IDIAP RESEARCH REPORT

IDIAP
Martigny - Valais - Suisse

Microphone Array Post-filter for Diffuse Noise Field

Iain A. McCowan [1] Hervé Bourlard [1, 2]

IDIAP–RR 01-39

November 2001

submitted for publication

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 − 27 − 721 77 11
fax +41 − 27 − 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

1 IDIAP—Dalle Molle Institute of Perceptual Artificial Intelligence, P. O. Box 592, CH-1920 Martigny, Switzerland, {mccowan, bourlard}@idiap.ch
2 EPFL, Lausanne

IDIAP Research Report 01-39

Microphone Array Post-filter for Diffuse Noise Field

Iain A. McCowan   Hervé Bourlard

November 2001

submitted for publication

Abstract. This paper proposes a novel technique for estimating the signal power spectral density to be used in the transfer function of a microphone array post-filter. The technique is a modification of the existing Zelinski post-filter, which uses the auto- and cross-spectral densities of the array inputs to estimate the signal and noise spectral densities. The Zelinski technique, however, assumes zero cross-correlation between noise on different sensors. This assumption is inaccurate in real conditions, particularly at low frequencies and for arrays with closely spaced sensors. In this paper we replace this with an assumption of a theoretically diffuse noise field, which is more appropriate in a variety of realistic noise environments. In experiments using noise recordings from an office of computer workstations, the modified post-filter results in significant improvement in terms of objective speech quality measures and speech recognition performance.

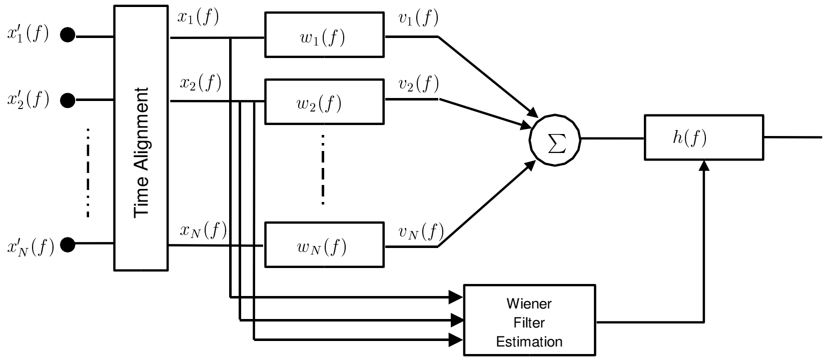

Figure 1: Filter-sum beamformer with post-filter

1 Introduction

Much research has been undertaken in recent years into the use of microphone arrays for the task of speech enhancement and robust speech recognition. Microphone arrays permit distant, hands-free signal acquisition and they provide directional discrimination, allowing for reduction of undesired noise sources and tracking of the speech source. The directional discrimination of the array is exploited by beamforming algorithms, and often the beamformer output is further enhanced by applying a post-filter. A thorough review of the motivation and theory of microphone array post-filtering techniques is presented in [1].

In this article we propose a new post-filter estimator for microphone array speech enhancement. The technique is based upon the commonly used post-filter proposed by Zelinski [2]. The Zelinski post-filter uses the input channel auto- and cross-spectral densities to estimate a Wiener post-filter to be applied to the beamformer output. The use of such a post-filter with a standard sub-array beamforming microphone array was thoroughly investigated by Marro *et al* [3], and has been used successfully in a number of speech enhancement and robust speech recognition applications.

While the Zelinski post-filter shows reasonable performance, its formulation is based upon a number of assumptions. In particular, in estimating the speech power spectral density, the assumption of zero correlation between the noise on different channels is made, corresponding to a perfectly incoherent noise field. In practice, such an incoherent noise field is seldom encountered, and the correlation of the noise between channels can be significant, particularly at low frequencies. This is especially true for closely spaced sensors, as is typically the case in speech enhancement applications.

In this article, we replace the incoherent noise assumption with an assumption of a diffuse noise field. A diffuse noise field has been shown to be a reasonable model for a number of practical noise environments, such as office and car noise, and is a common basis for superdirective beamforming techniques, such as [4]. By using the theoretically known coherence function of the noise field, we reformulate the speech power spectral density estimation in the post-filter transfer function. The behaviour of the new post-filter is investigated using a multi-channel office noise recording, and is shown to give significant performance improvement over the existing technique in terms of objective speech quality measures and speech recognition performance.

