SIR Beam Selector for Amazon Echo Devices Audio Front-End

Xianxian Zhang, Trausti Kristjansson, Philip Hilmes

Amazon Inc., Sunnyvale, California, USA
{xianxi, traustik, philmes}@amazon.com

Abstract

The Audio Front-End (AFE) is a key component in mitigating acoustic environmental challenges for far-field automatic speech recognition (ASR) on Amazon Echo family of products.

A critical component of the AFE is the Beam Selector, which identifies which beam points to the target user. In this paper, we proposed a new SIR beam selector that utilizes subband-based signal-to-interference ratios to learn the locations of the audio sources and therefore further improve the beam selection accuracy for multi-microphone based AFE system. We analyzed the performance of a Signal to Interference Ratio (SIR) beam selector with a comparison to classic beam selector using the datasets collected under various conditions. This method is evaluated and shown to simultaneously decrease word-error-rate (WER) for speech recognition by up to 46.20% and improve barge-in performance via FRR by up to 39.18%.

Index Terms: Beamforming, Adaptive Noise Canceller, Automatic Speech Recognition, Amazon Echo, Linearly Constrained Minimum Variance, Minimum Variance Distortionless response, Generalized Sidelobe Canceller

1. Introduction

Automatic speech recognition (ASR) systems have progressed to the point where humans can interact with computing devices using speech. However, the distance between a device and the speaker will cause a loss in speech quality and therefore impact the effectiveness of ASR performance. As such, there is a greater need to have reliable voice capture for far-field speech recognition. The launch of Amazon Echo devices prompted the use of far-field ASR in the consumer electronics space, as it allows its users to interact with the device from several meters away by using microphone array processing techniques.

The microphone array processing techniques aim to combine the signals received at individual microphones in such a way that a signal coming from a particular direction is enhanced while signals coming from the other directions are attenuated. The classic array beamforming method is delay-and-sum beamforming (DASB) [1], [2], [3], and is based on applying time shifts to a set of microphone array signals to compensate for the propagation delays in the arrival of the source signal at each microphone. These signals are time-aligned and summed together to form a single output signal. This method is very simple and robust if we know the direction of the speech source and the number of microphones, and microphone spacing is selected appropriately. However, if the source location changes during operation, this method will be less effective due to the mismatch in estimating the delays between the microphones. Another practical problem of DASB is that the theoretical maximum noise attenuation $10log_{10}M$ [4] (where M is the number of the microphones in the array) is not easy to obtain in noise environments due to the small microphone array, since the ambient noise is not entirely uncorrelated and traditional beamforming techniques with small standard arrays do not provide substantial improvement in signal-to-noise ratio (SNR) as compared to a single omnidirectional microphone [5]. The linearly constrained minimum variance (LCMV) beamforming proposed by Frost [6] in 1972 is one of the most promising beamforming techniques for noise reduction in acoustic environments and even for speech enhancement though the channel impulse responses need to be known [7], [8], [9]. The basic idea behind LCMV beamforming is to constrain the response of the beamformer so signals from the direction of interest are passed with specified gain and phase. The weights are chosen to minimize output variance or power subject to the response constraint [2]. By imposing multiple linear constraints, LCMV has the effect of preserving the desired signal while minimizing contributions to the output due to interfering signals and noise arriving from directions other than the direction of interest. As a special case of LCMV family, the minimum variance distortionless response (MVDR) beamforming that uses a single constraint towards the look direction has received more attention in the field. In order to avoid the constrained adaptation of MVDR beamformer suggested by Frost in [6], Griffiths and Jim [10] proposed the generalized sidelobe canceller (GSC) structure that separates the output power minimization and the application of the constraint. While Griffiths and Jim only considered one constraint in [10], it was later shown in [11] that the GSC structure can also be used in the case of multiple constraints.

The baseline system that we are considering in this study is based on highly specialized MVDR beamformers. Each beamformer is able to enhance the signals from a particular look direction and suppress the noise and interfering signals from the directions other than look direction. We referred the enhanced signals as beams here, with each beam pointing to its own look direction. In addition, we adopted a beam selector to analyze the beam signals and decide which beam will be used for backend processing, Wakeup Word (WW) and ASR engines. The wakeup word (WW) used in Amazon Echo devices is "*Alexa*", and it is required at the beginning of the voice command phrases in order to get the device's attention. A beam selection algorithm is critical for correct operation of the device and should be carefully designed to avoid missing the signal of interest. In this paper, we proposed a Signal to Interference Ratio (SIR) beam selector using a simple learning technique for improved beam selection.

