A LOG-MMSE ADAPTIVE BEAMFORMER USING A NONLINEAR SPATIAL FILTER

Michael L. Seltzer and Ivan Tashev

Speech Technology Group
Microsoft Research
Redmond, WA 98052 USA
{mseltzer,ivantash}@microsoft.com

ABSTRACT

In this paper, we present a new adaptive microphone array processing algorithm for hands-free sound capture. Most traditional adaptive beamforming techniques operate solely on the basis of the direction of arrival of the desired source and are blind to any knowledge about the signal itself. In contrast, the proposed algorithm uses a nonlinear spatial filter to generate an estimate of the magnitude of the source signal. This estimate is then used to drive the adaptation of a linear beamformer according to a log-MMSE criterion. By combining a nonlinear spatial filter with an adaptive beamformer in this manner, we are able to exploit the high SNR output from the nonlinear spatial filter to drive a linear beamformer that does not suffer from distortions or artifacts. A series of experiments demonstrate that the proposed method generates improvements in PESQ and SNR over conventional methods.

***Index Terms*—** adaptive beamforming, microphone arrays, spatial filtering

1. INTRODUCTION

High quality sound capture in hands-free applications has been an ongoing challenge for the last several decades. The growing use of mobile phones and computing devices, voice over IP (VoIP), and speech recognition, has increased the need for such technology. Microphone arrays have long been proposed as a means of obtaining high quality sound capture. The source signal is captured by multiple microphones and jointly processed to generate an enhanced output signal [1]. The most common form of microphone array processing is beamforming, which creates a linear spatial filter that can capture sounds from a desired direction and attenuate sources of unwanted noise. Most beamformers can be divided into two basic classes of algorithms, time-invariant and adaptive. Time-invariant beamformers are those whose parameters are created offline and then held constant during deployment. In contrast, adaptive beamforming algorithms have parameters that are updated during deployment in order to better react to a noise environment that is unknown *a priori*.

The most well known methods of both time-invariant and adaptive beamforming operate under the Minimum Variance Distortionless Response (MVDR) principle. That is, they seek to minimize the power of the array's output signal subject to the constraint that there should be no distortion in gain or phase of signals coming from the desired direction of interest. Both the delay-and-sum and superdirective beamformers operate under this principle (under different assumptions of the noise). The most well-known adaptive algorithms, the Frost beamformer [2] and the Generalized Sidelobe Canceler (GSC) by Griffiths and Jim [3] also operate under this criterion, but in an online manner.

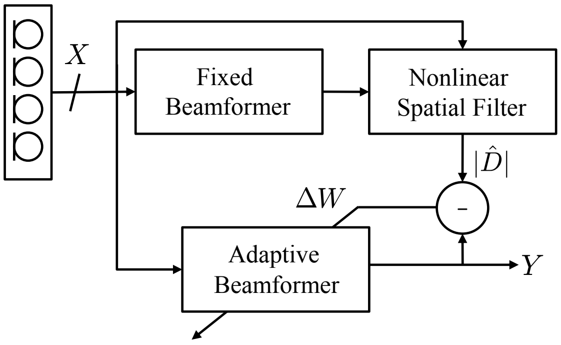

Fig. 1. Block diagram of proposed Log-MMSE adaptive beamforming algorithm that uses a nonlinear spatial filter to generate an estimate of the desired target signal.

One of the most curious and interesting aspects of these beamforming algorithms is that they are "signal agnostic". That is, they operate blind to any knowledge of the source signal and perform solely on the basis of a presumed or estimated direction-of-arrival (DOA). In this paper, we propose a new adaptive beamforming algorithm that attempts to use an estimate of the target signal to drive the beamformer adaptation. We use a nonlinear spatial filter as a preprocessor that generates an estimate of the source signal. The spatial filter's suppression rule is based on a time-varying gain function which generates an output signal that has significantly improved SNR. However, because the spatial filter estimates the magnitude of the target signal but not the phase, the output tends to have the same types of distortion and artifacts that plague other common noise suppression algorithms, e.g. Wiener filtering.

