BROADBAND BEAMFORMING WITH ADAPTIVE POSTFILTERING FOR SPEECH ACQUISITION IN NOISY ENVIRONMENTS*

Sven Fischer

Ericsson Eurolab Deutschland GmbH
Äußere Bayreuther Str. 350
90411 Nürnberg, Germany
e–mail: sven.fischer@eedn.ericsson.se

Karl–Dirk Kammeyer

University of Bremen, FB–1,
Department of Telecommunications
P.O. Box 33 04 40, D–28334 Bremen, Germany
e–mail: kammeyer@comm.uni-bremen.de

ABSTRACT

In this paper the implementation of a broadband beamformer which is built up by several harmonically nested subarrays for each octave band combined with optimal postfiltering is described. This method has the advantage of providing large sensor distances for the postfilter estimation by simultaneously controlling the directivity of the array. The selection of an optimal postfilter is discussed in detail and its estimation based on a Nuttall/Carter method for spectrum estimation is described. The resulting noise reduction system yields improved performance in diffuse noise fields and no distortions in the case of coherent direct path noise. Furthermore, the system is robust to steering misadjustment.

1. PROBLEM STATEMENT

To reduce the noise and reverberation in hands–free speech communication, microphone arrays consisting of multiple microphones can be used for sound pick–up. The performance of the microphone array depends on the number of sensors building up the array. Due to space limits and monetary cost only small microphone arrays can be carried out in many applications. To increase the performance of the microphone array without increasing the number of sensors adaptive signal processing techniques can be used. One class of adaptive techniques for microphone arrays is the use of an optimal filter in the beamformer output signal which transfer function is estimated from the spatial cross power densities of the microphone signals [1, 2]. The postfilter estimation is based on the assumption of spatially uncorrelated noise. This is fulfilled in most realizations by applying an undersampled array aperture, i.e. by using large sensor distances in the array [1, 2]. But this approach has two main disadvantages:

1. The large inter element spacing results in a very narrow beamwidth, especially at high frequencies and therefore, the array is very sensitive to steering misadjustment.
2. Using an undersampled aperture yields grating lobes in the visible region of the array. This can result in a severe distorted output signal in the case of coherent direct path noise.

To overcome these disadvantages we use a "true" broadband beamformer which is built up by several harmonically nested subarrays for each octave band [3–6]. This method has the advantage of providing large sensor distances for the postfilter estimation by simultaneously controlling the directivity of the array. This combination of subarray processing and adaptive postfiltering of the array output signal yields an increased noise reduction performance in diffuse noise fields and no distortions in the case of coherent direct path noise. Furthermore, the resulting noise reduction system is robust to steering misadjustment.

*This work was performed at the University of Bremen, Department of Telecommunications.

2. PRACTICAL REALIZATION

The detailed structure of the system is shown in figure 1. The array consists of nine omnidirectional microphones which are grouped into three harmonically nested linear subarrays. The array is focused to the desired signal source using time delay compensation. The microphone signals are windowed and transformed into the frequency domain to obtain the short time spectra $X_i(nL, v)$, were L is the overlapping of the analysis windows, n is the discrete time and v the frequency index. Aiming a telephone bandwidth, arrays for a low–frequency (LF), mid–frequency (MF) and high frequency (HF) section were realized, each consisting of five microphones with spacing 20 cm, 10 cm and 5 cm, respectively. As some microphones can be used for several frequency sections, the entire array consists of nine microphones. The conventional beamformer output spectrum $\overline{X}(nL, v)$ is then obtained by summing the three sections after frequency selective filtering. Each filter is designed as an 128th order FIR filter and the sum of the frequency responses of these filters approximates unity over the entire frequency range. The cut–off frequencies $f_{c,i}$ of these filters were chosen according to $f_{c,i} = c/2d_i$, were c is the speed of sound and d_i is the inter element spacing of subarray i (full steering range is allowed). The array power pattern and the directivity index of this conventional beamformer with broadside steering ($\phi_0 = \pi/2$) is shown in figure 2. Theoretically, the noise reduction of the array for localized coherent noise sources is the inverse of the array power pattern. In diffuse noise fields the noise reduction of the array is given by the directivity index, which expresses the ratio of sound energy received from the steered– (look–) direction to the average energy received from all directions.

