LEAST-SQUARES ERROR BEAMFORMING USING MINIMUM STATISTICS AND MULTICHANNEL FREQUENCY-DOMAIN ADAPTIVE FILTERING

Robert Aichner, Wolfgang Herbordt, Herbert Buchner, Walter Kellermann

University of Erlangen-Nuremberg
Multimedia Communications and Signal Processing
Cauerstr. 7, D-91058 Erlangen, Germany
{aichner,herbordt,buchner,wk}@LNT.de

ABSTRACT

In this paper we introduce a novel adaptive beamformer which also copes with incoherent background noise. After derivation of the optimum filter based on a weighted time-domain least-squares error criterion we present an efficient realization by applying a multichannel frequency-domain algorithm exhibiting RLS-like convergence. For the computation of this algorithm a simultaneous estimation of the power spectral density matrices of both, the noise signal and the noisy speech signal is necessary. Hence, we propose to use a novel approach based on minimum statistics to achieve this simultaneous estimation. Furthermore, the necessary estimate of a desired signal is generated by using single-channel spectral subtraction. The musical noise is avoided in our approach due to the inherent temporal and spatial averaging of our proposed beamformer. Experimental results show that the algorithm is well-suited for diffuse noise environments (e.g. car noise). Moreover, subjective listening tests confirm that a high speech quality can be obtained.

1. INTRODUCTION

Using microphone arrays for the task of speech enhancement has been a long-standing research interest in the signal processing community. Good results are already achieved in the predominantly coherent noise case (e.g. [1]) but there is still much ongoing research for diffuse noise environments (e.g. [2]).

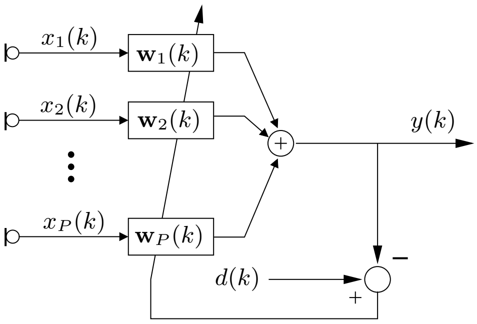


Fig. 1. LSE beamformer.

In this paper we propose an optimization criterion which is, in contrast to many traditional derivations, based on a time-domain least-squares criterion. After formulating a least-squares error (LSE) cost function with an exponential forgetting factor we can derive the optimum filters $\mathbf{w}_i$ by setting the gradient to zero. Fig. 1 shows the structure of the resulting LSE beamformer.

There are some problems to overcome when implementing this optimum filter. In the formulation of the cost function we assume the presence of a desired signal. As the clean speech signal is obviously not accessible, we have to use an estimate for the desired signal. This estimate is generated by applying a spectral subtraction based on minimum statistics [3, 4]. Due to the temporal and spatial averaging the generation of musical noise in the beamformer output is prevented. Therefore the spectral subtraction is tuned aggressively to obtain a high upper bound for the achievable SNR which is given by the SNR of the desired signal.

Moreover, the derivation of the optimum filter leads to a multichannel adaptive algorithm where often extremely ill-conditioned correlation matrices have to be inverted. As the input signals are not only auto-correlated but also highly cross-correlated, recursive least-squares (RLS) algorithms provide optimum convergence speed as they explicitly take the cross-correlations into account. To obtain a computationally feasible RLS algorithm, we apply a recently derived [5] frequency-domain multichannel algorithm that exhibits RLS-like convergence properties.

Additionally, to compute the optimum filter we have to estimate the correlation matrices of the noisy speech signal and of the noise signal simultaneously. As opposed to the approach in [6] where a voice activity detector is used to distinguish between speech and noise characteristics, we propose to use the minimum statistics approach [4] in our frequency-domain realization. Thus, we are able to simultaneously track the noise and noisy signal power spectral density (psd). To obtain also the cross-power spectral densities we detect the appearance of minima in each frequency bin, i.e., the moment when only the noise signal is present and hence we are able to estimate also the cross-power spectral densities.

Results of using the proposed algorithm in real-world scenarios are presented which show the capability of this method. Moreover, subjective listening tests confirm that a high speech quality can be obtained.