In this work, we use the estimate of the magnitude generated by the spatial filter as a target signal for an adaptive beamformer. By using only the magnitude from the spatial filter and not the noisy phase, we can learn the best linear filter that approaches the spatial filter output signal. Because it is learned in an online fashion in an iterative manner, it evolves smoothly and thus is free from the distortions and artifacts created by the spatial filter. The beamformer uses an MMSE objective function that operates on the log spectra of the estimated target signal and the array's output signal. Operating on the log spectrum has two advantages over the complex spectrum or the magnitudes: first, the log operation is similar to the compression that occurs in the human auditory system and as a result, log-domain optimization is believed to be more perceptually relevant than spectral optimization. Secondly, because of the compressive nature of the log operation at large values, large differences in magnitude produce relatively small differences in the log domain. As a result, the log domain optimization is robust to errors in the estimation of the target signal's magnitude. A block diagram of the proposed algorithm is shown in Figure 1.

The remainder of this paper is as follows. In Section 2 we review traditional adaptive beamforming, and in particular, Frost's algorithm. Section 3 describes the nonlinear spatial filter used in this work. The Log-MMSE adaptive beamforming algorithm is presented in Section 4. We describe the basic algorithm and two variants. In Section 5, the efficacy of the proposed beamforming algorithm is shown through a series of experiments on actual recordings. Finally, we present some conclusions in Section 6.

2. ADAPTIVE BEAMFORMING

We assume that a source signal $D_t(\omega)$ is captured by an array of M microphones. The received signals $X_t(\omega) = \{X_{1,t}(\omega), \ldots, X_{M,t}(\omega)\}$ are then segmented into a sequence of overlapping frames, converted to the frequency domain using a short-time Fourier transform (STFT) and processed by a set of beamformer parameters $W_t(\omega) = \{W_{1,t}(\omega), \ldots, W_{M,t}(\omega)\}$ to create an output signal $Y_t(\omega)$ as follows:

$$Y_t(\omega) = \sum_m W^*_{m,t}(\omega) X_{m,t}(\omega) = W_t^H(\omega) X_t(\omega) \qquad (1)$$

If a time-invariant beamformer is being employed, the weights do not vary over time, i.e. $W_t(\omega) = W(\omega)$.

In this paper, we assume that all frequency bins are processed independently. As a result, we will drop references to frequency bin ω from our notation for convenience and clarity.

In an adaptive beamformer, the goal is to learn the beamformer parameters W_t in an online manner as samples are received. Most adaptive beamformers have been developed by examining the derivation of time-invariant beamformers and substituting instantaneous estimates for long-term statistics. For example, the well-known Frost beamformer utilizes the same MVDR design criterion used to create the superdirective or delay-and-sum beamformers. Thus, the goal of the Frost beamformer is to minimize the power of the array's output signal, subject to a linear constraint that specifies zero distortion in gain or phase from the desired look direction. This results in the following objective function

$$H(W) = \frac{1}{2} W^H S_{XX} W + \lambda(W^H C - \mathcal{F}) \qquad (2)$$

where C describes the steering vector in the desired look direction θ and $\mathcal{F}$ defines the desired frequency response in this direction. To derive an online adaptive version of the MVDR beamformer, Frost used a gradient descent method whereby the weights at a given time instant are a function of the previous weights and the gradient of the objective function with respect to these weights.

$$W_{t+1} = W_t - \mu \nabla_W H(W_t) \qquad (3)$$

These updated weights must also satisfy the distortionless constraint, such that

$$W^H_{t+1} C = \mathcal{F} \qquad (4)$$

By taking the derivative of $H(W)$, substituting (3) into (4), and solving for W_{t+1}, it can be shown that the adaptive beamformer in Frost's algorithm has the following update equation

$$W_{t+1} = P(W_t - \mu Y_t^H X_t) + F \qquad (5)$$

where μ is the learning rate, $P = (I - C(C^H C)^{-1} C^H)$ and $F = C(C^H C)^{-1} \mathcal{F}$.

3. NONLINEAR SPATIAL FILTERING

Nonlinear spatial filtering was originally proposed as a post-filtering algorithm to achieve further noise reduction of the output channel of a time-invariant beamfomer [4]. We briefly review this algorithm here. In a given frame and frequency bin, an Instantaneous DOA (IDOA) vector Δ_t is formed from the phase differences between all microphone pairs.

The spatial filter is formed by computing a probability that an observed Δ_t originated from the desired look direction θ. This is done by first computing the Euclidean distance between Δ_t and Δ_θ, which is the IDOA vector generated by an ideal source originating from θ. This distance in IDOA space is then converted to a distance in physical space, denoted Γ^θ_t. For a linear array, this physical distance represents the absolute difference in radians between the angle of arrival of X_t and the desired look direction θ.

