Adaptive Beamforming with a Microphone Array

[1]Shah Mahdi Hasan, Mohammad Bin Monjil , Farhad Mohsin,
Md. Abul Hayat and [2]A.B.M. Harun-ur Rashid *Member, IEEE*

Department of Electrical & Electronic Engineering
Bangladesh University of Engineering & Technology
Dhaka-1000, Bangladesh
Email: [1]tanvir_dcc@live.com, [2]abmhrashid@eee.buet.ac.bd

***Abstract*—A robust and optimized system architecture has been developed and designed for adaptive beamformer with a Microphone Array. The system includes following subsystems - MMSE STSA Estimator, DOI (Direction of Interest) Estimator and an Adaptive Beamformer. This work is submitted to fulfill the requirement of Cadence Design Contest-2015. The system architecture has been implemented and tested for Xtensa Processor which was configured for HiFi-2 DSP Standard for audio processing.**

***Index Terms*—Adaptive Beamformer, Microphone Array, DOI Estimation, Xtensa, Hifi-2 DSP Standard.**

I. Introduction

Electronically steerable Microphone Arrays have become a rapidly emerging tool in speech data acquisition and processing. One of their prime applications is building adaptive beamformer for the tracking of active talkers while suppressing noise and interferences. Designing an adaptive beamformer primarily includes two important segments- a DOI estimator which localizes the speech source in the receptive area and a beamformer which produces directive gain toward the speech source and suppresses noise and interferences. Designing both of the segments has been considered as a classical and challenging problem because speech signals are wideband and highly non-stationary in nature.

A novel system architecture based on MMSE (Minimum Mean Square Error) STSA (Short Time Spectrum Amplitude) Estimator, MCCC (Multi Channel Cross Correlation) and Frost Adaptive Beamformer has been designed, developed and implemented using Xtensa processor. The system is designed with ample robustness so that it can satisfactorily meet the given requirement stated in the project statement of Cadence Design Contest-2015.

II. Theory

The system architecture includes three subsystems which employs three different algorithms.

A. The Gaussian Based MMSE STSA Estimator For Noise Suppression

In this algorithm the MMSE STSA [1] estimator is derived which is based on modeling speech and noise spectral components as statistically independent Gaussian RV. The reason behind using this method is primarily for dealing with non-stationary microphone noise which is 'Pink Noise' as per project statement. In order to derive the MMSE STSA estimator, the *a priori* probability distribution of the speech and noise Fourier expansion coefficients are assumed, as these are unknown in practice. Let $y(n) = x(n) + d(n)$ be the sampled noisy speech signal consisting of the clean signal $x(n)$ and the noise signal $d(n)$. Taking the short-time Fourier transform of $y(n)$, to have:

$$Y(w_k) = X(w_k) + D(w_k)$$

Where, $w_k = \frac{2\pi k}{N}, k = 0, 1, 2, \cdots, N-1$ and N is the frame length. The above equation can also be expressed in polar form as

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)}$$

As, the spectral components are assumed to be statistically independent, the MMSE amplitude estimator $\hat{X}_k$ can be derived from $Y(w_k)$ only. That is,

$$\begin{aligned}\hat{X}_k &= \mathbb{E}\{X_k \mid Y(w_0), Y(w_1), \cdots\} \\ &= \mathbb{E}\{X_k \mid Y(w_k)\} \\ &= \frac{\int_0^\infty \int_0^{2\pi} x_k p(Y(w_k) \mid x_k, \theta_k) p(x_k, \theta_k) d\theta_k dx_k}{\int_0^\infty \int_0^{2\pi} p(Y(w_k) \mid x_k, \theta_k) p(x_k, \theta_k) d\theta_k dx_k}\end{aligned}$$

where $\theta_k = \theta_x(k)$. Under the assumed Gaussian model $p(Y(w_k) \mid x_k, \theta_k)$ and $p(x_k, \theta_k)$ are given by

$$p(Y(w_k) \mid x_k, \theta_k) = \frac{1}{\pi\lambda_d(k)} exp\{-\frac{1}{\lambda_d(k)} |Y_k - X_k e^{j\theta_x(k)}|^2\}$$

$$p(x_k, \theta_k) = \frac{1}{\pi\lambda_d(k)} exp\{-\frac{X_k^2}{\lambda_d(k)}\}$$

