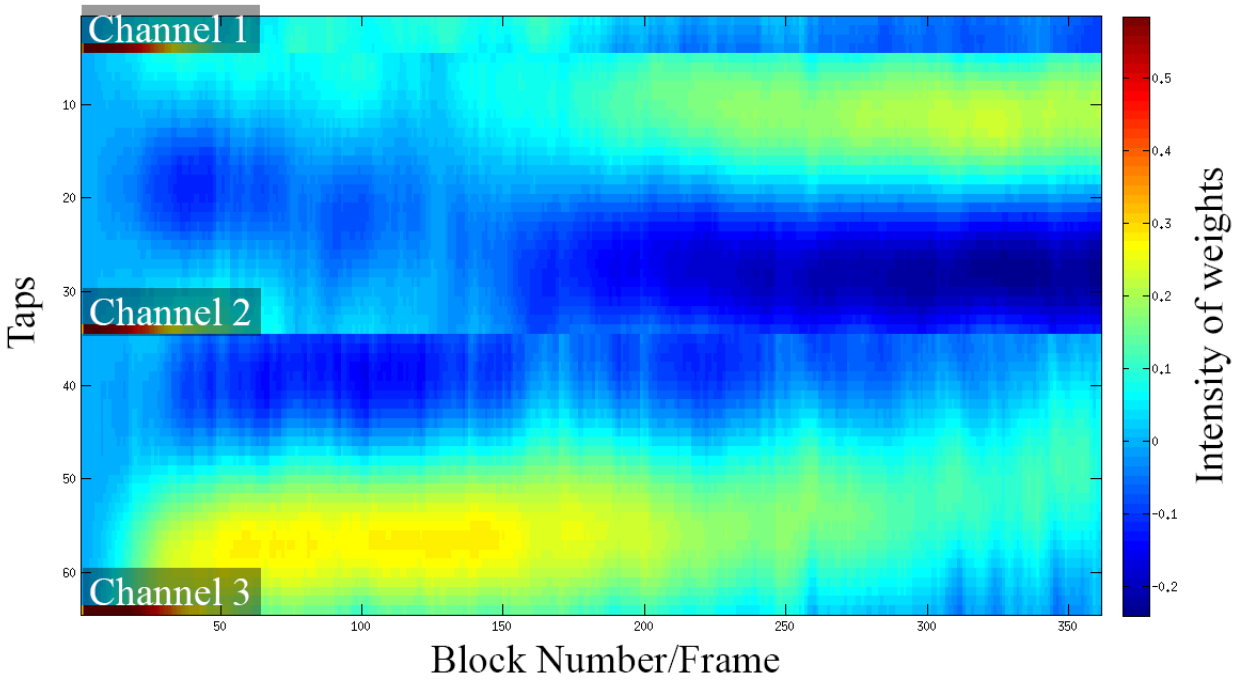


Figure 6.7: *Variations in weights vector (Block processed)*

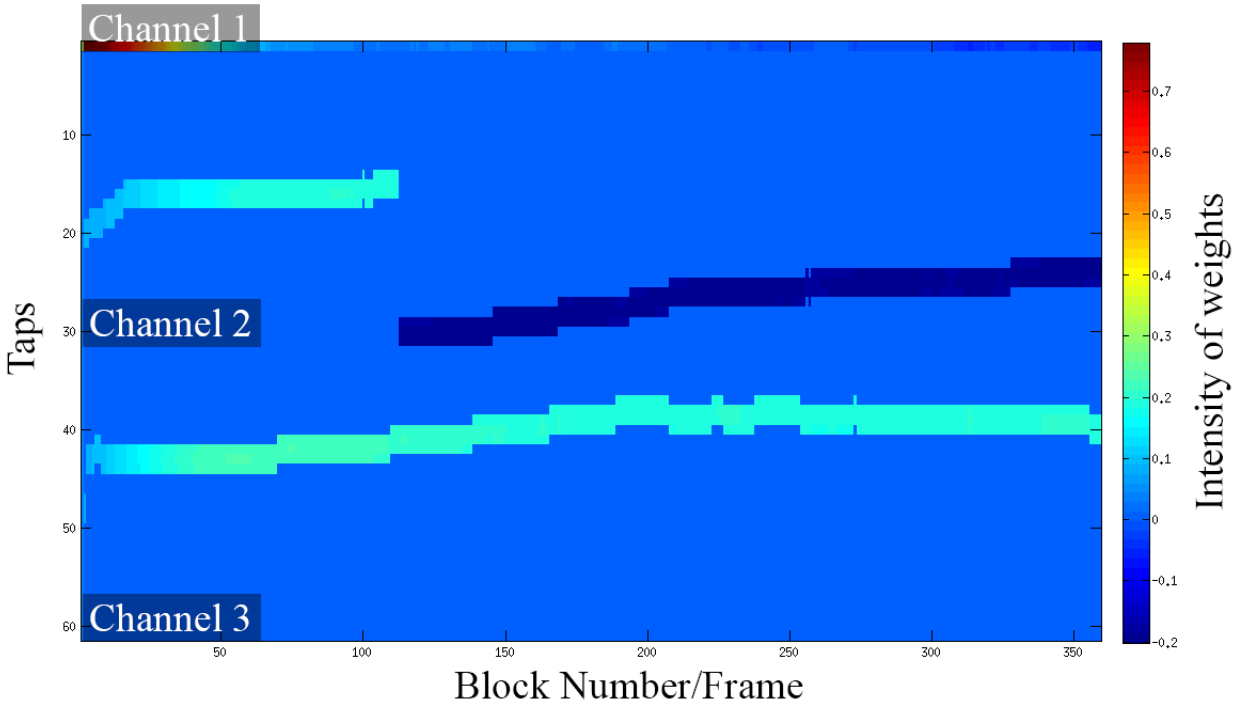


Figure 6.8: *Variations in weights vector (Optimal weights selection + Block processed)*

6.7 Evaluation of the proposed beamformer

Section 6.6 explains the system that performs blind beamforming with energy maximization in time domain. In order to evaluate the above system, audio samples were collected