2 Zelinski Post-filter Estimator

In this section we review the Zelinski post-filter on which the proposed technique is based [2, 3].

Consider the microphone array system shown in Figure 1. It is assumed that at the output of the time alignment module, the inputs have been aligned to account for the effect of the propagation vector $\mathbf{d}(f)$. The signals at the output of the time alignment can thus be modeled as

$$x_i(f) = s(f) + n_i(f) \tag{1}$$

where $n_i(f)$ is the noise signal on channel i after time alignment for the desired signal. Calculating the auto- and cross-spectral densities of the aligned signals on channels i and j, leads to

$$\phi_{x_i x_i}(f) = \phi_{ss}(f) + \phi_{n_i n_i}(f) + \phi_{s n_i}(f) + \phi_{n_i s}(f) \tag{2}$$

and

$$\phi_{x_i x_j}(f) = \phi_{ss}(f) + \phi_{n_i n_j}(f) + \phi_{s n_j}(f) + \phi_{n_i s}(f) \tag{3}$$

Under the assumptions that:

1. the signal and noise are uncorrelated,
2. the noise is uncorrelated between sensors, and
3. the noise power spectrum is the same on all sensors,

these reduce to

$$\phi_{x_i x_i}(f) = \phi_{ss}(f) + \phi_{nn}(f) \tag{4}$$

$$\phi_{x_i x_j}(f) = \phi_{ss}(f) \tag{5}$$

The single-channel Wiener post-filter transfer function for such a microphone array is given as [1]

$$h(f) = \frac{\phi_{ss}(f)}{\phi_{ss}(f) + \phi_{nn}(f)} \tag{6}$$

From the above equations, it is evident that we can estimate the numerator and denominator of the transfer function using the cross- and auto-spectral densities of the input channels, respectively. This estimate can be further improved by averaging the spectral densities over all possible sensor combinations. This results in the post-filter estimator

$$\hat{h}_1(f) = \frac{\frac{2}{N(N-1)} \Re\{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \phi_{x_i x_j}(f)\}}{\frac{1}{N} \sum_{i=1}^{N} \phi_{x_i x_i}(f)} \tag{7}$$

where $(\cdot)^*$ is the complex conjugate operator. The real operator $\Re\{\cdot\}$ is used as the term being estimated in the numerator, $\phi_{ss}(f)$, is necessarily real.

The auto- and cross-spectral densities of the time-aligned inputs, $\phi_{x_i x_i}(f)$ and $\phi_{x_i x_j}(f)$, are estimated using the standard recursive update formula [5]

$$\phi_{x_i x_j}(f) = \alpha \phi'_{x_i x_j}(f) + (1 - \alpha) x_i(f) x_j^*(f) \tag{8}$$

where $\phi'_{x_i x_j}(f)$ and $\phi_{x_i x_j}(f)$ are the spectral estimates for the previous and current frames respectively.

3 Proposed Technique

The above post-filter formulation makes the assumption that the noise between sensors is uncorrelated. This corresponds to a perfectly incoherent noise field, a situation that will not arise in practice for closely spaced microphone arrays. While the Zelinski post-filter estimation has been shown to work well in a variety of conditions, the performance would be improved if a more accurate model of the noise field were used. A diffuse noise field is a good approximation of a number of practical noise environments encountered in speech enhancement applications, such as office and car noise. A common measure used to characterise noise fields is the *complex coherence function*. The coherence between the signals at two points, i and j, is defined as

$$\Gamma_{ij}(f) = \frac{\phi_{ij}(f)}{\sqrt{\phi_{ii}(f)\phi_{jj}(f)}} \tag{9}$$

where ϕ_{ij} is the cross-spectral density between the signals at i and j. It can be shown that the coherence of a theoretical diffuse noise field is given by

$$\Gamma_{ij}(f) = \mathrm{sinc}\left(\frac{2\pi f d_{ij}}{c}\right) \tag{10}$$

where d_{ij} is the distance between sensors i and j, and c is the propagation speed of sound, given as $340\ ms^{-1}$ in air.

