ENHANCED MVDR BEAMFORMING FOR ARRAYS OF DIRECTIONAL MICROPHONES

Demba E. Ba

Dept. of Electrical Eng. and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
demba@mit.edu

Dinei Florêncio and Cha Zhang

Microsoft Research
One Microsoft Way
Redmond, WA 98052
{dinei,chazhang}@microsoft.com

ABSTRACT

Microphone arrays based on the minimum variance distortionless response (MVDR) beamformer are among the most popular for speech enhancement applications. The original MVDR is excessively sensitive to source location and microphone gains. Previous research has made MVDR practical by successfully increasing the robustness of MVDR to source location, and MVDR-based microphone arrays are already commercially available. Nevertheless, MVDR performance is still weak in cases where microphone gain variations are too large, e.g., for circular arrays of directional microphones. In this paper we propose an improved MVDR beamformer which takes into account the effect of sensors (e.g. microphones) with arbitrary, potentially directional responses. Specifically, we form estimates of the *relative* magnitude responses of the sensors based on the data received at the array and include those in the original formulation of the MVDR beamforming problem. Experimental results on real-world audio data show an average 2.4 dB improvement over conventional MVDR beamforming, which does not account for the magnitude responses of the sensors.

***Index Terms*—** Microphone arrays, MVDR, sound capture, directional microphones, circular arrays.

1. INTRODUCTION

As globalization continues to spread throughout the world economy, it is increasingly common to find projects where team members reside in different time zones. To provide a means for distributed groups to work together on shared problems, there has been an increasing interest in building special purpose devices and even "smart rooms" to support distributed meetings. These multimedia devices often contain multiple microphones and cameras, and fancy features, such as automatic speaker localization [1]. An example device called RoundTable [2] is shown in Figure 1(a). It has a six-element circular microphone array at the base, and five video cameras at the top. The captured videos are stitched into a 360 degree panorama, which gives a global view of the meeting room. The RoundTable device enables remote group members to hear and view the meeting live online. In addition, the meetings can be recorded and archived, allowing people to browse them afterward.

As in many other communication devices, sound quality is of foremost importance. Microphone arrays are a common way to improve sound quality; the diversity in the received signals is exploited by setting different gains to each mic, depending on the location of the source and the interference. This is generally referred to as *beamforming*. Early designs were generally "fixed" beamformers (e.g., delay-and-sum), adapting only to the location of the desired source.

Fig. 1. RoundTable and its captured images. (a) The RoundTable device. (b) Captured images.

More recent designs are based on "null-steering", and adapt to characteristics of the interference as well. The minimum variance distortionless response (MVDR) beamformer and its associated adaptive algorithm, the generalized sidelobe canceller (GSC) [3, 4], are probably the most widely studied and used beamforming algorithms, and are basis to some commercially available arrays [5]. Assuming the direction of arrival (DOA) of the desired signal is known, the MVDR beamformer estimates the desired signal while minimizing the variance of the noise component of the formed estimate. In practice, however, the DOA of the desired signal is not known exactly, which significantly degrades the performance of the MVDR beamformer [6]. A lot of research has been done into a class of algorithms known as robust MVDR [7, 8]. As a general rule, these algorithms work by extending the region where the source can be located. Nevertheless, even assuming perfect sound source localization (SSL), the fact that the sensors may have distinct, directional responses adds yet another level of uncertainty that the MVDR beamformer is not able to handle well. Commercial arrays solve this by using a linear array of microphones, all pointing at the same direction, and therefore with similar directional gain. Nevertheless, for the circular geometry used in the RoundTable, this directionality is accentuated: each microphone will have a significantly different direction of arrival in relation to the desired source. This has not been exploited much in the existing literature, and presents new challenges. Experiments showed that MVDR and other existing algorithms perform well when omnidirectional mics are used, but do not provide much enhancement when directional mics are used. For example, we will show in section 6 that MVDR provides only an 1.1dB improvement over the best microphone.