from 5 different sessions where in each of the sessions, 3 microphones were used for recording. Figure 6.9 tabulates the SNRs of the outputs obtained using the proposed speech enhancement system.

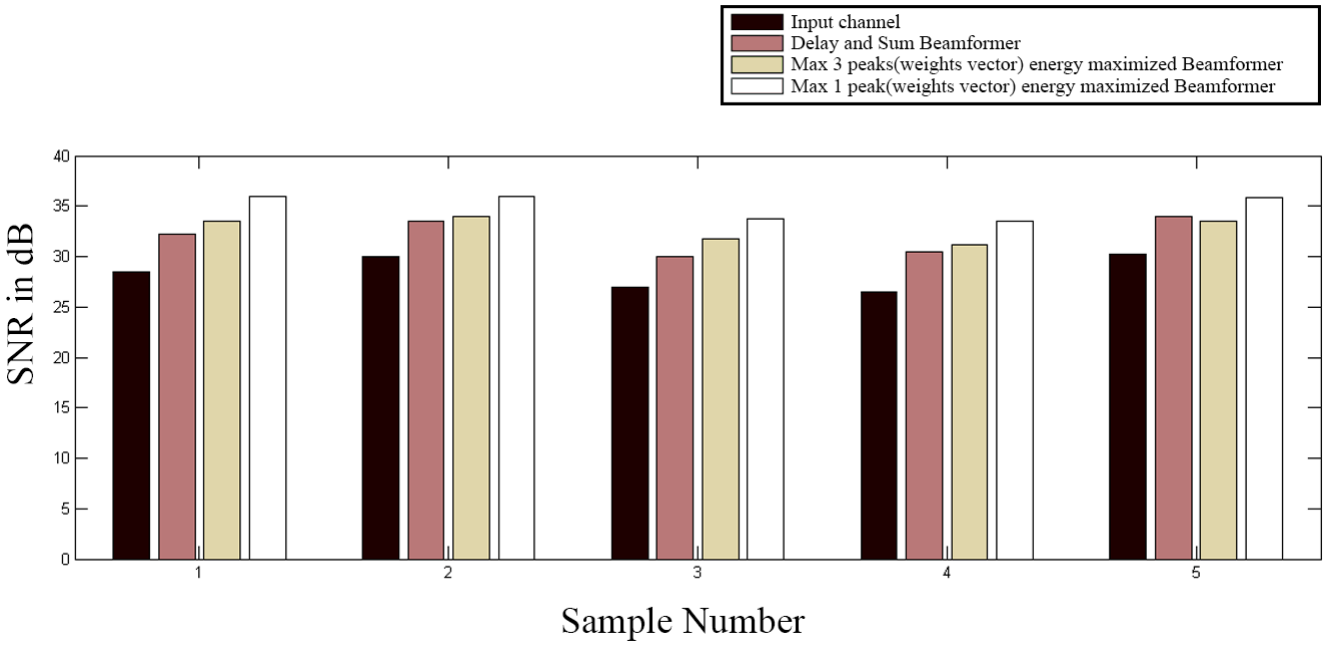


Figure 6.9: *Plot of SNR*

As can be observed from the above figure, the SNRs of the output of proposed system (1 Maximum peak) has an improvement of around 3dB in comparison to the TDOA based beamfomer.