2.1. Choice of the Postfilter

To increase the diffuse noise reduction performance of the array, postfiltering of the conventional beamformer output signal is applied. The optimal postfilter is derived from Wiener theory, which solves the problem of optimal signal estimation in the presence of additive noise [7]. If the signal $s(n)$ and the noise $n(n)$ have zero mean and are uncorrelated, the optimal filter is given by

$$W(\mathrm{e}^{\mathrm{j}\Omega}) = \frac{\Phi_{ss}(\mathrm{e}^{\mathrm{j}\Omega})}{\Phi_{xx}(\mathrm{e}^{\mathrm{j}\Omega})} = \frac{\Phi_{ss}(\mathrm{e}^{\mathrm{j}\Omega})}{\Phi_{ss}(\mathrm{e}^{\mathrm{j}\Omega}) + \Phi_{nn}(\mathrm{e}^{\mathrm{j}\Omega})} \quad , \qquad (1)$$

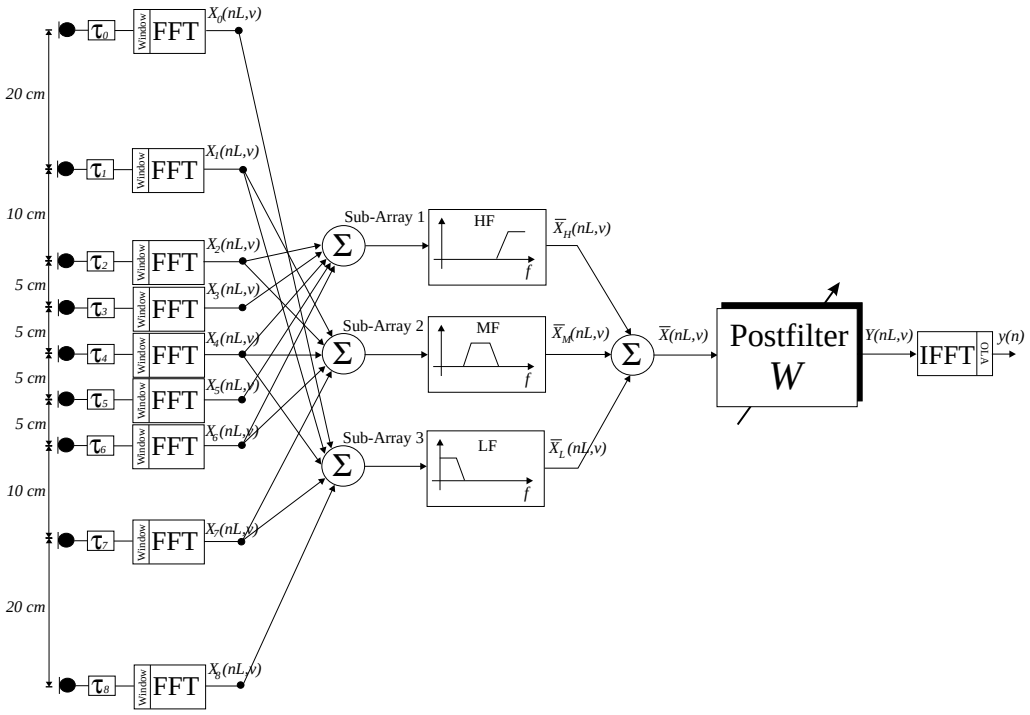

Figure 1: Microphone array with adaptive postfilter.