In this paper, we introduce an enhanced MVDR (eMVDR) beamformer that can be applied to a circular array of directional microphones. The proposed beamformer distinguishes itself from previous approaches in that it explicitly factors in the effect of microphones with arbitrary, potentially directional responses. Specifically, we form estimates of the *relative* magnitude responses of the microphones based on the data received at the array and include those in

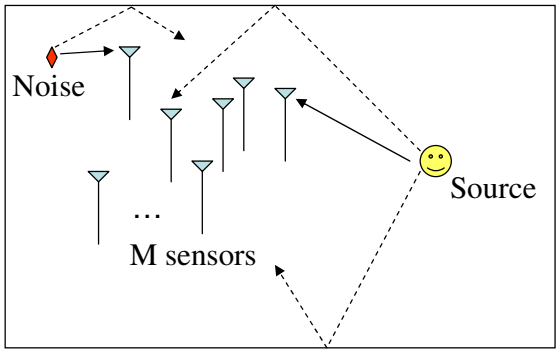

Fig. 2. A source incident on an array of *M* sensors in the presence of noise and multipath

the original formulation of the MVDR beamforming problem. We show through experiments on real audio data that the proposed approach indeed yields an average improvement of 2.4 dB over traditional MVDR beamforming, which does not account for directional microphone responses.

In the next section, we describe the general observation model considered in this paper. This sets the stage for a review of traditional MVDR beamforming in Section 3 and the introduction of the eMVDR beamformer in Section 4. Some details pertaining to our implementation of the eMVDR beamformer are discussed in Section 5. In Section 6, we describe the data set and experimental set up that were used, as well as results of experiments we conducted. We provide concluding remarks in Section 7.

2. OBSERVATION MODEL

Consider a signal $s(t)$, impinging on an array of M sensors as shown in Figure 2. The positions of the sensors are assumed to be known. We model the received signal $x_i(t), i \in \{1, ..., M\}$ at each sensor as:

$$x_i(t) = \alpha_i s(t - \tau_i) + h_i(t) \otimes s(t) + n_i(t). \quad (1)$$

where α_i is a parameter that includes the intrinsic gain of the corresponding sensor as well as its directionality and the propagation loss from the source to the sensor; τ_i is the time delay of propagation associated with the direct path of the source, which is a function of the source and the sensor's location; $h_i(t)$ models the multipath effects to the source, often referred to as *reverberation*; "$\otimes$" denotes convolution; $n_i(t)$ is the sensor noise at each microphone. We can re-write Eq. (1) in the frequency domain as:

$$X_i(\omega) = \alpha_i(\omega)S(\omega)e^{-j\omega\tau_i} + H_i(\omega)S(\omega) + N_i(\omega), \quad (2)$$

where we also allowed for the α_i to vary with frequency. Since multiple sensors are involved, we can express the overall system in vector form:

$$\mathbf{X}(\omega) = S(\omega)\mathbf{d}(\omega) + \mathbf{H}(\omega)S(\omega) + \mathbf{N}(\omega), \quad (3)$$

where

$$\begin{aligned} \mathbf{X}(\omega) &= [X_1(\omega), \cdots, X_M(\omega)]^T, \\ \mathbf{d}(\omega) &= [\alpha_1(\omega)e^{-j\omega\tau_1}, \cdots, \alpha_M(\omega)e^{-j\omega\tau_M}]^T \\ \mathbf{N}(\omega) &= [N_1(\omega), \cdots, N_M(\omega)]^T, \\ \mathbf{H}(\omega) &= [H_1(\omega), \cdots, H_M(\omega)]^T. \end{aligned}$$

The primary source of uncertainty in the above model is the array response vector $\mathbf{d}(\omega)$ and the reverberation filter $\mathbf{H}(\omega)$. The same problem appears in sound source localization, and various methods to approximate $\mathbf{H}(\omega)$ have been proposed [9, 10]. However the effect of $\mathbf{d}(\omega)$, and in particular its dependency on the characteristics of the sensors, has been largely ignored in the literature. While some may argue that the microphone response may be pre-calibrated, this may not be practical in all cases. For instance, in the RoundTable device, the microphones used are directional, which means different gains along different directions of arrival. In addition, microphone gain variations on the order of 4 dB are common due to manufacturing tolerances. Measuring the gain of each microphone, at every direction, for each device would be time-consuming and expensive. The goal of this paper is to increase the robustness of the MVDR beamformer to the uncertainty associated $\mathbf{d}(\omega)$, in order to enhance the performance of circular arrays.