The paper is organized as follows: First we derive the optimum filter based on the least-squares error cost function. Second we describe the RLS-like multichannel frequency-domain algorithm. This is followed by a description of how to estimate

This work was partly supported by the ANITA project funded by the European Commission under contract IST-2001-34327 and by a grant from Intel Corp., Hillsboro, OR.

the cross-spectral density matrices and the desired signal by using minimum statistics. In the end we show some experimental results.

2. OPTIMUM LEAST-SQUARES ERROR BEAMFORMING

In this section we introduce the signal model and propose a least-squares error (LSE) criterion for our beamformer setup as shown in Fig. 1. In our notation lower case boldface and upper case boldface denote vectors and matrices, respectively. Superscripts ^{T} and ^{H} denote vector or matrix transposition and complex conjugate transposition. We assume the presence of a desired speech signal s and a wideband noise signal n. Thus the sensor signals in each channel $i = 1, \ldots, P$ may be written as

$$x_i(k) = s_i(k) + n_i(k). \tag{1}$$

The output of the beamformer is obtained by convolution of the sensor data with time-varying FIR filter impulse responses and summation

$$y(k) = \mathbf{w}^T(k)\mathbf{x}(k) \tag{2}$$

where the P time-varying beamformer filters $\mathbf{w}_i(k) = [w_{0,i}(k), w_{1,i}(k), \ldots, w_{L-1,i}(k)]^T$ are combined in a tap-stacked $PL \times 1$ weight vector

$$\mathbf{w}(k) = \left[\mathbf{w}_1^T(k), \mathbf{w}_2^T(k), \ldots, \mathbf{w}_P^T(k)\right]^T. \tag{3}$$

Accordingly, the $PL \times 1$ sensor data vector $\mathbf{x}(k)$ is defined as

$$\begin{aligned} \mathbf{x}(k) &= \left[\mathbf{x}_1^T(k), \mathbf{x}_2^T(k), \ldots, \mathbf{x}_P^T(k)\right]^T, & (4) \\ \mathbf{x}_i(k) &= [x_i(k), x_i(k-1), \ldots, x_i(k-L)]^T. & (5) \end{aligned}$$

Similarly to (1) we can write the sensor data vector as combination of speech signal $\mathbf{s}(k)$ and noise signal $\mathbf{n}(k)$

$$\mathbf{x}(k) = \mathbf{s}(k) + \mathbf{n}(k) \tag{6}$$

2.1. Weighted error criterion

We derive an optimum beamformer for non-stationary signals in the time domain in a least squares error (LSE) sense. To take the non-stationarity into account we formulate the beamformer using time averages over finite data blocks instead of using stochastic expectations. As the speech signal $s(k)$ is not accessible, we can only obtain an estimate for the speech component $s(k)$ which is denoted as desired signal $d(k)$. Hence, we define the estimation error $e(k)$ as the difference between the multichannel filter output and the desired signal $d(k)$ (see Fig. 1)

$$\begin{aligned} e(k) &= d(k) - \mathbf{w}^T(k)\mathbf{x}(k) & (7) \\ &= d(k) - \mathbf{w}^T(k)\left(\mathbf{s}(k) + \mathbf{n}(k)\right). \end{aligned}$$

To express the estimation error $e(k)$ (7) in a block-by-block manner we introduce the desired signal vector $\mathbf{d}(m)$ of size $L \times 1$ and the $PL \times L$ data matrix $\mathbf{X}(m)$

$$\begin{aligned} \mathbf{d}(m) &= [d(mL), d(mL+1), \ldots, d(mL+L-1)]^T & (8) \\ \mathbf{X}(m) &= [\mathbf{x}(mL), \mathbf{x}(mL+1), \ldots, \mathbf{x}(mL+L-1)]. & (9) \end{aligned}$$

where m denotes the block index. The data matrix $\mathbf{X}(m)$ can be splitted according to (1) into a matrix with speech signal components $\mathbf{X}_s(m)$ and a matrix with interference components $\mathbf{X}_n(m)$. Additionally, similar to [6] we introduce a weighting factor β to allow a trade-off between signal distortion and residual noise. We can now write the block error $\mathbf{e}(m)$ as

$$\mathbf{e}(m) = \mathbf{d}(m) - \left(\mathbf{X}_s^T(m) + \beta\mathbf{X}_n^T(m)\right)\mathbf{w}(mL), \tag{10}$$