This paper is organized as follows. In Section 2, we introduce the baseline system, i.e., multiple microphone-based audio front-end (AFE). Next, we present the proposed SIR beam selector in detail in Section 3. Introduction to the datasets and an extensive series of evaluations are then performed and presented in Section 4. Finally, we draw conclusions in Section 5.

2. AFE Baseline System

Figure 1 illustrates a block diagram of the multiple microphone array processing baseline system (also called audio front end AFE) investigated by this paper. It consists of an acoustic echo canceller (AEC), a fixed-beamformer (FBF), a specialized adaptive noise canceller (S-ANC), and a classic beam selector. The AEC is designed to remove the acoustic echoes of the sound played from the device. The FBF is designed to form a set of beams representing different look directions, and S-ANC is designed to remove the ambient noise and the interfering signals coming from directions other than the look direction. The classic beam selector is used to select the proper beam as AFE output. The reason that we named it as "classic" is in order to differentiate it from the newly proposed SIR beam selector. The audio processing block illustrated in Figure 1 is only a portion of the entire AFE, and the overall AFE architecture is described in [12]. Since our primary focus of this paper is to analyze the audio signals on the microphone path and the impact of different beam selection schemes on AFE performance, we will not discuss the case when there is severe acoustic echo here. Our assumption is that the acoustic echoes have been significantly reduced by the acoustic echo canceller (AEC) before FBF.

As illustrated in Figure 1, the signal flow of the baseline system is as follows:

1) The multi-microphone signals are first processed by a AEC block to remove the echoes and then processed by an FBF block. The FBF block employs a subband-based filter-and-sum structure and its weights are specially designed so that the formed beams are able to cover all the look directions of interest. The outputs of FBF block, referred as preliminary beam outputs in this paper, can be considered as a set of directional beams and each beam represents a certain range of spatial angles. In this block, the audio signals that originate from certain directions are enhanced while audio signals that coming from other directions are largely attenuated. The preliminary beam output signals $y_k(n,l)$ can be expressed as:

$$y_k(n,l) = \sum_{m=1}^{I}\sum_{i=0}^{L-1} w_{m,i}^{l,k} \cdot x_{m,i}^{l,k}(n), \qquad (1)$$

Where x denotes the microphone input signals and w denotes the filter weights. k denotes the beam index, and $1 \leq k \leq K_{max}$. K_{max} can be a fixed number or adaptive. We used a fixed number in our simulation. l denotes the subband index, n denotes the discrete time index, m denotes the number of microphones, and i denotes the filter tap.

2) The preliminary beam outputs are then further processed by a specialized multi-channel subband-based adaptive noise canceller (i.e., S-ANC), where the enhanced beam outputs will be generated.

3) A beam selection mechanism is applied to analyze the enhanced candidate beam outputs and therefore choose the proper beam as AFE output for WW and ASR engines.

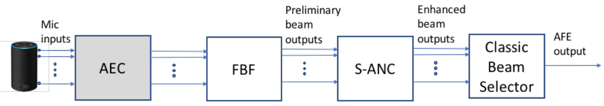

Figure 1: *Block Diagram of Baseline System*

In the baseline AFE system, the classic beam selector is used and placed after the S-ANC block. The classic beam selector is based on signal-to-noise-ratio (i.e., SNR), where SNR of each beam is directly estimated. It is possible that the adaptive noise cancellation may degrade a quality of the signal and/or suppress a desired signal, resulting in beam selection component not selecting a desired beam for WW or ASR engines. A poorly selected beam may reduce the effectiveness of Wakeford detection and speech processing performance.

3. SIR Beam Selector

In this section, we will describe in details the working scheme of the proposed SIR beam selector for multi-microphone audio front-end system.

The SIR beam selector is designed to utilize the Signal-to-Interference Ratios (i.e., SIRs) before S-ANC for beam selection. It consists of two steps:

1) Calculate instantaneous SIR ratios for each beam and select the beams that have strong SIRs.

2) Apply a simple learning algorithm to learn the location of the audio source candidates based on the selected beam indexes.

Figure 2 illustrates the flowchart of SIR beam selector. Figure 3 illustrates the block diagram of the improved AFE baseline system with a SIR beam selector. As shown in Figure 3, the SIR beam selection also uses the same input signals as used by S-ANC.