3. TRADITIONAL MVDR BEAMFORMING

The goal of beamforming is to estimate the desired signal s as a linear combination of the data collected at the array. In other words, we would like to determine an $M \times 1$ vector weights $\mathbf{w}(\omega)$ such that $\mathbf{w}^H(\omega)\mathbf{X}(\omega)$ is a good estimate of $S(\omega)$. The beamformer that results from minimizing the variance of the noise component of $\mathbf{w}^H\mathbf{X}$, subject to a constraint of gain 1 in the look direction, is known as the MVDR beamformer. The corresponding weight vector $\mathbf{w}$ is the solution to the following optimization problem:

$$\min_{\mathbf{w}} \mathbf{w}^H\mathbf{Q}\mathbf{w} \;\; s.t. \;\; \mathbf{w}^H\mathbf{d} = 1, \quad (4)$$

where

$$\begin{aligned} \mathbf{N}_c(\omega) &= \mathbf{H}(\omega)S(\omega) + \mathbf{N}(\omega), & (5) \\ \mathbf{Q}(\omega) &= E[\mathbf{N}_c(\omega)\mathbf{N}_c^H(\omega)] & (6) \end{aligned}$$

Here $\mathbf{N}_c(\omega)$ is the combined noise (reflected paths and auxiliary sources). $\mathbf{Q}(\omega)$ is the covariance matrix of the combined noise. It is estimated from the data and therefore inherently contains information about the location of the sources of interference, as well as the effect of the sensors on those sources.

The optimization problem in Eq. (4) has an elegant closed-form solution [7] given by:

$$\mathbf{w} = \frac{\mathbf{Q}^{-1}\mathbf{d}}{\mathbf{d}^H\mathbf{Q}^{-1}\mathbf{d}} \quad (7)$$

Note that the denominator of Eq. (7) is merely a normalization factor which enforces the gain 1 constraint in the look direction.

The MVDR beamforming algorithm has been very popular in the literature. In most previous works, the sensors are assumed to be omnidirectional or all pointing in the same direction (and assumed to have the same directional gain). Namely, the α_i's in $\mathbf{d}$ are assumed to be equal to 1 (or measurable beforehand). However this may not always be true. For instance, the array in the RoundTable device, uses highly directional, uncalibrated microphones [11]. This design has shown to produce better sound capture than, for example, omnidirectional mics followed by beamforming. Therefore, the α_i's are unknown and have to be estimated from the perceived signals.

4. MVDR WITH SENSOR GAIN COMPENSATION

We assign a weight g_i, $i \in 1, ..., M$, to each of the components of $\mathbf{d}$ based on the *relative* strength of the signal recorded at sensor i

compared to all other sensors. In this way we can compensate for the effect of sensors with directional gain patterns. In the following, we describe how the g_i's are computed based on the data received at the array.

Assume that $S(\omega)$ and $N_i(\omega)$ are uncorrelated. The energy in the reflected paths of the signal (second term in Eq. (2)) is very complex. Following existing work such as [10], we assume it is a proportion γ of $|X_i(\omega)|^2 - |N_i(\omega)|^2$, then,

$$\begin{aligned} E[|X_i(\omega)|^2] &= |\alpha_i(\omega)|^2|S(\omega)|^2 \\ &+ \gamma|X_i(\omega)|^2 + (1-\gamma)|N_i(\omega)|^2. \end{aligned}$$