CHAPTER 7

Integrated Blind Beamforming (Kurtosis Maximization)

7.1 Introduction

As already stated, speech has a higher kurtosis content as compared to others. In the previous chapter, we proposed a blind beamformer with a energy maximization constraint. In this chapter, we intend to modify the algorithm mentioned in the previous chapter by maximizing kurtosis instead of energy.

7.2 Approach

The algorithm of this system is very similar to the algorithm mentioned in section 6.3. Here, instead of maximizing energy we will be maximizing kurtosis.

The kurtosis measure of a signal y(n) is given in equation 5.4. The gradient of kurtosis with respect to the reconstruction weights is given in equation 5.9. The gradient thus computed is used to obtain the new weights vector w_{opt} according to equation 6.4

7.3 Evaluation

In order to evaluate the above system, audio samples were collected from 5 different sessions where in each of the sessions, 3 microphones were used for recording. Figure 7.1 tabulates the SNRs of the outputs obtained using the proposed speech enhancement system.

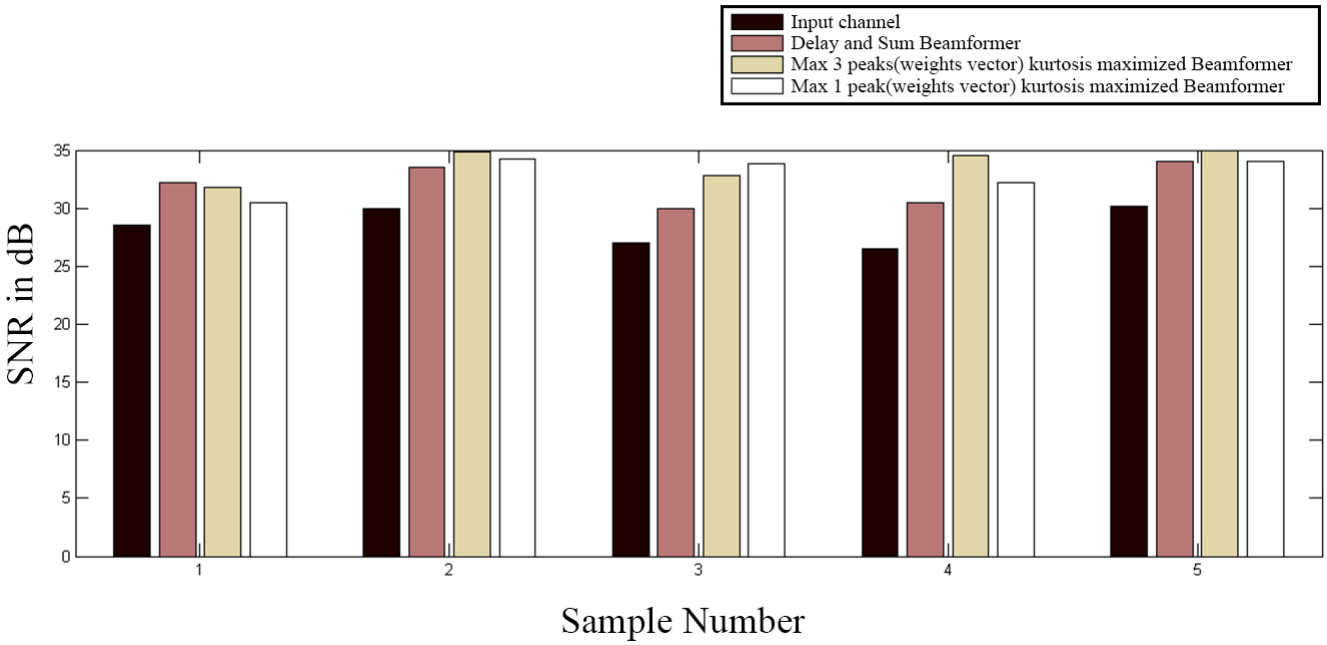


Figure 7.1: *Plot of SNR*

CHAPTER 8

Conclusions

8.1 Analysis

Initially, we passed all our samples through the BeamformIt toolkit in order to set a benchmark value for reference. The outputs obtained from BeamformIt did not have an increase in SNR as compared to all the input channels in a few cases. In order to overcome this issue instead of processing all the channels together, we concentrated on processing the individual channels using a de-reverberation system via maximizing the kurtosis before being passed as inputs to *beamformIt*.