In the absence of noise, the distance Γ^θ_t would be equal to zero if $\Delta_t = \Delta_\theta$. To reflect the presence of noise, we assume that Γ^θ_t follows a Gaussian distribution with zero mean and variance σ^2_θ, i.e. $p(\Gamma^\theta_t) \sim \mathcal{N}(0; \sigma^2_\theta)$. Estimates of the variance σ^2_θ are made online during non-speech segments for a discrete set of look directions.

The nonlinear spatial filter Λ^θ_t is computed as the ratio of the probabilities of Γ^θ_t and $\Gamma^\theta_{\text{max}}$, defined as the distance that generates the highest probability for the given look direction. This can be written as

$$\Lambda^\theta_t = \frac{p(\Gamma^\theta_t)}{p(\Gamma^\theta_{\text{max}})} \qquad (6)$$

Note that Λ^θ_t is a real-valued function between 0 and 1. Thus, the filter, applied to the array output signal, controls the gain only. Because the phase is not compensated, this time-varying filter shares the same properties as other gain-based noise suppression algorithms. It can significantly increase the output SNR, but also cause significant distortion and artifacts.

4. LOG-MMSE ADAPTIVE BEAMFORMING

The adaptive beamformer described in Section 2 assumes no prior knowledge of the desired source signal D_t. However, the spatial filter described in Section 3 generates an estimate of the *magnitude* of the desired source signal $|\hat{D}|$. Thus, in this section we derive an adaptive beamformer that utilizes this estimate.

4.1. Unconstrained Log-MMSE Beamforming

The first beamformer we describe is a minimum mean squared error beamformer in the log domain. As described in Section 1, operating in the log domain rather than in the magnitude or power spectral domains has advantages related to perceptual relevance and robustness to errors in estimated spectral magnitudes.

We now define the error function simply as the mean squared error of the log spectra of the desired signal and the array output:

$$W = \underset{W}{\operatorname{argmin}}\; E\left[(\log(|D|^2) - \log(|Y|^2))^2\right] \qquad (7)$$

Since we are interested in online adaptation, we replace the expectation with the instantaneous error ϵ_t.

$$\epsilon_t = \frac{1}{2}(\log(|D_t|^2) - \log(|Y_t|^2))^2 \qquad (8)$$

We can now take the derivative of (8) with respect to the filter parameters.

$$\frac{\partial \epsilon}{\partial W} = -\frac{(\log(|D_t|^2) - \log(|Y_t|^2))}{|Y_t|^2} X_t X_t^H W_t \tag{9}$$

$$= -(\log(|D_t|^2) - \log(|Y_t|^2)) \frac{X_t}{Y_t} \tag{10}$$

Using (9) the gradient descent update rule can be written as

$$W_{t+1} = W_t - \mu \left[\log(|Y_t|^2) - \log(|D_t|^2)\right] \frac{X_t}{Y_t} \tag{11}$$

The update equation (11) defines an unconstrained adaptive beamformer. If we have reliable estimates of the desired signal this approach may be sufficient. However, if the desired signal approaches zero, an unconstrained adaptive beamformer may approach the degenerate solution $W_t = 0$. Therefore, it may be desirable to impose a constraint on the adaptation.

4.2. Linearly Constrained Log-MMSE Beamforming

By following Frost's derivation, we can impose a linear constraint on the unconstrained beamformer described in Section 4.1. We now assume that our beamformer is operating with a desired look direction that specifies C and a desired array response in that direction that specifies $\mathcal{F}$.

Thus, in this case, our objective function becomes:

$$\epsilon_t = \frac{1}{2}(\log(|D_t|^2) - \log(|Y_t|^2))^2 + \lambda(W^H C - \mathcal{F}) \tag{12}$$

Taking the gradient of (12) produces the following gradient expression:

$$\nabla_W \epsilon_t = -(\log(|D_t|^2) - \log(|Y_t|^2)) \frac{X_t}{Y_t} + \lambda C \tag{13}$$

This produces the following constrained update expression:

$$W_{t+1} = W_t - \mu \left[(\log(|Y_t|^2) - \log(|D_t|^2)) \frac{X_t}{Y_t} + \lambda C\right] \tag{14}$$

which must satisfy the linear constraint

$$C^H W_{t+1} = \mathcal{F} \tag{15}$$

where we have assumed a real-valued function for $\mathcal{F}$ so that $C^H W = W^H C$.