Replacing the assumption of zero cross-correlation of the noise between sensors with the assumption of a diffuse noise field, we can reformulate our knowledge as

$$\phi_{x_i x_j}(f) = \phi_{ss}(f) + \phi_{n_i n_j}(f) \tag{11}$$

$$\phi_{x_i x_i}(f) = \phi_{ss}(f) + \phi_{n_i n_i}(f) \tag{12}$$

$$\phi_{x_j x_j}(f) = \phi_{ss}(f) + \phi_{n_j n_j}(f) \tag{13}$$

$$\Gamma_{n_i n_j}(f) = \frac{\phi_{n_i n_j}(f)}{\sqrt{\phi_{n_i n_i}(f)\phi_{n_j n_j}(f)}} \tag{14}$$

Given that we can calculate approximations of all the quantities in the left-hand side of the above equations, we see that they form a set of four equations with four unknown variables, ϕ_{ss}, $\phi_{n_i n_i}$, $\phi_{n_j n_j}$ and $\phi_{n_i n_j}$. Applying standard algebra techniques, these equations can be solved for any of the variables. For the post-filter transfer function we are only interested in re-estimating the numerator of Equation 6 (as the denominator estimation remains unchanged under the modified assumption), and so solving for ϕ_{ss}, we obtain a second order equation with solution

$$\hat{\phi}_{ss}(f) = \frac{-b(f) \pm \sqrt{b^2(f) - 4a(f)c(f)}}{2a(f)} \tag{15}$$

where

$$a(f) = \Gamma^2_{n_i n_j}(f) - 1 \tag{16}$$

$$b(f) = 2\phi_{x_i x_j}(f) - \Gamma^2_{n_i n_j}(f)\left(\phi_{x_i x_i}(f) + \phi_{x_j x_j}(f)\right) \tag{17}$$

$$c(f) = \Gamma^2_{n_i n_j}(f)\phi_{x_i x_i}(f)\phi_{x_j x_j}(f) - \phi^2_{x_i x_j}(f) \tag{18}$$

The ambiguity of the solution is resolved by noting that on substitution, one of the solutions corresponds to a negated coherence function and can be discarded.

A problem arises if $\Gamma_{n_i n_j}(f) = 1$. Theoretically, the solution is indeterminite in this case, as unity coherence indicates that the noise cross-spectral density is the same as the noise auto-spectral densities, and thus any value of $\phi_{ss}(f)$ will satisfy the above equations. In practice, due to variations

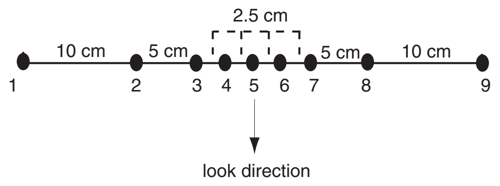

Figure 2: Microphone array geometry

low freq	high freq	microphones
0 Hz	1 kHz	1,5,9
1 kHz	2 kHz	1,2,5,8,9
2 kHz	4 kHz	2,3,5,7,8
4 kHz	8 kHz	3,4,5,6,7

Table 1: Beamformer sub-arrays

from the assumed unity coherence, the above solution may in fact result in large negative values of $\hat{\phi}_{ss}(f)$. For a diffuse noise field, this is only problematic for a few low frequency coefficients of the FFT, and we thus choose to resolve the problem in our experiments by simply applying a minimum threshold of zero to the solution of $\hat{\phi}_{ss}(f)$.

As was the case for the Zelinski post-filter, the estimate may be improved by averaging the solution over all different sensor combinations

$$\hat{h}_2(f) = \frac{\frac{2}{N(N-1)}\Re\{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\phi}_{ss}(f)\}}{\frac{1}{N}\sum_{i=1}^{N}\hat{\phi}_{x_ix_i}(f)} \tag{19}$$

While the above estimation of ϕ_{ss} seems involved, as the solution reduces to an analytical expression involving known quantities, it can be coded directly. Thus thus the added computational expense required (compared to the Zelinski post-filter) is negligible.

4 Experiments and Results

4.1 Configuration

Figure 2 shows the array geometry used in the experiments. It consists of 9 microphones arranged in a nested sub-array structure. Different inter-element spacings are used in four different frequency bands in a standard sub-array beamformer, as shown in Table 1. Within each sub-array, the channel filters were calculating using standard superdirectivity, as detailed in [4].