were Φ_{ss} and Φ_{nn} are the (a priori known) power density spectra of the signal and the noise respectively. This filter applied to the noisy observation $x(n)$ leads to an optimal estimate of the signal $s(n)$ in the mean–square error sense. It should be noted that $W(\mathrm{e}^{\mathrm{j}\,\Omega})$ is a real and thus zero–phase function. The Wiener filter weights the spectral components of the disturbed signal according to the signal–to–noise power density ratio at individual frequencies. In frequency regions where there is no signal power the spectral components are entirely suppressed; if there is no noise power, the components are entirely passed. However, in regions where the signal and noise spectra overlap, the filter not only affects the noise components, but the signal components as well. In general, the Wiener filter is therefore a biased estimator. It reduces the variance in $x(n)$ but at the cost of an increased bias, that is, a systematic reduction of the amplitude of the signal components. By minimizing the mean squared error the filter attains a compromise between these two factors.

Applying the Wiener filter theory to microphone array data, two main approaches exist: Zelinski's [1] transfer function can theoretically be expressed as

$$W_1(\mathrm{e}^{\mathrm{j}\,\Omega}) = \frac{\Phi_{ss}(\mathrm{e}^{\mathrm{j}\,\Omega})}{\overline{\Phi_{xx}}(\mathrm{e}^{\mathrm{j}\,\Omega})} \quad , \tag{2}$$

were $\overline{\Phi_{xx}}$ is the average of the power density spectra of the individual input signals x_i. Simmer and Wasiljeff's [2] transfer function on the other hand can be expressed as

$$W_2(\mathrm{e}^{\mathrm{j}\,\Omega}) = \frac{\Phi_{ss}(\mathrm{e}^{\mathrm{j}\,\Omega})}{\Phi_{\bar{x}\bar{x}}(\mathrm{e}^{\mathrm{j}\,\Omega})} \quad , \tag{3}$$

were $\Phi_{\bar{x}\bar{x}}$ is the power density spectrum of the average signal (output signal of the conventional beamformer). Since the postfilter is applied to the output signal of the conventional beamformer, W_2 is the correct filter to apply. The conventional beamformer already causes a reduction of the input noise. Applying W_1 to the beamformer output signal yields a stronger reduction of the remaining noise but at the cost of an increased bias, especially when the number of microphones N is large. This filter attains no compromise between variance and bias of the estimate and therefore, it is not optimal in the Wiener sense. This discrepancy is more relevant in our application. For single frequencies, the conventional beamformer can yield a perfect noise reduction. For these frequencies the postfilter should not affect the signal. This is only fulfilled by applying the filter W_2.

In practice, the power density spectrum of the desired speech signal Φ_{ss} can not be estimated directly. The estimation of Φ_{ss} is based on the assumption of spatially uncorrelated noise [1]. In this case the spatial cross power density of the disturbed input signals x_i equals the auto power density of the desired speech signal. Therefore, we obtain an estimate for the transfer function of the postfilter:

$$\widehat{W_2}(\mathrm{e}^{\mathrm{j}\,\Omega}) = \frac{\Phi_{x_i x_j}(\mathrm{e}^{\mathrm{j}\,\Omega})}{\Phi_{\bar{x}\bar{x}}(\mathrm{e}^{\mathrm{j}\,\Omega})} = \frac{\Phi_{ss}(\mathrm{e}^{\mathrm{j}\,\Omega}) + \Phi_{n_i n_j}(\mathrm{e}^{\mathrm{j}\,\Omega})}{\Phi_{\bar{x}\bar{x}}(\mathrm{e}^{\mathrm{j}\,\Omega})} \quad , \tag{4}$$

were $\Phi_{x_i x_j}(\mathrm{e}^{\mathrm{j}\,\Omega})$ is the spatial cross power density spectrum between microphone signals i and j (after time delay compensation of the desired signal). In the case of zero spatial correlation of the noise signals ($\Phi_{n_i n_j} = 0$) $\widehat{W_2}$ equals W_2 and we can estimate the transfer function of the postfilter by using the disturbed input signals only. But in practice, the noise field can be at best diffuse with a spatial coherence function Γ_{ij} given by a sinc function [8]:

$$\Gamma_{ij}(d_{ij}, f) = \mathrm{sinc}\left(2\frac{d_{ij}}{c}f\right) \quad , \tag{5}$$