Where, $\lambda_x(k) \triangleq \mathbb{E}\{|X_k|^2\}$, and $\lambda_d(k) \triangleq \mathbb{E}\{|D_k|^2\}$ are variences of the k^{th} spectral component of the speech and noise the equation gives

$$\tilde{X}_k = \Gamma(1.5) \frac{\sqrt{v_k}}{\gamma_k} \exp\{-\frac{v_k}{2}\} \left[(1 + v_k) I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})\right]$$

Where $\Gamma(\cdot)$ donates the Gamma function and $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel function of zero and first order,respectively. The variable, v_k is defined by

$$v_k \triangleq \frac{\xi_k}{1 + \xi_k} \gamma_k$$

Where ξ_k and γ_k are interpreted as the a priori and a posteriori signal-to-noise ratio (SNR), respectively and are defined by

$$\xi_k \triangleq \frac{\lambda_x(k)}{\lambda_d(k)}$$
$$\gamma_k \triangleq \frac{Y_k^2}{\lambda_d(k)}$$

At high SNR, $\xi_k \gg 1$ and $\gamma_k \gg 1$; therefore, the estimator can be simplified as

$$\tilde{X}_k \triangleq \frac{\xi_k}{1+\xi_k}\gamma_k$$

The above is called Weiner estimator.

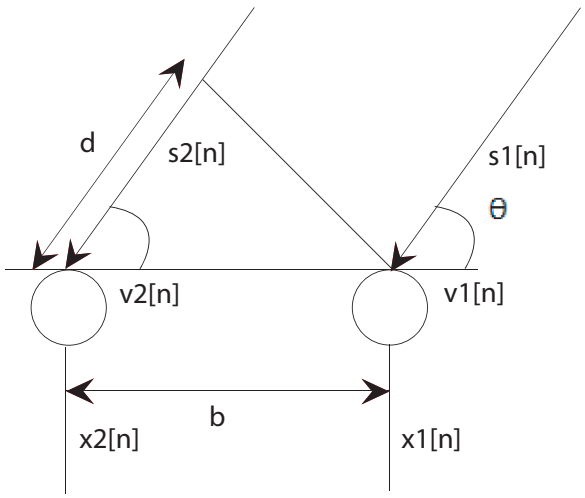

Fig. 1: TDOA Estimation

B. DOI Estimator Subsystem

The problem of finding direction of arrival of a speech source can be solved by estimating TDOA (Time Delay of Arrival) estimation. In the simple case of two microphones, the TDOA estimation is depicted in Fig.1. In this figure the two dimensional geometry of TDOA estimation problem is portrayed. The signal $s_1[n]$ and $s_2[n]$ are speech signals including ambient noise. The $v_1[n]$ and $v_2[n]$ are microphone noise, which is in our case is pink noise. The noises are assumed uncorrelated with both speech signals and noises at other microphones. The acoustical path difference experienced by speech signals is therefore $d = b\cos\theta$, where θ is the incident angle as well as the direction of source. Now knowing the microphone spacing b and determining d using TDOA (Time Delay of Arrival) estimation, the incident angle θ can be calculated.

For the estimation of TDOA we use Multi Channel Cross Correlation Method (MCCC) [2]. The microphone array consists of L microphones in a linear equidistantly spaced array, from the 1^{st} to the L^{th} microphone.The delay between the 1^{st} and the L^{th} microphones is then given by

$$f_l = (l-1)\tau$$

where τ is the time delay between two neighboring microphones.

For the application of the MCCC algorithm, we consider the column vector of the aligned signals at the L microphones

$$\boldsymbol{x}_{1:L}[n - f_L(m)] = [x_1[n - f_L(m) + f_1(m)] \\ x_2[n - f_L(m) + f_2(m)] \cdots x_L[n]]$$

with $m/f_s = \hat{\tau}$ as a guess for the delay, where f_s is the sampling frequency. The corresponding spatial correlation matrix of the microphone signals is then

$$\boldsymbol{R}_{m,1:L} = \mathbb{E}\{\boldsymbol{x}_{1:L}[n - f_L(m)] \cdot \boldsymbol{x}_{1:L}^T[n - f_L(m)]\} \\ = \begin{bmatrix} r_{m,11} & \cdots & r_{m,1L} \\ \vdots & \ddots & \vdots \\ r_{m,L1} & \cdots & r_{m,LL} \end{bmatrix}$$