Thus we can formulate the weighted LSE cost function with the exponential forgetting factor λ $(0 \leq \lambda \leq 1)$

$$\mathcal{J}_{LSE}(m) = (1-\lambda)\sum_{b=1}^{m} \lambda^{m-b}\mathbf{e}^T(b)\mathbf{e}(b). \tag{11}$$

2.2. Optimum filter

In the following we assume that the speech signal and the noise signal are mutually orthogonal, i.e. $\mathbf{X}_s(m)^T\mathbf{X}_n(m) = 0$. The noise signal, however, may be correlated between the different sensors. We can now derive an optimum filter $\mathbf{w}_{LSE,o}(mL)$ by taking the derivative of (11) with respect to $\mathbf{w}(mL)$ and by setting the gradient to zero. Thus we obtain

$$\mathbf{w}_{LSE,o}(mL) = \left(\Phi_x(m) + \rho\Phi_n(m)\right)^{-1}\mathbf{X}(m)\mathbf{d}(m) \tag{12}$$

where the weighting factor $\rho = \beta^2 - 1$ and the recursive estimates of the cross-correlation matrices with respect to (w.r.t.) the sensor signals $\Phi_x(mL)$ and w.r.t. interference $\Phi_n(mL)$ are given by

$$\begin{aligned} \Phi_x(m) &= \lambda\Phi_x(m-1) + (1-\lambda)\mathbf{X}(m)\mathbf{X}^T(m) & (13) \\ \Phi_n(m) &= \lambda\Phi_n(m-1) + (1-\lambda)\mathbf{X}_n(m)\mathbf{X}_n^T(m). & (14) \end{aligned}$$

3. REALIZATION IN THE FREQUENCY DOMAIN

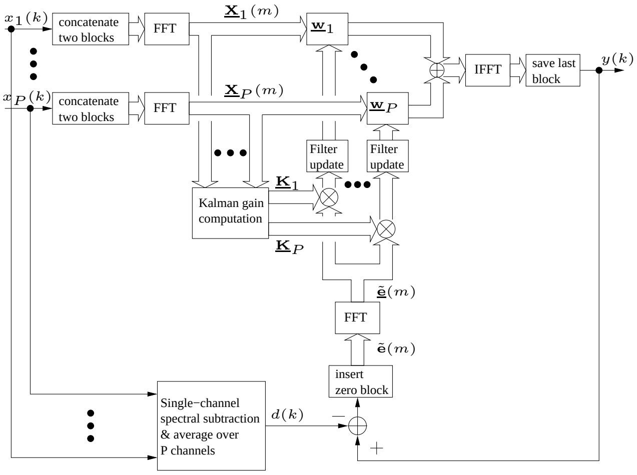
The direct realization of (12) requires the computation of a matrix inverse of size $PL \times PL$ (typically: $L = 64 \ldots 512$ and $P = 2 \ldots 16$). In order to reduce the computational complexity, we propose to determine (12) in the discrete Fourier transform domain (DFT) using an efficient multichannel frequency-domain algorithm [5]. The estimation of the cross-correlation matrices Φ_x, Φ_n and of the desired signal $\mathbf{d}(m)$ will be based on minimum statistics.

3.1. Efficient RLS-like multichannel frequency-domain algorithm

The performance of multichannel adaptive filtering algorithms depends strongly on the choice of the adaptation algorithm. This is due to the very ill-conditioned nature of the underlying normal equation (12) of the optimization problem to be solved iteratively. For such applications, the recursive least-squares (RLS) algorithm is known to be the optimum choice in terms of convergence speed as it exhibits properties that are independent of the eigenvalue spread.