Rearranging the above equation, we have that

$$|\alpha_i(\omega)|.|S(\omega)| = \sqrt{(1-\gamma)(|X_i(\omega)|^2 - |N_i(\omega)|^2)} \quad (8)$$

In Eq. (8), $|X_i(\omega)|^2$ can be directly computed from the data collected at the array. $|N_i(\omega)|^2$ can be determined from the silence periods of $X_i(\omega)$. Note that $|\alpha_i(\omega)|$ on its own cannot be estimated from the data; only the product $|\alpha_i(\omega)|.|S(\omega)|$ is observable from the data. However, this is not an issue because we are interested only in the *relative* gain of a given sensor with respect to other sensors. Therefore, we define g_i as follows:

$$g_i = \frac{|\alpha_i(\omega)|.|S(\omega)|}{\sum_{j=1,...,M} |\alpha_j(\omega)|.|S(\omega)|}, i \in 1, ..., M. \quad (9)$$

The resulting array response vector $\mathbf{d}$ is given by:

$$\mathbf{d}(\omega) = [g_1(\omega)e^{-j\omega\tau_1}, \cdots, g_M(\omega)e^{-j\omega\tau_M}]^T \quad (10)$$

The corresponding weight vector $\mathbf{w}$ is obtained by substituting Eq. (10) in the closed-form solution to the MVDR beamforming problem (Eq. (7)). Note that g_i as defined in Eq. (9) compensates for the *gain* response of the sensors only. The problem of compensating for the *phase* response of the microphones is the subject of future work.

5. IMPLEMENTATION DETAILS

Figure 3 shows a block diagram of our implementation of eMVDR beamforming using an array of directional microphones. The microphone array considered uses a 16 KHz sampling rate. The data is processed in frames of length 40 ms (overlapped), which corresponds to 640 samples per frame. Consecutive frames have a 20 ms overlap, i.e. 320 samples. The working frequency band was chosen to be between 200 and 7200 Hz.

1. **Working in the frequency domain**: Each frame first undergoes a transformation to the frequency domain using the modulated complex lapped transform (MCLT). The MCLT has been shown to be useful in a variety of audio processing applications [12]. Alternatively, the discrete fourier transform could be used. A beamformer is computed as in Eq. (7) for each discrete frequency bin. After beamforming, the time domain estimate of the desired signal is computed from its frequency domain estimate through inverse MCLT transformation (I-MCLT).
2. **Voice Activity Detection and Q computation**: Once in the frequency domain, each frame goes through a voice activity detector (VAD). The VAD used in our implementation is a classical energy-based VAD. However, unlike conventional binary (speech or noise) VADs, our VAD classifies a given frame as one of three possible choices, namely *Speech*, *Noise* or *Not Sure*. We found that a decision of noise has to be made conservatively in order to avoid leakage of the speech signal into the estimate of $\mathbf{Q}$. Recall from Eq. (7) that the logic behind MVDR beamforming is to suppress projections of the array data on $\mathbf{Q}$, i.e. that are correlated with the noise. If the estimate of $\mathbf{Q}$ were to contain contributions from the source signal, this would lead to (at least partial) cancellation of the signal by the beamformer. The noise covariance matrix $\mathbf{Q}$ is obtained from frames classified as *Noise* by computing its sample mean.
3. **SSL and eMVDR beamforming**: The DOA of the source s is determined from frames classified as *Speech* through SSL. This is followed by eMVDR beamforming. In our simulations we used $\gamma = 0.2$, but performance was reasonable for a range of values of γ, from 0.1 to 0.5. The SSL algorithm we use is based on time delay of arrival and maximum likelihood estimation [13]. It turns out that maximum likelihood based SSL is equivalent to forming multiple MVDR beamformers and choosing that with the best output signal to noise ratio [14]. We also perform beamforming on frames that were classified as *Noise* or *Not Sure* using the beamformer weights of the last frame classified as *Speech*. This was merely a heuristic choice which seemed to work well in practice.