The experiments were conducted in an office room containing a number of computer workstations. The room has a measured reverberation time of approximately RT60 $\approx 250ms$. Multi-channel recordings of the room noise were made, consisting mainly of computer noise, air-conditioning noise, and a variable level of background speech. In addition, the impulse responses of the acoustic path between the desired speaker and each microphone were calculated from real recordings. These impulse responses were used to generate the multi-channel desired speech input for the test database. The desired speaker was located directly in front of the centre microphone, at a distance of 70 cm.

To verify that the diffuse noise assumption is valid for this noise recording, we compared the actual coherence function with the theoretical sinc function. While we found that significant differences exist in the instantaneous values, in general the actual coherence follows the trend of the theoretical values quite closely, showing that indeed a diffuse noise assumption is much more appropriate than an

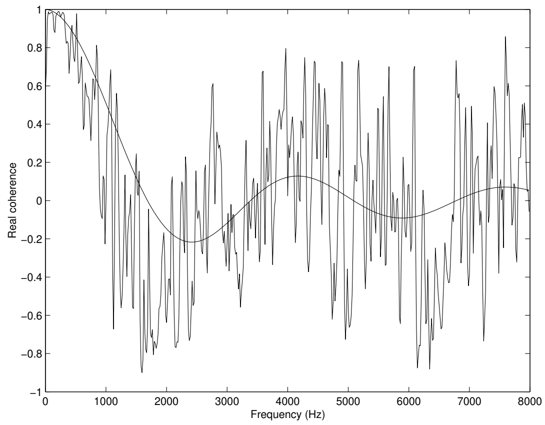

Figure 3: Example of actual and theoretical coherence functions ($d_{ij} = 0.2m$)

signal	**SNRE (dB)** 10 dB	5 dB	0 dB
beamformer output	0.2	0.3	0.5
Zelinski post-filter	3.6	3.0	2.0
proposed post-filter	12.6	11.8	9.5

Table 2: SNRE results

incoherent noise assumption in this situation, particularly for low frequencies. To illustrate this, the real part of the actual and theoretical coherence functions are compared for a frame in Figure 3 ($d_{ij} = 0.2m$).

4.2 Speech Enhancement Experiments

A first set of experiments was conducted in which the multi-channel speech signal was corrupted with the multi-channel noise recordings at average segmental SNR levels of 0, 5 and 10 dB. The noisy input, beamformer output, Zelinski post-filter output and the output from the new proposed post-filter were each assessed for the task of speech enhancement across 10 different input speech files, taken from the male adult portion of the TIDIGITS database. Each speech signal was corrupted with randomly selected portions of the noise recordings. Figure 4 plots these signals as well as the clean input signal for an utterance corresponding to the digit string '388'.

To obtain objective measures of the speech enhancement performance of the technique, we calculated the SNR enhancement (SNRE) and the log area ratio (LAR) distortion across the 10 file database. The SNRE is defined as the difference in segmental SNR (actually signal-plus-noise to noise ratio) between the noisy input and the enhanced output. The SNRE for the beamformer and both post-filters is shown in Table 2.

While the SNRE is a good measure of the reduction of the noise level, the LAR distortion is an objective speech quality measure that has been shown to be more highly correlated with speech quality as assessed subjectively by humans [6]. The LAR distortion for the noisy input, and each of the processed outputs is shown in Table 3.

These results clearly show that the proposed estimation of the signal power spectral density results in better speech enhancement performance than the standard Zelinski post-filter when the noise field is approximately diffuse. The signal plots of Figure 4, as well as the SNRE results, indicate that the proposed technique succeeds in achieving significant additional noise reduction compared to the