The adaptive filters in our system are efficiently updated in the frequency (DFT) domain in a block-by-block fashion, using the Fast Fourier Transform (FFT) as a powerful vehicle. As a result of this block processing, the arithmetic complexity is significantly reduced compared to time-domain adaptive algorithms while desirable RLS-like properties are maintained. The possibility to exploit the efficiency of FFT algorithms is due to the Toeplitz structure of the matrices $\mathbf{X}_s$ and $\mathbf{X}_n$ involved in (10), which results from the time-shift properties of the input signals. Consequently, by rewriting the original time-domain block error signal (10) in the frequency domain and then introducing an analogous frequency-domain cost function allows a mathematically rigorous derivation of single- and multichannel frequency-domain adaptive algorithms, as shown in [5].

In Fig. 2 the frequency-domain LSE beamforming algorithm is depicted and Table 1 summarizes the necessary steps to compute the algorithm. There $\underline{\mathbf{w}}_i$ and $\underline{\mathbf{d}}$ denote the frequency-domain representations of the estimated beamformer weights of channel i and of the desired signal, respectively. The block output signal is written as $\mathbf{y}(m) = [y(mL), \ldots, y(mL+L-1)]^T$ and the matrix $\mathbf{F}$ is the DFT matrix of size $2L \times 2L$. Compared to (12) we now only have to invert a $P \times P$ matrix in each frequency bin for computing the frequency-domain Kalman gain $\underline{\mathbf{K}}(m)$. This can be done very efficiently as shown, e.g., in [5].



$x_1(k)$
concatenate two blocks
FFT
$\underline{\mathbf{X}}_1(m)$
$\underline{\mathbf{w}}_1$
IFFT
save last block
$y(k)$
$x_P(k)$
concatenate two blocks
FFT
$\underline{\mathbf{X}}_P(m)$
$\underline{\mathbf{w}}_P$
Filter update
Filter update
$\underline{\mathbf{K}}_1$
Kalman gain computation
$\underline{\mathbf{K}}_P$
$\underline{\tilde{\mathbf{e}}}(m)$
FFT
$\tilde{\mathbf{e}}(m)$
insert zero block
Single−channel spectral subtraction & average over P channels
$d(k)$
−
+

Fig. 2. Frequency-domain implementation of the LSE beamformer.

The main difficulty in implementing this algorithm is the need to estimate the desired signal and the cross-power spectral density matrix w.r.t. sensor signals and w.r.t. noise signals. This will be addressed in the next section.

3.2. Estimation of spectral density matrices and of the desired signal with minimum statistics

In this paper we propose to use the minimum statistics approach [4] for the estimation of the noise characteristics. This method is based on the observation that the power of a noisy speech signal frequently decays to the power of the background noise. Hence the general idea is to track the minima of the psd of the sensor signals in each frequency bin without any distinction between speech activity and speech pause. The minima are estimates of the psd of the noise. In [4] it was also shown that for an accurate estimate of the noise psd a time and frequency-dependent psd smoothing and bias compensation is necessary. Thus unlike in [6] where a voice activity detector is used to subsequently estimate either the noise psd or the noisy speech psd, we can perform this estimation simultaneously.