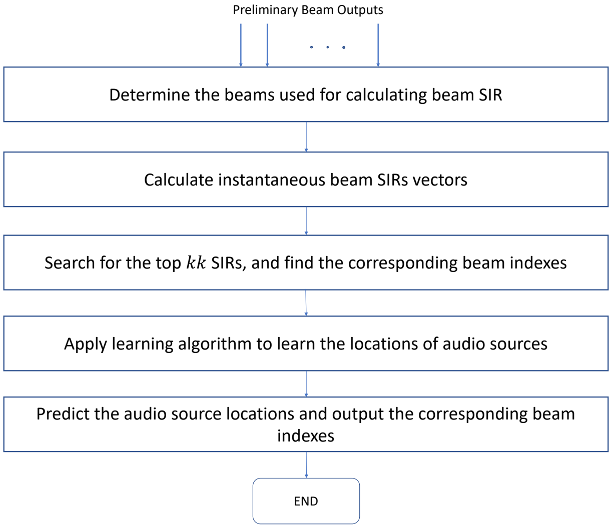
Preliminary Beam Outputs
Determine the beams used for calculating beam SIR
Calculate instantaneous beam SIRs vectors
Search for the top *kk* SIRs, and find the corresponding beam indexes
Apply learning algorithm to learn the locations of audio sources
Predict the audio source locations and output the corresponding beam indexes
END

Figure 2: *Dataflow of SIR Beam Selector*

It is well known that adaptive noise canceller (ANC) has two types of inputs – primary and reference. As shown in Figure 4, the primary input receives a signal s from the signal source that is corrupted by the presence of noise n uncorrelated with the signal, where $y = s + n$. The reference input receives a noise n_0 uncorrelated with the signal but correlated in some way with the noise n. The noise n_0 passes through a filter to produce an output $\hat{n}$ that is a close estimate of primary input noise. This noise estimate is subtracted from the corrupted signal to produce an estimate of the signal at $\hat{s}$, the ANC system output. The concept of ANC can also be found in [13]. Figure 4 illustrates the block diagram of traditional ANC.

In our system, each S-ANC module takes a single beam as a primary signal and one or more other beams as reference signals. The SIR ratios were calculated in subband domain by comparing the smoothed energy of both primary and the reference beams and can be illustrated as below:

$$SIR_k(l,n) = \frac{B_{YY}(l,n)}{\sum_{p=1}^{P} N_{ZZ,p}(l,n) + \delta}, \quad l \in [l_{LB,}l_{UB}] \quad (2)$$

where, k denotes the beam index, l_{LB} denotes the lower bound for the subband range bin and l_{UB}denotes the upper bound for the subband range bin under consideration and δ is a regularization factor. Further, $B_{YY}(l,n)$ denotes the average powers of the primary input signal and $Y(l,n)$ denotes the instantaneous powers of the primary input signal. $N_{ZZ,p}(l,n)$ denotes the average powers of the pth reference input signals and $Z_p(l,n)$ denotes the instantaneous powers of the pth reference input signals. P denotes the total number of reference signals. The powers are calculated using first order recursive averaging as shown below:

$$B_{YY}(l,n) = \alpha B_{YY}(l,n-1) + (1-\alpha)|Y(l,n)|^2, \qquad (3)$$

$$N_{ZZ,p}(l,n) = \alpha BN_{ZZ,p}(l,n-1) + (1-\alpha)|Z_p(l,n)|^2 \quad (4)$$

where, $\alpha \in [0,1]$ is a smoothing parameter.

The frame-based beam SIR is then obtained by averaging across the subband range:

$$SIR_k(n) = \frac{1}{(l_{UB} - l_{LB} + 1)} \sum_{l=l_{LB}}^{l_{UB}} SIR_k(l,n) \qquad (5)$$

A higher beam SIR ratio indicates that there is higher probability of having an audio source coming from the direction that this beam is representing.