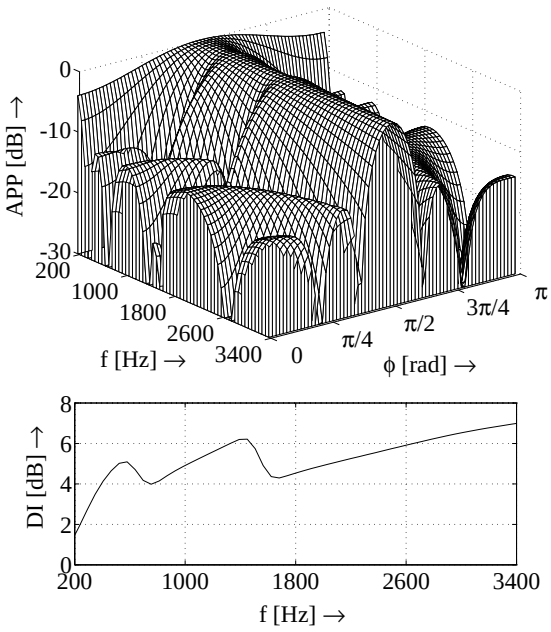

Figure 2: Beam power pattern and directivity index of the conventional beamformer.

were d_{ij} is the spacing between microphones i and j. The spatial coherence has its first zero at frequency $f = c/2d_{ij}$. Sampling the noise field at spatial locations separated by $d_{ij} > c/2f_{min}$ would yield approximately uncorrelated noise components for frequencies above f_{min}. Unfortunately, for sensor distances $d_{ij} > c/2f_{min}$ grating lobes appear in the visible region of the array. To obtain an "all–purpose" noise reduction system, i.e. which is useful in incoherent as well as in coherent noise fields, both, the beamformer and the postfilter have to be designed carefully. By applying the subarray as described above, grating lobes are avoided and sensor distances from 5 cm up to 80 cm are available for the estimation of the postfilter. Averaging the spatial cross power densities over all microphone pairs ij ($i < j$) reduces the remaining correlation of the noise and we obtain the following expression for the transfer function estimate of the postfilter:

$$\widehat{W}(\mathrm{e}^{\mathrm{j}\Omega}) = \frac{\frac{2}{N(N-1)} \sum_{i=0}^{N-2} \sum_{j=i+1}^{N-1} \Phi_{x_i x_j}(\mathrm{e}^{\mathrm{j}\Omega})}{\Phi_{\bar{x}\bar{x}}(\mathrm{e}^{\mathrm{j}\Omega})} \quad , \qquad (6)$$

were N denotes the number of microphones ($N = 9$ in our case).

2.2. Estimation of The Transfer Function $\widehat{W}$

The postfilter estimation and implementation is performed with the OLA method. The window length is 256 samples (8 kHz sampling frequency) and the overlap of the data windows is 128 samples. Each data block is transformed into the frequency domain with a FFT of size 512. To calculate the transfer function $\widehat{W}$ according to equation (6) the power density spectra $\Phi_{x_i x_j}$ and $\Phi_{\bar{x}\bar{x}}$ have to be estimated from the short time spectra. Due to the nonstationarity of the signals, only short data segments are available for spectrum estimation. The power spectra are estimated by using a short–time Nuttall/Carter spectral estimation method [9, 10]. This method can be viewed as a combination of an exponentially averaged Welch periodogram (which can be computed by a simple recursion) and the Blackman/Tuckey correlogram method. This combined method smoothes the power spectra in time and frequency and yields improved estimates within a few data segments. Additionally, the estimated transfer function approximately satisfies a linear convolution constraint when multiplied with the short time spectrum $\overline{X}(nl, v)$.

2.2.1. Postprocessing

As stated in section 2.1, the transfer function of a Wiener filter is a zero–phase function. But in general, this is not guaranteed by the estimate according to equation (6). In addition, the transfer function estimate may even take negative values despite the fact that the numerator and denumerator of the original filter (eqn. 3) are real and positive quantities. These ill effcts are due to the way of estimating the auto power density of the speech signal from the spatial cross power density of the disturbed input signals x_i. In practice, these ill effects are increased due to estimation errors.