where the cross-correlation between the two signals $x_k[n - f_l(m)]$ and $x_l[n - f_k(m)]$ is given by

$$r_{m,kl} = \mathbb{E}\{x_k[n - f_l(m)]x_l[n - f_k(m)]\}$$

The spatial correlation matrix $\boldsymbol{R}_{m,1L}$ can be factored as

$$\boldsymbol{R}_{m,1:L} = \boldsymbol{D}\tilde{\boldsymbol{R}}_{\boldsymbol{m,1:L}}\boldsymbol{D}$$

with the diagonal matrix

$$\boldsymbol{D} = \begin{bmatrix} \sqrt{\mathbb{E}\{x_1^2[n]\}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\mathbb{E}\{x_1^2[n]\}} \end{bmatrix}$$

the symmetric matrix

$$\tilde{\boldsymbol{R}}_{m,1:L} = \begin{bmatrix} 1 & \cdots & \rho_{m,1L} \\ \vdots & \ddots & \vdots \\ \rho_{m,L1} & \cdots & 1 \end{bmatrix},$$

and the cross-correlation coefficients between $x_k[n - f_l(m)]$ and $x_l[n - f_k(m)]$

$$\rho_{m,kl} = \frac{\mathbb{E}\{x_k[n - f_l(m)]x_l[n - f_k(m)]\}}{\sqrt{\mathbb{E}\{x_k^2[n]\}\mathbb{E}\{x_l^2[n]\}}} \qquad (1)$$

with k and $l = 1, 2, \cdots, L$. In the case of two microphones, the two-channel cross-correlation coefficient is given by

$$\rho_{m,12}^2 = 1 - \det \tilde{\boldsymbol{R}}_{m,1:2}$$

Similarly, the multichannel cross-correlation coefficient is defined as [1]

$$\rho_{m,1L}^2 = 1 - \det \tilde{\boldsymbol{R}}_{m,1:L}$$

The delay estimation is then based on maximizing the cross-correlation coefficient $\rho_{m,1:L}^2$ or by minimizing the determinant of the matrix $\tilde{R}_{m,1L}$ with respect to the guessed delay m.