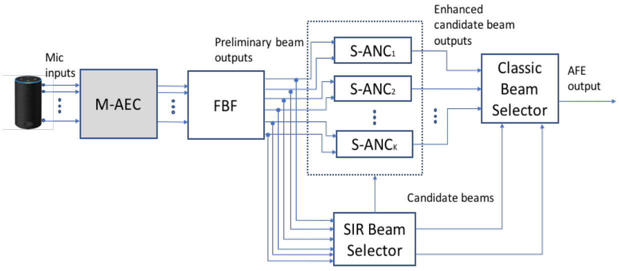

Figure 3: *Improved Baseline System with SIR Beam Selector*

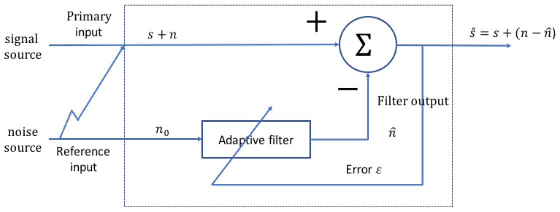

Figure 4: *Adaptive Noise Canceller (ANC)*

As illustrated in Figure 2, the instantaneous SIR ratios (i.e., the ratios between the primary and its corresponding reference signals) are calculated every frame and the beam indexes with stronger SIRs (i.e., potential audio source directions) are found at frame basis accordingly. These selected beam indexes are then utilized by a simple learning algorithm to estimate the audio source locations. The learning algorithm keeps tracking the beams with strong SIRs, analyzes the beam indices and finds how often each beam index was selected. The beams that are selected frequently and constantly indicate that an audio source may locate in the direction that this beam is pointing to. The number of audio sources kk that can be learned by the proposed SIR beam selector needs to be less than the maximum number of beams that FBF blocks have generated. In addition, the value of kk could be adaptive or fixed. Finally, these selected beams will be passed through S-ANC for further audio processing. The S-ANC outputs are referred to as enhanced candidate beam outputs in this study. The learned beam indexes that represent potential audio source locations will be utilized by the classic beam selector and used as a constraint for a refining selection. The classic beam selector will choose the final beam output for WW and ASR engines. The beams that are not selected by the SIR beam selector will not be considered by our classic beam selector.

The advantage of SIR beam selector is to utilize more than one preliminary beam output signals (i.e., speech and noise beam signals) to assist beam selection and locate all the audio sources in the surrounding of the device. This selection process happens before S-ANC, so it is free of speech distortion that might be introduced by S-ANC. Since both S-ANC and classic beam selector will only consider the beams selected by SIR beam selector, the entire AFE algorithm complexity has been greatly reduced.

Next, we will provide our evaluation results.

4. Simulation Results

4.1. Performance Metrics

The entire system processing can be divided into three parts:

i) Audio front-end processing, referred as AFE in this paper.
ii) Wake-word engine, referred as WW in this paper.
iii) Post processing infrastructure, which hosts the ASR, NLU and TTS engines, along with other Alexa services.

The multiple microphone data is first processed by the AFE algorithms and its output is then sent to the WW engine and post processing modules. In our evaluation, we used the following metrics to evaluate our newly proposed SIR system are:

i) Word-error-rate (WER), which is a key metric for the ASR engine and it is defined as the ratio of the decoding errors (insertion, deletion and substitution) and the total number of valid words [14].
ii) False-rejection-rate (FRR), which measures the percentage of missed WW commands.

4.2. Datasets

We used two different internal datasets for performance evaluation. The first dataset that we used was collected in a typical household environment. We used multiple Amazon Echo devices for the data collection and distributed them in each room at various places. Multiple loudspeakers were used to play utterances that contain Wakeford “Alexa”. We placed the loudspeakers at various locations in order to cover all the possible use-device positions. Several common home appliances were used during the data collection to simulate the actual household noise environments. Table 1 lists the test conditions that we selected form this dataset for evaluating the proposed SIR beam selector. For each test condition, we used more than 10,000 utterances leveraging both male and female speech.

Table 1: *List of Selected Test Conditions from the 1st Dataset*

Test Case	**Music Playback**	**Home Appliance**
vol_0	No	off
vol_3	volume 3	off
vol_4	volume 4	off
vol_5	volume 5	off
vol_6	volume 6	off
vol_7	volume 7	off
Home Appliance #1	Various	on
Home Appliance #2	Various	on

The second dataset that we used for evaluating WER performance is a larger dataset. This dataset includes 49 speakers that both male and female speakers were invited for the data collection and seven different Amazon devices were used during that data collection. In addition, this dataset was manually labelled. In this paper, we used a portion of this dataset that has around 40,000 utterances under a variety of ambient noise conditions.

4.3. Simulation Results

Table 2 provides the beam selection accuracy improvement under both clean and noisy conditions after integrating the SIR beam selector to our baseline system. The experimental results demonstrate that SIR beam selector improved the beam selection accuracy under both clean and noisy conditions.

Table 2: *Beam Selection Accuracy Improvement*

Test Condition	Clean	Noisy
Detection Accuracy Improv. Over Baseline System	52.51%	62.20%

Table 3 provides the relative FRR and WER improvement of the improved AFE system with a SIR beam selector over our baseline system under various test conditions.

Table 3: *Relative FRR and WER Improvement over Baseline System*

Test Condition	FRR relative Improvement	WER relative Improvement
vol_0	21.9%	17.78%
vol_3	23.37%	17.95%
vol_4	20.5%	22.34%
vol_5	15.86%	20%
vol_6	11.46%	18.52%
vol_7	8.67%	11.36%
Home appliance #1	15.31%	16.67%
Home Appliance #2	8.65%	11.52%
Overall	12.49%	15.73%

Table 4 provides the relative WER improvement of overall system over raw microphone data under various noise conditions. For relative FRR improvement, we achieved 39.18% over raw microphone.