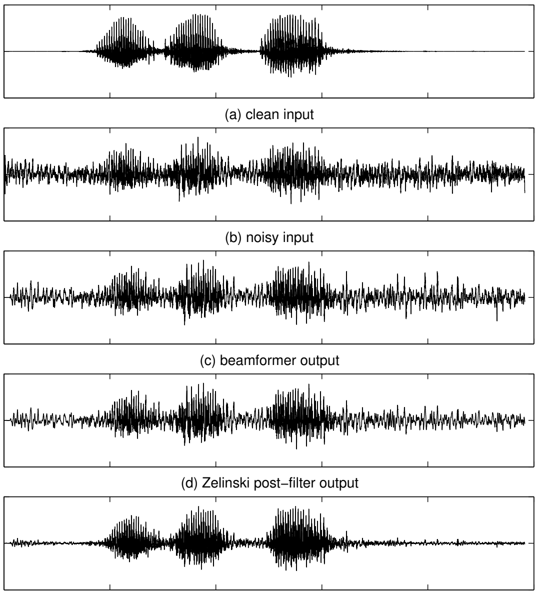

Figure 4: Signal plots

	LAR		
signal	10 dB	5 dB	0 dB
noisy input	2.7	3.4	3.9
beamformer output	2.3	3.0	3.7
Zelinski post-filter	2.4	3.2	4.0
proposed post-filter	1.8	2.3	2.8

Table 3: LAR results

standard post-filter. In addition, the technique gives lower LAR distortion than the other techniques, indicating that the additional noise reduction does not come at the expense of any significant distortion to the desired speech signal.

4.3 Speech Recognition Experiments

Speech recognition experiments were conducted on the male adult portion of the TIDIGITS database using the same experimental configuration. HMM's were trained on the clean TIDIGITS training data set, with no adaptation performed for the noise conditions or for the effect of the distant microphone. Standard MFCC parameters with energy, delta and acceleration parameters were used. The various signals were tested using the clean HMM's, at different input SNR's. The speech recognition results are given in Table 4 in terms of the percentage word error rate.

These results underline the success of the new technique in enhancing the speech signal in the presence of a high level of diffuse noise. While the Zelinski post-filter provides significant improvement in speech recognition performance, particularly at lower SNR's, the new post-filter closely matches the performance of the clean speech signal for an SNR of 10 dB, and the performance degrades gracefully as the noise is increased. Naturally, better speech recognition performance could be attained by adapting the recognition models to the noise conditions, to the distant microphone, and also to the enhanced output, and this would be done in any practical system. However, the purpose of these experiments

	WER (%)			
signal	clean	10 dB	5 dB	0 dB
noisy input	3.7	18.4	52.2	98.5
beamformer output	4.0	12.6	38.7	97.9
Zelinski post-filter	3.5	8.5	26.2	64.9
proposed post-filter	3.5	3.8	8.9	23.8

Table 4: Speech recognition results

is to simply show the degree of noise robustness that can be achieved by the enhancement technique in isolation. In this respect, the results clearly demonstrate that the proposed post-filter is successful in approaching both the quality and intelligibility of the clean speech signal in reasonable levels of diffuse noise.

5 Conclusions

In this article, we have presented a microphone array post-filter formulated specifically for a diffuse noise field. The technique builds upon the existing Zelinski array post-filter by replacing the assumption of incoherent noise with a diffuse noise assumption. We show how knowledge of the complex coherence function of the noise field can be used to solve a set of equations to obtain a more accurate estimate of the signal power spectral density, which is then used in a Wiener filter transfer function. Using real multi-channel noise recordings from an office environment, the proposed technique has been shown to give significant improvement over the existing post-filter in terms of signal to noise ratio, log area ratio distortion and word recognition rate. While this paper has focussed on a diffuse noise field, the technique may equally be applied to any noise field where the complex coherence function can be modeled.

References

[1] K. Uwe Simmer, J. Bitzer, and C. Marro. Post-filtering techniques. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 3, pages 36–60. Springer, 2001.

[2] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proceedings of ICASSP-88*, volume 5, pages 2578 –2581, 1988.

[3] Claude Marro, Yannick Mahieux, and K. Uwe Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Transactions on Speech and Audio Processing*, 6(3):240–259, May 1998.

[4] M. Doerbecker. Speech enhancement using small microphone arrays with optimized directivity. In *Proc. Int. Workshop on Acoustic Echo and Noise Control*, pages 100–103, September 1997.

[5] J. B. Allen, D. A. Berkley, and J. Blauert. Multimicrophone signal-processing technique to remove room reberberation from speech signals. *Journal of the Acoustical Society of America*, 62(4):912–915, October 1977.

[6] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements. *Objective Measures of Speech Quality*. Prentice-Hall, NJ, 1988.