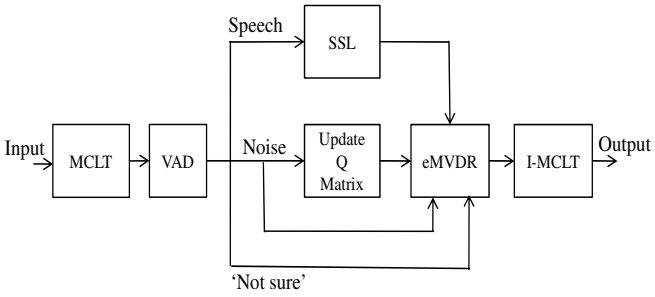

Fig. 3. Block diagram of the eMVDR beamforming

6. EXPERIMENTAL RESULTS

In this section, we present the results of experiments using data collected by a circular array of directional microphones in the RoundTable device. The data set has 10 audio sequences collected by the RoundTable device in various rooms. Each sequence is about 4 minutes long. The test data includes recordings of single speakers, multiple speakers, as well as multiple speakers with cross-talk. The speech frames are manually segmented for the computation of the SNR.

We compare the eMVDR beamformer to an MVDR algorithm without gain compensation as well as to a simple microphone selection scheme. This latter scheme chooses the microphone with the highest signal to noise ratio (SNR) as the estimate of the desired signal s, and has shown to perform surprisingly well for the array in question. The SNRs of the output signals are compared against each other to evaluate performance.

Table 1 summarizes the SNR results of the experiments. The column on SSL accuracy is included to help understand some of the factors influencing performance. It reports the fraction of frames where the SSL [13] correctly pointed at (one of) the speaker(s) with a 6 degrees tolerance. Note that SSL performance is affect by both

Table 1. Comparisons, based on SNR (dB), of Best mic selection, MVDR and eMVDR on 10 audio clips.

Clip ID	Best Mic	MVDR	eMVDR	SSL Accuracy
A	10.6	12.7	13.9	92.5%
B	19.8	21.5	25.5	95.5%
C	16.2	16.8	19.6	72.6%
D	22.6	24.2	25.2	98.3%
E	23.3	22.6	22.9	73.2%
F	18.8	21.8	24	93.9%
G	13.4	14.2	17.7	82.1%
H	20.2	21.1	23	45.1%
I	19.3	18.8	24.4	97.6%
J	14	14.9	16.4	54.4%
Avg.	17.8	**18.9**	**21.3**	80.5%

SNR and reverberation. As it can be seen in Table 1, the eMVDR algorithm always outperforms traditional MVDR beamforming in terms of SNR. The average performance gain is 2.4 dB. The eMVDR beamformer also outperforms the best mic selection scheme by an average of 3.5 dB. It is interesting to note that there are two cases (E and I) where best mic selection does better than traditional MVDR beamforming and one (E) where it does better than eMVDR. Sequence I corresponds to a case with high SSL accuracy, which shows that not compensating for the directionality of the microphones can turn out to be expensive at times in terms of degrading the performance of the beamformer. Case E is a scenario were SSL accuracy is not as good as in I. This suggests that, in case E, one of the directional microphones might have been pointing directly at the source (meaning that the best mic selection scheme might have had a better estimate of the DOA of the source in case I). This may explain why, even after compensating for the gain pattern of the microphones, eMVDR still does slightly worse than best mic selection in case E. The performance loss in this case could be attributed to SSL accuracy.

The results presented in Table 1 highlight several important points. First, they underline the importance of compensating for the gain pattern of directional microphones when using MVDR beamforming for speech enhancement. The enhancement of the proposed algorithm was 3.5 dB, compared to the 1.1 dB enhancement produced by the traditional MVDR. This gain compensation method for MVDR is the main contribution of this paper. Second, very much to our surprise, the best mic selection scheme does not seem to be such a bad algorithm after all. It has very low computational complexity and has performance comparable to that of traditional MVDR beamforming and not very far from that of eMVDR (at least in the high SNR cases). However, we believe that the advantage of eMVDR beamforming over best mic selection comes from the fact that the average improvement in SNR that was recorded can allow us to be more conservative in nonlinear post-processing operations. Since the post-processing usually distorts the signal when SNR is not high enough, a few dBs of improvement in SNR through eMVDR beamforming could actually result in better perceptual audio quality after post-processing.