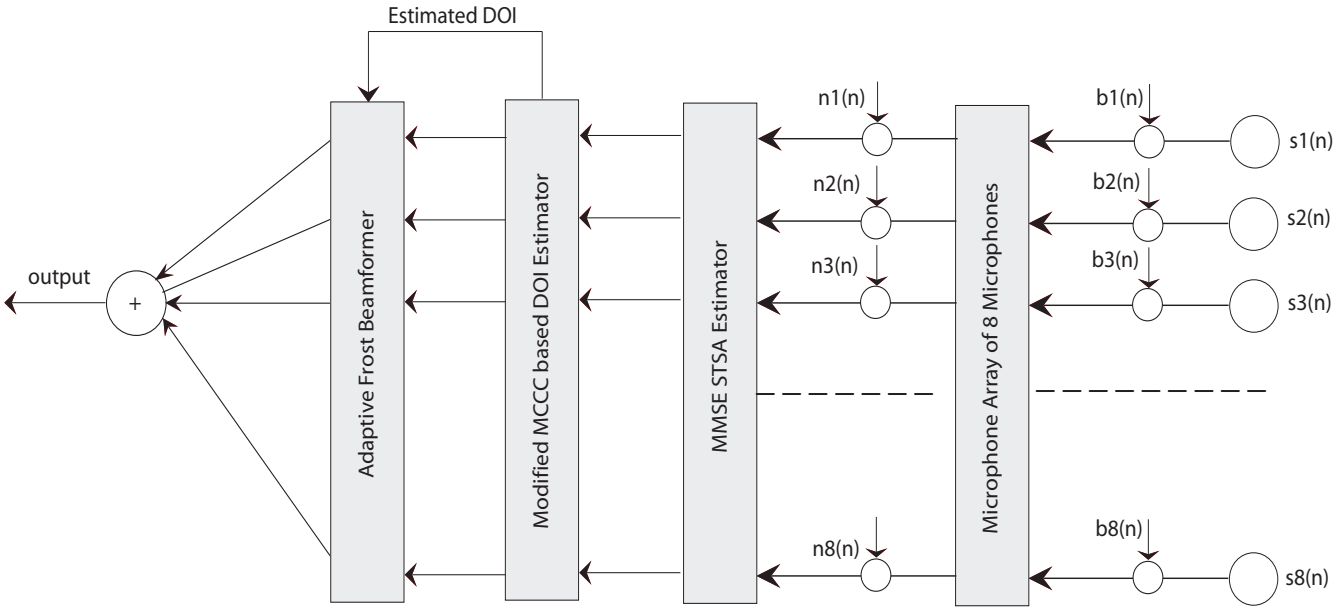

Fig. 2: System Architecture

C. The Adaptive Frost Beamformer

Frost beamformer [3] is conventional beamforming algorithm for wideband signal e.g. speech. In our case we have assumed that at 0° the beamformer has gain of 1 and at other direction is minimizes the power. At the beginning, we set the weight vector to

$$\boldsymbol{w}[0] = \boldsymbol{C}(\boldsymbol{C}^H\boldsymbol{C})^{-1}\boldsymbol{f}$$

for initialization, which satisfies the constraint $\boldsymbol{C}^H\boldsymbol{w} = \boldsymbol{f}$. At each iteration, the vector $\boldsymbol{w}$ is updated in the direction of the negative gradient by a step proportional to a scaling factor μ according to

$$\boldsymbol{w}[n+1] = \boldsymbol{w}[n] - \mu(\boldsymbol{R}_{xx}\boldsymbol{w}[n] + \boldsymbol{C}\boldsymbol{\lambda}[\boldsymbol{n}])$$

Since, $\boldsymbol{w}[n+1]$ must satisfy the constraint, we can substitute this Equation into $\boldsymbol{C}^H\boldsymbol{w} = \boldsymbol{f}$ and solve for the Lagrange multipliers $\boldsymbol{\lambda}[\boldsymbol{n}]$.Then we substitute $\boldsymbol{\lambda}[\boldsymbol{n}]$ into the iteration equation and arrive at

$$\begin{aligned}\boldsymbol{w}[n+1] = \boldsymbol{w}[n] - \mu(\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{C}^H\boldsymbol{C})^{-1}\boldsymbol{R}_{xx}\boldsymbol{w}[n] \\ + \boldsymbol{C}(\boldsymbol{C}^H\boldsymbol{C})^{-1}(\boldsymbol{f} - \boldsymbol{C}^H\boldsymbol{w}[n])\end{aligned}$$

now defining the short-hand of $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{C}(\boldsymbol{C}^H\boldsymbol{C})^{-1}\boldsymbol{C}^H$, the algorithm in Equation now can be rewritten as

$$\boldsymbol{w}[n+1] = \boldsymbol{C}(\boldsymbol{C}^H\boldsymbol{C})^{-1}\boldsymbol{f} + \boldsymbol{P}(\boldsymbol{w}[n] - \mu\boldsymbol{R}_{xx}\boldsymbol{w}[n])$$

Not knowing the true second order statistics $\boldsymbol{R}_{xx}$, the correlation matrix can be replaced by its simple approximation $\boldsymbol{R}_{xx} = \boldsymbol{x}\boldsymbol{x}^H$. This results in the minimization of the instantaneous square error rather than the mean square error, and leads to the following stochastic constrained algorithm

$$\boldsymbol{w}[n+1] = \boldsymbol{C}(\boldsymbol{C}^H\boldsymbol{C})^{-1}\boldsymbol{f} + \boldsymbol{P}(\boldsymbol{w}[n] - \mu e^*[n]\boldsymbol{x}[n])$$

which is also known as the Frost's algorithm. To increase robustness against DOI error a small quantity to the diagonal elements of the estimated covariance matrix is added and the modified update equation becomes

$$\boldsymbol{w}[n+1] = \boldsymbol{C}(\boldsymbol{C}^H\boldsymbol{C})^{-1}\boldsymbol{f} + \boldsymbol{P}(\boldsymbol{w}[n] - \mu e^*[n]\boldsymbol{x}[n] - \mu\delta\boldsymbol{w}[n])$$

where δ is the diagonal loading factor.