If the noise field is purely diffuse (and if there is no misadjustment in the beam steering unit), an imaginary part of the spatial cross power density can only occur due to estimation errors, since both the auto power density of the speech signal and the spatial cross power density of a diffuse noise field are real functions. The estimation error can then be reduced by taking the real part of the numerator of equation (6) [1] (this result would also be obtained by averaging over microphone pairs $i > j$ in the numerator of equation (6) additionally). Further more, the estimate can be improved by setting the negative values of the spatial cross power density to zero. This postprocessing scheme leads to an improved estimate of $\widehat{W}$ in purely diffuse noise fields [1].

If the noise field is not purely diffuse, an imaginary part in the spatial cross power density is not only due to estimation errors. The spatial coherence of a combined noise field consisting of coherent direct path noise and diffuse noise is a complex function [11]. Then taking only the real part of the spatial cross power density leads again to an improved estimate for the auto power density of the speech signal, but at the cost of a systematic error due to the real part of the spatial coherence of the noise field. In addition, the estimate can contain negative values over a wide frequency range. Applying this estimated filter to the output signal of the conventional beamformer may result in a severe distorted output signal [11]. Instead of taking the real part of the spatial cross power density, we take the modulus of the spatial cross power densities:

$$\widehat{W}(\mathrm{e}^{\mathrm{j}\Omega}) = \frac{\left| \frac{2}{N(N-1)} \sum_{i=0}^{N-2} \sum_{j=i+1}^{N-1} \Phi_{x_i x_j}(\mathrm{e}^{\mathrm{j}\Omega}) \right|}{\Phi_{\bar{x}\bar{x}}(\mathrm{e}^{\mathrm{j}\Omega})}. \qquad (7)$$

To reduce the estimation error, we try to improve the spectral estimation by using the Nuttall/Carter method as mentioned in the beginning of subsection 2.2 (and not by neglecting any imaginary part). Therefore, the numerator of the postfilter is built by an estimate of the spatial cross spectrum of the propagating wavefield (neglecting any phase). If the noise field is spatially uncorrelated, this spectrum contains only the power density of the speech signal. If the noise field is spatially correlated, this spectrum contains the auto power density of the speech as well as that of the noise and therefore, the transfer function $\widehat{W}$ approximates unity. Due to this approach, the postfilter is an estimate for the Wiener filter for

spatially uncorrelated noise only. For coherent noise the transfer function tends to be one and the noise reduction performance of the system is only due to the conventional beamformer.

3. SIMULATION RESULTS

To evaluate the performance of the method, some computer simulations of the system were performed. To simulate the acoustical properties of the enclosure, we used Allen's image method [12] to compute the room impulse responses from the source to the microphones. The input signals were obtained by filtering the speech and noise signal with the corresponding room impulse responses. For performance evaluation we used the Log Area Ratio (LAR) distance (l_1 norm without energy weighting) as objective measure for speech quality [13]. Figure 3 shows the LAR improvement as function of the reverberation time of the enclosure (low LAR $\widehat{=}$ high speech quality). The solid line in figure 3 shows the input LAR (noise source hair drier, input SNR=3 dB, sampling frequency 8 kHz). The dash–dotted line in figure 3 shows the output LAR of the beamformer without postfiltering, the dashed line shows the output LAR with postfiltering as described above. As can be seen from this figure, for low reverberation times the performance of the system is only due to the conventional beamformer. For higher reverberation times the noise field becomes more and more diffuse. Then the postfilter yields an extra reduction of the noise and therefore an improved speech quality is obtained. For comparison, the dotted line in figure 3 shows the performance of a microphone array with postfilter using an undersampled aperture [2]. For this experiment, only the subarray with the largest sensor distance is used (subarray 3 in figure 1). This method yields a poor performance for low reverberation times (see also [9]and [11]), because of the undersampled aperture used and because of using only the real part in the numerator of the Wiener filter estimate.