Additionally, a few extra modifications were made on *beamformIt* which further improved the output SNR values of the complete system by approximately 4dB when we chose β=0.9.

We then implemented an RLS based beamformer. This method employs a constraint on the array geometry and the output thus, obtained had a lower SNR in comparison to the Delay & Sum beamformer by around 1dB. In order to remove the constraint on geometry, we implemented a kurtosis based blind beamforming system. It was observed that there was a decrease in SNR by 2-3dB even in this method.

The previous methods tried to maximize kurtosis in the frequency domain and this introduced artefacts in the output. In order to get rid of these artefacts we implemented an integrated blind beamforming algorithm with energy maximization. The main difference between the above algorithm and the kurtosis based blind beamformer was that the above algorithm is implemented in time domain while the latter was implemented in frequency domain. It was observed that the SNR values of the outputs from this algorithm were better in comparison to all the previous stated algorithms by around 3dB.

The integrated blind beamforming algorithm was later modified so that the weights vector was updated with a kurtosis maximization constraint instead of energy maximization. Hence, we have been successful in developing a speech enhancement system which provides a better SNR performance, on an average of 3dB as compared to TDOA beamformer.

8.2 Future Work

We have seen that using a parameter like kurtosis for designing the filter has resulted in a successful speech enhancement system. A dictionary based learning might be helpful since the number of non-zero coefficients in the weights vector are less. Hence, it might be useful to implement a dictionary based algorithm (by using a parameter like kurtosis) for processing the input files before subjecting them to beamforming.

REFERENCES

[1] M. Wolfel and J.McDonough, "*Distant Speech Recognition*", Wiley, 2009.

[2] S. Haykin, "*Adaptive filter theory*", Upper Saddle River, New Kersey: Prentice Hall, 2002.

[3] M.Wolfel, "Enhanced speech features by single-channel joint compensation of noise and reverberation", *IEEE Transactions Audio, Speech and Language Processing*, vol. 17, pp. 312-323, 2009.

[4] Barry D. Van Veen and Kevin M. Buckley, "Beamforming: a versatile approach to spatial filtering", IEEE ASSP Magazine, vol.5, Issue 2, 1998.

[5] L.C. Parra and C.V. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming", *IEEE Transactions on Audio and Speech Processing*, vol.10, pp. 352-362, 2002.

[6] Bradford W. Gillespie , Henrique S. Malvar and Dinei A. F. Florncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering", ICASSP 2001, vol.6, pp. 3701-3704, 2001.

[7] Zohra Yermeche, Benny Sallberg, Nedelko Grbic and Ingvar Claesson, "Real-Time DSP Implementation of a Subband Beamforming Algorithm for Dual Microphone Speech Enhancement", *IEEE International Symposium on Circuits and Systems*, pp. 353-356, 2007.

[8] Wenqing Jiang and Henrique Malvar, "Adaptive Noise Reduction of Speech Signals", Technical Report, Microsoft Research.

[9] Flanagan, J., Johnson, J., Kahn, R. and Elko, G., "Computer-steered microphone arrays for sound transduction in large rooms", Journal of the Acoustic Society of America 78, 15081518, 1994.

[10] X. Anguera, C. Wooters, J. Hernando, "Acoustic Beamforming for Speaker Diarization of Meetings", *IEEE Transactions on Audio, Speech and Language Processing*, vol.15, issue 7, pp. 2011-2022, 2007.

[11] http://labrosa.ee.columbia.edu/projects/snreval/

[12] B. Yegnanarayana and P. Satyanarayana Murthy, "Enhancement of reverberant speech using LP residual signal", *IEEE Transactions on Audio and Speech Processing*, vol. 8, pp. 267-281, 2000.

[13] O. Tanrikulu and A.G. Constantinides, "Least-mean kurtosis: a novel higher-order statistics based adaptive filtering algorithm", *Electronics Letters*, vol. 30, pp. 189190, 1994.

[14] S. Haykin, Adaptive Filter Theory. New Jersey: Prentice- Hall, 1996.

[15] Z. Yermeche, P.Marquez, N. Grbic and I. Claesson, *"A Calibrated Sub-band Beamforming Algorithm for Speech Enhacement"*, published in SAMSP Workshop proceedings, pp. 485-489, USA, 2002

[16] D. Johnson and D. Dudgeon, *Array Signal Processing - Concepts and Techniques*, Prentice Hall, 1993

[17] Daniel C Klinger, "*Kurtosis-based blind beamforming: An adaptive, Sub-band implementation with a convergence improvement"*, MS Thesis report, UIUC, 2013