The value of for λ can be found by substituting (14) into (15). Finally, by substituting this value back into (14) and rearranging terms, we obtain the final update expression:

$$W_{t+1} = P\left[W_t - \mu(\log(|Y_t|^2) - \log(|D_t|^2)) \frac{X_t}{Y_t}\right] + F \tag{16}$$

where P and F are defined as in Section 2.

4.3. Using a variable constraint

During processing, there may be times we would like to have the constraint active or inactive. For example, in long periods of silence we would like to run the beamformer in a constrained mode to prevent the filter weights from degenerating to the zero solution, while during periods of desired signal activity, we would like the beamformer to best match the estimated log spectrum of the desired signal, irrespective of any constraints. By comparing (11) and (16), it is obvious that these two modes of operation can be combined into a single update equation given by

$$W_{t+1} = \tilde{P}\left[W_t - \mu(\log(|Y_t|^2) - \log(|D_t|^2)) \frac{X_t}{Y_t}\right] + \tilde{F} \tag{17}$$

where

$$\tilde{P} = \begin{cases} I - C(C^H C)^{-1} C^H, & \text{if VAD} = 0 \\ I, & \text{if VAD} = 1 \end{cases} \tag{18}$$

and

$$\tilde{F} = \begin{cases} C(C^H C)^{-1} \mathcal{F}, & \text{if VAD} = 0 \\ 0, & \text{if VAD} = 1 \end{cases} \tag{19}$$

4.4. Nonlinear NLMS updates

Because we are operating on log spectral values, finding an optimal value of W_t requires a nonlinear iterative optimization method. However, because of the nonlinearity between the log spectral observations and the linear beamformer weights, the objective function is no longer quadratic. As a result, methods for improving the convergence of LMS algorithms, e.g. Normalized LMS (NLMS), cannot be applied.

In order to improve convergence, we utilize the Nonlinear NLMS algorithm introduced in [5]. In this, method, the step size is normalized by the norm of the gradient of the output signal, $\log(|Y|^2)$, with respect to the parameters being optimized, W. This results in the following normalized step size expression:

$$\mu = \frac{\tilde{\mu}}{\left(\frac{\partial \log(|Y_t|^2)}{\partial W_t}\right)^H \left(\frac{\partial \log(|Y_t|^2)}{\partial W_t}\right)} = \frac{\tilde{\mu}}{\frac{X^H X}{|Y|^2}} \tag{20}$$

where $0 < \tilde{\mu} < 1$.

5. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed Log-MMSE adaptive beamforming algorithm, we performed a series of experiments on microphone array data recorded in an office environment with a reverberation time of 270 ms. We used a 4-element linear microphone array with a length of 190 mm. The microphones in the array are electret directional elements with a cardioid directivity pattern. Incoming audio was sampled at 16 kHz and segmented into 20 ms frames using a Hann window with a 10 ms overlap. The frames were then converted to the frequency domain using a STFT.

Two recordings were used for the evaluation. The first has a user speaking directly in front of the array (0 degrees) at a distance of 1.5 m. There is a high degree of ambient noise due to the presence of several computers and the air conditioning. In addition, there is a white noise interference source located at a distance of 2 m at an incident angle of -40^o. In the second recording, the white noise point source was replaced by a radio playing pop music. In experiments that required a Voice Activity Detector (VAD), the well-known algorithm by Kim et al. was used [6].

We compared the performance of the conventional delay and sum beamformer (D&S), the Frost beamformer, the Nonlinear Spatial Filter (NSF) described in Section 3 applied to the output of the D&S beamformer, and the proposed Log-MMSE adaptive beamformer. The Log-MMSE beamformer was evaluated in three modes discussed in Section 4: no constraints, a linear constraint,

	PESQ		
	WN	MUS	AVG
D&S	2.26	2.14	2.20
Frost	2.26	2.10	2.18
D&S + NSF	2.14	1.99	2.06
D&S + NSF + Log-MMSE (Unc)	2.35	2.47	2.41
D&S + NSF + Log-MMSE (Con)	2.27	2.09	2.18
D&S + NSF + Log-MMSE (Var)	2.27	2.09	2.18

Table 1. PESQ values obtained by various processing algorithms for speech captured in an office environment with a white noise interference source (WN) and a music interference source (MUS). The average performance is also shown

	SNR (dB)		
	WN	MUS	AVG
D&S	12.98	12.59	12.73
Frost	13.90	13.38	13.64
D&S + NSF	20.81	21.59	21.20
D&S + NSF + Log-MMSE (Unc)	17.40	18.89	18.15
D&S + NSF + Log-MMSE (Con)	14.37	13.80	14.09
D&S + NSF + Log-MMSE (Var)	14.33	13.70	14.02

Table 2. SNR values in dB obtained by various processing algorithms for speech captured in an office environment with a white noise interference source (WN) and a music interference source (MUS). The average performance is also shown

and a VAD-dependent constraint. In all constrained beamformer algorithms, a distortionless constraint was used, i.e. $\mathcal{F} = 1$.