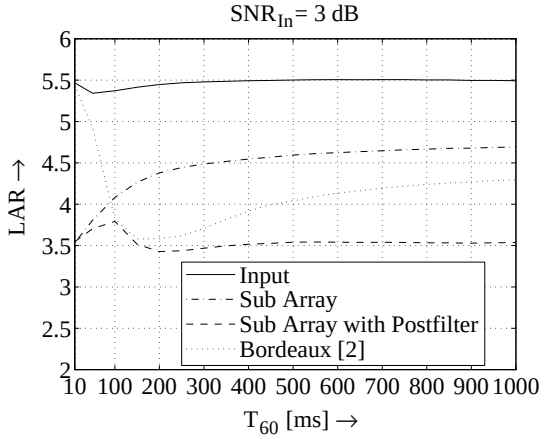

Figure 3: Log–Area–Ratio distance (LAR) as function of reverberation time (T_{60}) of the enclosure.

4. CONCLUSION

Microphone arrays with adaptive postfiltering were previously considered for diffuse noise fields only. This is because of the contradiction in the sensor element spacing requirements for the microphone array on the one hand and for the postfilter estimation procedure on the other hand. In this paper it is shown that by using a broadband beamformer which is built up by several harmonically nested linear subarrays for each octave band and by carefully designing a postfilter estimation method, the resulting noise reduction system performance is nearly independent of the correlation properties of the noise field, i.e. the system is applicable for diffuse as well as for coherent direct path noise.

5. REFERENCES

[1] R. Zelinski, "A microphone array with adaptive post–filtering for noise reduction in reverberant rooms," in *Proc. of the Internat. Conference on Acoustics, Speech and Signal Processing ICASSP–88*, (New York), pp. 2578–2581, Apr. 1988.

[2] K.U. Simmer and A. Wasiljeff, "Adaptive microphone arrays for noise suppression in the frequency domain," in *Second Cost 229 Workshop on Adaptive Algorithms in Communications*, (Bordeaux, France), pp. 185–194, 30.9.–2.10 1992.

[3] J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, and M.M. Sondhi, "Autodirective microphone systems," *Acustica*, vol. 73, pp. 58–71, 1991.

[4] W. Kellermann, "A self–steering digital microphone array," in *Proc. of the Internat. Conference on Acoustics, Speech and Signal Processing ICASSP–91*, pp. 3581–3584, 1991.

[5] Y. Mahieux, A. Gilloire, and G. Le Tourneur, "A microphone array for multimedia applications," in *1993 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, (Mohonk Mountain House New Paltz, New York), October 17–20 1993.

[6] C. Marro, Y. Mahieux, and K.U. Simmer, "Performance of adaptive dereverberation techniques using directivity controlled arrays," in *Proc. European Signal Processing Conf. EUSIPCO–96*, (Trieste, Italy), pp. 1127–1130, Sep. 1996.

[7] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. Cambridge, MA: MIT, 1964.

[8] F. Jacobsen and T.G. Nielsen, "Spatial correlation and coherence in a reverberant sound field," *J. Sound and Vibration*, vol. 118, no. 1, pp. 175–180, 1987.

[9] S. Fischer and K.U. Simmer, "Beamforming microphone arrays for speech acquisition in noisy environments," *Speech Communication*, Dec. 1996.

[10] A. H. Nuttall and G.C. Carter, "Spectral estimation using combined time and lag weighting," *Proc. IEEE*, vol. 70, no. 9, pp. 1115–1125, Sep. 1982.

[11] S. Fischer, *Adaptive Mehrkanalgeräuschunterdrückung bei gestörten Sprachsignalen unter Berücksichtigung der räumlichen Kohärenz des Geräuschfeldes*. PhD thesis, Universität Bremen, Mai 1996. Verlag Shaker, Aachen (in German).

[12] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small–room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[13] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete–Time Processing of Speech Signals*. New York: Macmillan Publishing Company, 1993.