7. CONCLUSION

It is well-known that the performance of traditional MVDR beamforming significantly degrades in the presence of mismatches between the actual DOA of the desired signal (e.g. speech) and that estimated through SSL. In speech enhancement applications, robustness of the beamformer to pointing errors is often increased by exploiting information provided by a video channel, that is, using so-called joint audio-video SSL [15]. We argued that, even with 100% SSL accuracy, assumption of uniform gain when directional sensors (e.g. microphones) are used also degrades the performance of the traditional MVDR beamformer. To address this latter issue, we proposed an enhanced MVDR (eMVDR) beamformer which compensates for the effect of sensors with directional magnitude responses. Specifically, the proposed method forms estimates of the *relative* magnitude responses of the sensors based on the data received at the array and incorporates those in the formulation of the traditional MVDR beamforming problem. We showed the superiority of our approach to traditional MVDR beamforming through experiments conducted on real audio data.

In the future, we plan to look into compensating for the phase response of directional microphones as it also varies significantly from one directional microphone to another. Another interesting problem would be to determine the relative importance of increased SSL accuracy versus microphone directionality compensation in improving the performance of the MVDR beamformer.

8. REFERENCES

[1] C. Zhang, P. Yin, Y. Rui, R. Cutler, and P. Viola, "Boosting-based multimodal speaker detection for distributed meetings," *MMSP*, 2006.

[2] http://www.microsoft.com/presspass/presskits/uc/gallery.mspx.

[3] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. on Anttenas and Propagation*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[4] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Sig. Proc.*, vol. 47, no. 10, pp. 2677–2684, Oct. 1999.

[5] http://bitwave.com.sg/products.php?currentpage=computers.

[6] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Sig. Proc. Mag.*, pp. 4–24, Apr. 1988.

[7] H. Cox, R. M. Zeskind, and M. H. Owen, "Robust adaptive beamforming," *IEEE Trans. ASSP*, pp. 1365–1376, Oct. 1987.

[8] A. El-Keyi, T. Kirubarajan, and A. Gershman, "Robust adaptive beamforming based on the kalman filter," *IEEE Trans. Sig. Proc.*, vol. 53, no. 8, pp. 3032–3041, Aug. 2005.

[9] H. Wang and P. Chu, "Voice source localization for automatic camera pointing system in videoconferencing," *ICASSP*, 1997.

[10] Y. Rui and D. A. Florencio, "Time delay estimation in the presence of correlated noise and reverberation," *ICASSP*, 2005.

[11] R. Cutler, Y. Rui, A. Gupta, J. Cadiz, I. Tashev, L.W. He, A. Colburn, Z. Zhang, Z. Liu, and S. Silverbert, "Distributed meetings: a meeting capture and broadcasting system," in *Proc. ACM Conf. on Multimedia*, 2002.

[12] H. S. Malvar, "A modulated complex lapped transform and its application to audio processing," *ICASSP*, pp. 1421–1424, 1999.

[13] C. Zhang, Z. Zhang, and D. Florêncio, "Maximum likelyhood sound source localization for multiple directional microphones," *ICASSP*, pp. 125–128, 2007.

[14] K Harmanci, J. Tabrikian, and J. Krolik, "Relationships between adaptive minimum variance beamforming and optimal source localization," *IEEE Trans. Sig. Proc.*, vol. 48, no. 1, pp. 1–12, Jan. 2000.

[15] N. Strobel, S. Spors, and R. Rabenstein, "Joint audio-video object localization and tracking," *IEEE Sig. Proc. Mag.*, vol. 18, pp. 22–31, Jan. 2001.