Definitions
$i, j = 1, \cdots, P$ (number of microphone channels) $\mathbf{W} = \text{diag}\{[\mathbf{0}_{1\times L}\mathbf{1}_{1\times L}]\}$ $\mu \leq 2$
Algorithm
Microphone signals: $\underline{\mathbf{X}}_i(m) = \text{diag}\left\{\mathbf{F}\left[x_i(mL-L+1)\cdots x_i(mL+L)\right]^T\right\}$ $\underline{\mathbf{X}}(m) = [\underline{\mathbf{X}}_1(m), \ldots, \underline{\mathbf{X}}_P(m)]$ $\underline{\mathbf{X}}(m) = \underline{\mathbf{X}}_s(m) + \underline{\mathbf{X}}_n(m)$ *Power spectrum estimation w.r.t. sensor and noise signals:* $\mathbf{S}(m) = \lambda\mathbf{S}(m-1) + (1-\lambda)\underline{\mathbf{X}}^H(m)\underline{\mathbf{X}}(m)$ $\mathbf{S}_n(m) = \lambda\mathbf{S}_n(m-1) + (1-\lambda)\underline{\mathbf{X}}_n^H(m)\underline{\mathbf{X}}_n(m)$ $\tilde{\mathbf{S}}(m) = \mathbf{S}(m) + \rho\mathbf{S}_n(m)$ *Kalman gain computation:* $\underline{\mathbf{K}}(m) = (1-\lambda)\tilde{\mathbf{S}}^{-1}\underline{\mathbf{X}}^H(m)$ *Filtering:* $\underline{\mathbf{y}}(m) = \sum_{i=1}^{P}\underline{\mathbf{X}}_i(m)\underline{\mathbf{w}}_i(m)$ $\tilde{\mathbf{e}}(m) = \left[\mathbf{0}_{1\times L}\mathbf{d}^T(m)\right]^T - \mathbf{W}\mathbf{F}^{-1}\underline{\mathbf{y}}(m)$ $\underline{\tilde{\mathbf{e}}}(m) = \mathbf{F}\tilde{\mathbf{e}}(m)$ $\underline{\mathbf{w}}_i(m+1) = \underline{\mathbf{w}}_i(m) + \mu\underline{\mathbf{K}}_i(m)\underline{\tilde{\mathbf{e}}}(m)$ *Output signal:* $\mathbf{y}(m) = \mathbf{W}\mathbf{F}^{-1}\underline{\mathbf{y}}(m)$

Table 1. $P \times 1$-channel frequency-domain adaptive filtering

However, in the frequency-domain counterpart of (13), (14) described in Table 1, not only the psd but also the cross-power spectral densities of the noisy signal x_i and the background noise n_i are required. They are estimated and averaged recursively for each frequency bin whenever we detect a minimum (i.e. speech pause) of the noisy speech signal. This method gives an accurate estimate of the noise spectral density matrix for slowly time-varying noise statistics.

The other difficulty we have to deal with is the need of a desired signal $d(k)$. In [6] this was circumvented by selecting a one microphone channel as a reference channel and then simply subtracting the estimated noise signal from the reference channel and using the result as estimate of the desired signal.

Here we propose to use the noise characteristics estimated with minimum statistics and then apply a single-channel spectral subtraction rule [3]. This allows a more sophisticated estimation of the desired signal than a simple subtraction in the time-domain. To increase robustness we average the desired signal over all P channels. In single-channel noise reduction we can often perceive musical noise resulting from the nonstationarity of the noise background. This effect can be decreased by the introduction of an oversubtraction factor and a limitation of the maximum subtraction by a spectral floor constant. However, the residual musical noise is still audible.

In our algorithm, the LSE beamformer is trying to adapt to the desired signal $d(k)$ (i.e. the single-channel spectral subtraction solution) in the least-squares sense. Therefore the upper bound of the achievable SNR by using LSE beamforming is given by the SNR of the desired signal. The desired signal is generated by single-channel noise reduction, where always a tradeoff between musical noise and SNR enhancement exists. To obtain a high upper bound we will tune the spectral subtraction parameters aggressively to obtain a desired signal $d(k)$ with high SNR leading to high musical noise.

Due to the temporal averaging introduced by using the recursive estimation of the cross-power spectral density matrices of sensor and noise signals $\mathbf{S}(m), \mathbf{S}_n(m)$, we can avoid the generation of musical noise in our multichannel frequency-domain adaptive algorithm. Additionally also the spatial averaging over all channels P is contributing to avoid musical noise in the output $y(k)$. This means we can combine in our LSE beamformer a high noise suppression while avoiding musical noise.

Additionally it was shown in [1, 7] that we can decompose our system into an MVDR beamformer and a frequency-dependent postfilter. Thus if the spatial degrees of freedom are already sufficient for noise suppression, then the output of the LSE beamformer would be undistorted. If any residual noise remains then it is further reduced by the frequency-dependent postfilter which by nature also introduces desired signal distortion.

4. EXPERIMENTAL RESULTS

We have evaluated our algorithm on real data which was recorded with an 8-channel microphone array mounted at the sun visor position of the co-driver in a Nissan Patrol 4×4 car driving in a suburban area with 80km/h. The inter-element spacing was chosen to $d = 4$cm and the sampling rate f_s is 8 kHz.