III. Innovation in the Algorithm

A. Modification in MCCC

The aforementioned MCCC algorithm used for TDOA estimation implies a poor resolution in determining DOI because of some inherent physical constraints in the system. These constraints are primarily imposed by required inter-microphone distance in the array. The required inter-microphone spacing comes from both system specification and for avoiding spatial aliasing. Moreover fractional delay between microphones is most common phenomena in the system. The fundamental inability of MCCC to bring the role of fractional delay makes it a poor choice for high resolution TDOA estimation.

For overcoming these constraints and ensuring high resolution DOI estimation, we bring a modification in MCCC. This is done by applying interpolation on cross correlation coefficient. To accomplish the most perfect interpolation we choose the *'Not a Knot spline'* interpolation method. There are two reasons behind choosing this interpolation. Firstly, for speech signals which are highly non-stationary in nature, it requires higher order statistics to reveal useful information for interpolation. Secondly, it can resemble the continuous time auto-correlation sequence with minimum mean square error. Moreover, the computational complexity is also moderate for this very particular method of interpolation. Hence the frac-

tional TDOA estimation becomes more accurate and precise with low convergence time.

B. MMSE STSA

The conventional MMSE STSA requires voice activity detection to estimate silent zones. As the noise can vary over time so it tries to detect the silent zones and thus update the noise parameters accordingly. In our case we have $1/f$ noise which is fixed for all the time, so we have estimated the noise parameters only at the starting 1s taking it as a silent zone.

IV. System Architecture

The system works on window by window basis. For windowing the speech samples we use hamming window. The reason behind using this particular window is its ability of reducing the correlation between widely separated spectral components. For precise DOI estimation 40 percent overlap between two consecutive windows are allowed. We chose window length to be 512. Choosing this length ensures faster performance in FFT, short time stationary constraint of speech signal and good estimate of covariance in MCCC algorithm.

After windowing each of the 8 channel with noisy signal, FFT is taken. Phase of these FFT values are saved and the magnitude are sentto MMSE STSA subsystem. Taking IFFT of the estimated amplitude combined with previously saved phase, enhanced signal is obtained. This step minimizes the $1/f$ noise generated at each of the microphone channel.

The enhanced signals go into MCCC block for DOI estimation. FFT is used here to for faster determination of covariance coefficients. DOI of the desired signal is estimated here.

Using the estimated DOI, delays required for the each of the channels are calculated. The signals are then appropriately delayed to perform the pre-steering. As a result signals coming from DOI are now in phase in all 8 channels.

Pre-steered signals are feed into beamforming block. The constraint matrix and vector was constructed such that the beam former will minimize power and maintain unity response towards broadside. As the desired source is now at broadside(due to pre steering) , the frost beamformer adaptively enhances the desired source and suppresses the interferences.

V. Test Setup

We test the system at various noises and interferences to justify its robustness and response to various real time scenarios. The test setup is configured as per our system design and requirement which is given below

1) *Number of microphones in the array:* 8 (specified as design criteria)
2) *Length of array:* 35 cm (specified as design criteria)
3) *Spacing between microphones:* $d = 5$ cm (uniformly spaced microphones)
4) *Talkers:* Two male talkers, one female talker. All of them are simultaneously active in the test scenario. We randomly choose a main talker and consider the others as randomly put discrete sources of interference. The interferences can be up to -3 dB in power of the main talker.
5) *Noise:* Two types of additive noises are present in the system. One is stationary pink noise, which can be up to -3 dB of Microphone signals. Another is ambient white Gaussian noise, which can be up to -10 dB of Speech signals.

The Processor was configured in the built in HiFi-2 [4] standard mode. To simulate the results in the Xtensa Instruction Set Simulator (ISS), we opted for the '*Cycle Accurate Mode'* since it gives the best estimate of performance for Hardware implementation. However, '*TurboXim'* mode was used for input and output periods; profiling was also turned off for this period.

The code was run on one window of 512 samples. Each such run gives corrected output for 307 new samples. The code takes input from three text files. One text file carries the new window data, one carries forwarded data from the last window, and one carries noise estimate from the initial silent zone.

VI. Results & Discussions

In this section we discuss performance of different blocks of our system. At first performance of MCCC block to properly identity DOI of the desired signal is evaluated. We run a simulation with three speakers, one being the main talker and other two are interferences. The main speaker and the interferences are at 60, 110 and 150 degree. The simulation was run using speech signal of 12 seconds, using window of 1024 we got a total of 335 windows. Following table shows total number of different DOI's estimated by our MCC block within the resolution of 15 degrees for different interference power. When no speaker is active in a particular window, $90°$ is found as the DOI.