Table 4: *Relative WER Improvement over Raw Microphone*

Raw Data SNR Range	Number of Utterances	Relative WER Improvement
[-20dB ~ 0dB)	5029	44.10%
[0dB ~ 4 dB)	14319	46.20%
[4dB ~ 8dB)	12982	27.47%
[8dB ~ 12dB)	5651	17.75%
[12dB ~ 30dB]	1046	16.37%
Overall	39027	39.62%

The results provided in Table 2 and 3 used the 1st dataset and Table 4 and Figure 3 results used the 2nd dataset.

5. Conclusions

In this paper, we have proposed a SIR beam selector for multi-microphone audio front-end (AFE) system for far-field automatic speech recognition in Amazon echo devices. The proposed SIR beam selector utilized the signal-to-interference ratios to learn the locations of the audio sources and further improved the accuracy of the desired beam selection. We demonstrated that the AFE with a SIR beam selector outperforms the baseline AFE system that utilized only a classic beam selector. Compared with the baseline system, the proposed system can improve the beam selection accuracy about 52.51% for clean condition and 62.20% for noisy condition, speech recognition performance by decreasing relative WER by 11.36% - 22.34% under various test conditions and barge-in performance by decreasing relative FRR by 8.65% - 23.37% under various test conditions using the internal datasets. We have also shown that the AFE with a SIR beam selector outperforms a single channel raw microphone by decreasing overall relative WER by 39.62% and overall relative FRR by 39.18% using a larger dataset. Finally, SIR beam selection requires neither a calibration signal nor *a prior* knowledge of speech or noise sources.

While the SIR beam selection algorithm has been shown to be effective for keyword barge-in and automatic speech recognition in Amazon echo devices, there are areas for further research. With the increasing demand for various Amazon devices, robustness is a key criterion for the AFE algorithms; we continue to explore better models and algorithms to improve the robustness of AFE by adaptively learning the environmental parameters.

6. References

[1] M. Brandstein and D. Ward, *Microphone arrays: signal processing techniques and applications*, Springer Science & Business Media, 2001.

[2] B. D. Van Veen and K. M. Buckley, "*Beamforming: A Versatile Approach to Spatial Filtering*", IEEE ASSP Magazine, pp. 4-24, April, 1989.

[3] I. A. McCowan," Robust *Speech Recognition using Microphone Arrays*," PhD Thesis, Queensland University of Technology, Australia, 2001.

[4] S. Haykin, J. H. Justice, N. L. Owsley, J. L. Yen, and A. C. Kab, *Array Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1985, pp. 119–123.

[5] V. Galanenko and A. Kalyuzhny, "Investigation of effectiveness of microphone arrays for in car use based on sound field

simulation," in *Proc. IEEE ICASSP 2001*, vol. 5, UT, May 2001, pp. 3017–3020.

[6] O. L. Frost, III, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[7] J. Benesty, J. D. Chen, Y. Huang, and J. Dmochowski, "*On Microphone-Array Beamforming from a MIMO Acoustic Signal Processing Perspective*", IEEE Trans. On Audio, Speech, and Language Processing, vol. 15, No. 3, March 2007.

[8] M. Souden, J. Benesty, and S. Affes, "*On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction*", IEEE Trans. On Audio, Speech, and Language Processing, vol. 18, no. 2, Feb. 2010.

[9] E. Habets, J. Benesty, S. Gannot, P. Naylor, and I. Cohen, "On the Application of the LCMV Beamformer to Speech Enhancement", *in Proc. Of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 18-21, New Paltz, NY, 2009, pp. 141 – 144.

[10] L. J. Griffiths and C. W. Jim, "*An alternative approach to linearly con- strained adaptive beamforming*," IEEE Trans. Antennas Propagat., vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[11] K. M. Buckley, "Broad-band beamforming and the general- ized sidelobe canceller," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, pp. 1322–1323, Oct. 1986.

[12] A. Chhetri, P. Hilmes, T. Kristjansson, R. Ayrapetian, W. Chu, M. Mansour, X. Li, and X. Zhang, "*Multichannel Audio Front-end for Far-field Automatic Speech Recognition*," Accepted by EUSIPCO'2018, Rome, Italy, Sept. 3-7, 2018.

[13] B. Widrow, J. Glover, J. McCool, *et. al.* "Adaptive Noise Cancelling: Principles and Applications", Proceedings of the IEEE, vol. 63, issue 12, pp. 1692-1716, Dec. 1975.

[14] M. Wolfel and J. McDonough, "*Distant Speech Recognition*," John Wiley & Sons, 2009.