For a diffuse environment the magnitude squared coherence (MSC) is given by [8]

$$\Gamma^2(\Omega) = \frac{\sin^2(\Omega f_s d c^{-1})}{(\Omega f_s d c^{-1})^2}, \tag{15}$$

where c is the velocity of sound. As it can be seen in Fig. 3, our recorded data exhibits relatively diffuse noise components. The reverberation time T_{60} of the car was approximately 80ms. We compared our algorithm to a filter & sum beamformer with Dolph-Chebyshev windowing by using the segmental signal-to-noise ratio (SSNR) as an objective performance measure. The SSNR was calculated by using a frame size of 16 ms and excluding frames whose power was below the long-term average of the whole clean speech signal. The input signal-to-noise ratio (SNR) was varied between 0 dB and 10 dB. The length of the beamformer filters was chosen to $L = 256$. There was no audible degradation of the speech signal observed. Speech signals are available for listening in [9].

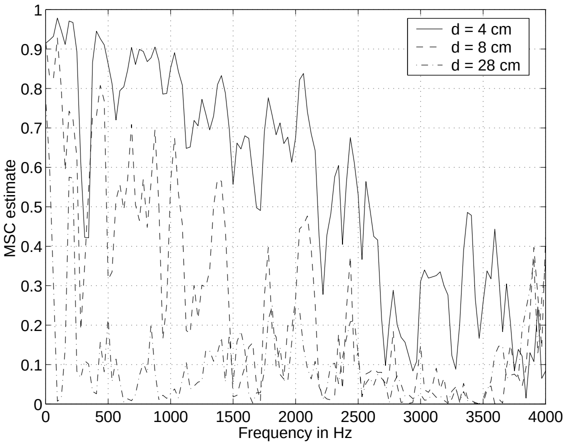


Fig. 3. Magnitude squared coherence of background noise (FFT length 256).

SNR Input	0 dB	5 dB	10 dB
SSNR Input in dB	-3.7	1.3	6.3
SSNR of Filter & Sum in dB	-1.2	3.8	8.8
SSNR of LSE beamformer in dB	1.4	6.4	11.0

Table 2. Segmental SNR results.

5. CONCLUSIONS

We have proposed to apply a weighted least-squares optimization criterion which can be implemented efficiently as a multichannel frequency-domain adaptive algorithm with RLS-like convergence properties. We applied a minimum statistics method to obtain a simultaneous estimation of the noise signal and noisy speech signal spectral density matrices. Based on the estimated noise psd we generated a desired signal with a spectral subtraction rule that was adjusted to give high SNR enhancement. Musical noise was avoided in the LSE beamformer due to the inherent temporal and spatial averaging. Experimental results on real data show the effectiveness of our method.

6. REFERENCES

[1] W. Herbordt and W. Kellermann, "Adaptive beamforming for audio signal acquisition," in *Adaptive signal processing: Application to real-world problems*, J.Benesty and Y.Huang, Eds., pp. 155–194. Springer, Berlin, Jan. 2003.

[2] I. McCowan and H. Bourlard, "Microphone array post-filter for diffuse noise field," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Orlando, FL, USA, May 2002, vol. 1, pp. 905–908.

[3] R. Martin, "Spectral subtraction based on minimum statistics," in *Proc. EUSIPCO*, Sept. 1994, pp. 1182–1185.

[4] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, July 2001.

[5] H. Buchner, J. Benesty, and W. Kellermann, "Multichannel frequency-domain adaptive filtering with application to acoustic echo cancellation," in *Adaptive signal processing: Application to real-world problems*, J.Benesty and Y.Huang, Eds., pp. 95–128. Springer, Berlin, Jan. 2003.

[6] D.A. Florêncio and H.S. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Salt Lake City, Utah, USA, May 2001, vol. 1, pp. 197–200.

[7] K. U. Simmer, J. Bitzer, and C. Marro, "Post-filtering techniques," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., pp. 39–60. Springer, Berlin, 2001.

[8] H. Kuttruff, *Room Acoustics*, Elsevier Science, 3rd edition, 1990.

[9] `http://www.LNT.de/~aichner/iwaenc03.html`