TABLE I

Interference Average Power (below main talker, in dB)	$60°$	$110°$	$150°$	$90°$
3	85	41	48	88
4	93	32	43	88
5	101	33	35	92
6	122	32	26	97
7	128	29	25	102

When the main talker is not speaking but any other interference is active, the MCCC identifies that interference as the main speaker as the power of that interference is maximum in that window. As the simulation is run using interference with average power below the main speaker, there occurs many windows where interference has greater power that the main speaker and as a result the resultant DOI points at that direction.

Now we show the tracking capability of our algorithm. A simulation was run for 12 second where the main talker

changes angular position with time. It is in 60°, 110° and 160° respectively and no other source is present in this simulation. As seen from Fig. 3, our algorithm can successfully track the main talker when it changes its angular position.

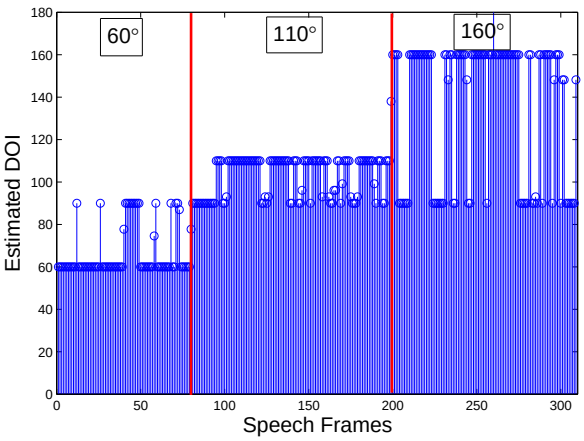

Fig. 3: Tracking capability of DOI estimator

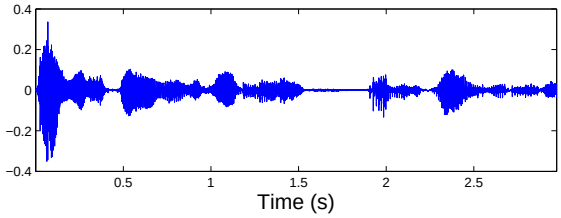

Fig. 4: Main Talker

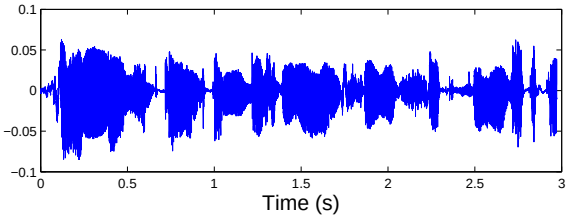

Fig. 5: Interference #1

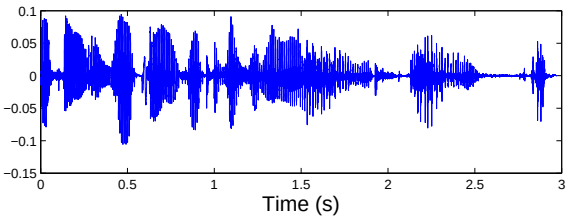

Fig. 6: Interference #2

Now we will show the enhancement of signal due to beamforming. A 3 second frame of main talker, two interference,