Tables 1 and 2 show the PESQ scores [7] and SNR obtained by the various algorithms. The tables show that the Frost algorithm generates a small increase in SNR but a negligible gain in perceptual quality, as indicated by the PESQ score. The Nonlinear Spatial Filter provides a 7 dB increase in SNR but results in a degradation of the PESQ score. The proposed unconstrained Log-MMSE beamformer that uses the spatial filter to generate an estimate of the target signals log spectrum generates both an increase in SNR of about 5.5 dB and a 0.2 absolute increase in the PESQ score. The Log-MMSE beamformer with a linear constraint or a VAD-dependent constraint resulted in a small increase in SNR but negligible difference in PESQ.

We also performed an experiment to highlight the difference between using the complex spectrum of the target signal and using the log spectrum of the target signal. We compared the performance of an unconstrained adaptive beamformer that was running under an MMSE objective function, i.e. $\epsilon_t = (D_t - Y_t)^2$ versus the Log-MMSE objective function defined in (8). We also compared the performance when the target signal was perfectly known (using a close-talking microphone signal) and when the target signal was estimated using the NSF. The PESQ scores are shown in Table 3. As the table shows, when the reference is known completely, using a MMSE criterion is significantly better than a Log-MMSE criterion. This makes intuitive sense as a global optimum can be found to the MMSE optimization, whereas the nonlinear optimization required in the Log-MMSE case can only guarantee a local optimum. However, when the the target signal is unknown and must be estimated, the performance of MMSE optimization degrades significantly, losing almost 1.2 points in PESQ score. In contrast, the Log-MMSE criterion is much more robust, only losing 0.28 compared to the oracle case.

		PESQ		
Criterion	Target	WN	MUS	AVG
MMSE	Closetalk	3.07	3.14	3.11
Log-MMSE	Closetalk	2.65	2.73	2.69
MMSE	D&S + NSF	1.92	1.96	1.94
Log-MMSE	D&S + NSF	2.35	2.47	2.41

Table 3. Log-MMSE vs. MMSE adaptive beamforming with oracle and estimated target signals

6. CONCLUSIONS

In this paper we have presented a new adaptive beamforming algorithm that operates according to a Log Minimum Mean Squared Error (Log-MMSE) criterion. Most existing adaptive beamforming algorithms operate in a "signal agnostic" manner and do not use any knowledge about the target signal. In the proposed algorithm, an estimate of the log spectrum of the desired signal is generated by a nonlinear spatial filter (NSF). This estimate is then used to drive the adaptation of a linear beamformer. This algorithm enables us to use the high SNR output from the NSF in a linear beamformer that does not produce artifacts or musical noise. The efficacy of the proposed method was shown in a series of experiments that showed gains in both SNR and PESQ.

In the future, we would like to further investigate the reasons why the constrained or VAD-dependent modes of the Log-MMSE beamformer did not perform as well as the unconstrained mode. In addition, the constrained version of the algorithm can be cast into the framework of the Generalized Sidelobe Canceler, which will enable unconstrained nonlinear optimization. By doing so, we will be able to explore alternative methods for estimating the optimal step size.

7. REFERENCES

[1] M. Brandstein and D. Ward, Eds., *Microphone Arrays - Signal Processing Techniques and Applications*, Springer-Verlag, New York, 2001.

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

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

[4] I. Tashev and A. Acero, "Microphone array post-processor using instantanous direction of arrival," in *Proc. IWAENC*, Paris, France, 2006.

[5] S. Kalluri and G. R. Arce, "A general class of nonlinear normalized adaptive filtering algorithms," *IEEE Trans. Sig. Proc.*, vol. 47, no. 8, pp. 2262–2272, Aug. 1999.

[6] N. S. Kim J. Sohn and W. Sung, "A statistical model-based voice activity detection," *IEEE Sig. Proc. Lett.*, vol. 6, no. 1, Jan. 2000.

[7] "ITU-T recommendation P.892. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Geneva, Switzerland 2001.