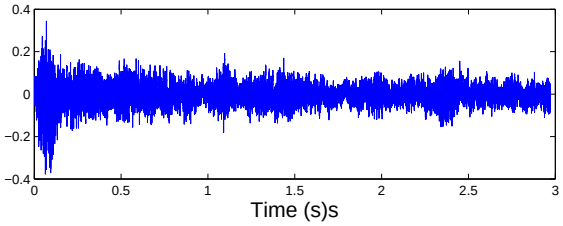

Fig. 7: Signal at Microphone

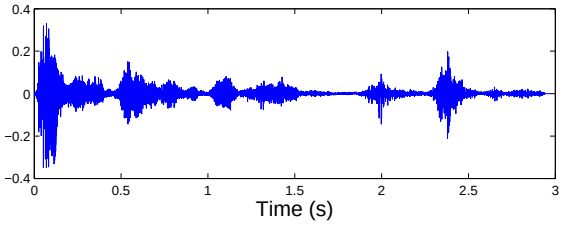

Fig. 8: Final Output

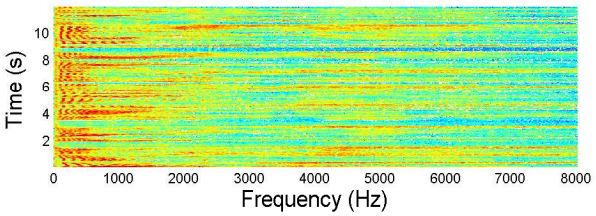

Fig. 9: Clean speech signal

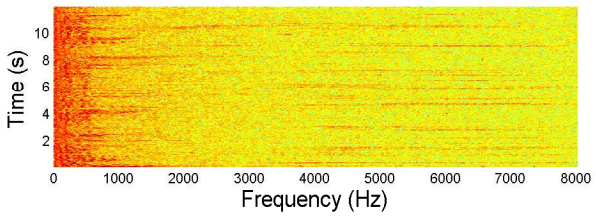

Fig. 10: Signal at microphone

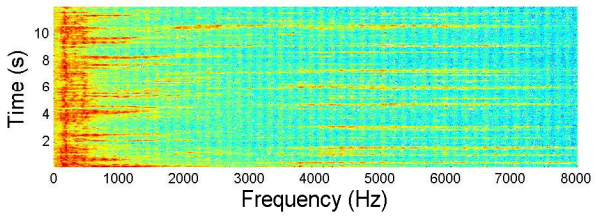

Fig. 11: Final output

signal at 4^{th} microphone channel and output of beamformer is given in Fig. 4, 5, 6, 7 & 8 respectively.

As can be seen from the signal at microphone channel, the signal of the main talker is buried in interference and noise. After beamforming the output signal closely resembles the original main signal. Noise and interference have been suppressed. Although a little of distortion can be seen, this signal is almost identical to the original signal when the audio is played. This little amount of distortion is introduced at MMSE STA block which is used to remove the $1/f$ noise of microphone.

For further analysis of our system's performance in frequency domain we present the spectrogram of signal at various stages of the system. Fig. 9, 10 & 11 shows the spectrogram of clean speech signal, signal at microphone & final output of beamformer respectively. The spectrograms clearly reveals the SNR improvement of the target speech signal and precision of DOI estimator & the beamformer.

From the profile of one such run in Xtensa, we get that the total number of committed instructions were $5,123,135,636$ requiring a total of $7,108,774,623$ cycles. Some data from the profile is presented below in Table II.

TABLE II

Operation	Number of Cycles	% of Cycles
[TOTAL]	7,108,774,623	100
_muldf3 & _muldf3_aux	1,911,132,210	26.87
_adddf3 & _adddf3_aux	1,041,984,429	14.64
divdf3	366,978,850	5.16
subdf3	754,333,195	10.61

The arithmetic operations thus require more than a fair share of cycles/instructions. Aforementioned performance does not make use of Hifi-2 Audio Engines built-in functions and the optimization occurs in the algorithm portion only. HIFi-2 processor's standard configuration may handle high speed audio signal processing norms like multiplication and addition of 24-bit numbers efficiently in parallel process.

VII. Conclusions

As per project statement, we have designed a complete Adaptive beamformer, simulated and evaluated its performance in Xtensa Xplorer. During the course of project development we did bring some major modifications and optimizations in conventional speech processing algorithms which are highly specific for the system. We are working on further optimization to drastically reduce the required cycle number exploiting the Hifi-2 DSP Standard's features. We already have configured two of subsystems. We are working on the final of the three and expecting a major optimization taking the advantage of Hifi-2 DSP Standard.

VIII. Acknowledgement

The authors would like to thank Cadence Design Systems,India for providing the software & license required for the work.

References

[1] Y. Ephraim et. al., "Speech Enhancement Using a Minimum Mean Square Short-Time Spectral Amplitude Estimator," *IEEE Trans. on Acoustic Speech & Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec 1984.

[2] J. Benesty et. al., "Microphone Array Signal Processing",Springer, 2008.

[3] O. L. Frost, "An algorithm for linear constrained adaptive array processing,"*Proceedings of the IEEE*, Vol. 60, Number 8, Aug. 1972, pp. 925-935.

[4] Users Guide, "HiFi 2/EPAudio Engine", *Cadence*, June 2013.