IDIAP RESEARCH REPORT

Microphone Array Post-filter based on Noise Field Coherence

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

IDIAP–RR 01-40

December 2002

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 01-40

Microphone Array Post-filter based on Noise Field Coherence

Iain A. McCowan Hervé Bourlard

December 2002

Abstract. This article introduces 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 generalisation 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 the noise on different sensors. This assumption is inaccurate, particularly at low frequencies and for arrays with closely spaced sensors, and thus the corresponding post-filter is sub-optimal in realistic noise conditions. In this article, a more general expression of the post-filter estimation is developed based on an assumed knowledge of the complex coherence of the noise field. This general expression can be used to construct a more appropriate post-filter in a variety of different noise fields. In experiments using real noise recordings from a computer office, the modified post-filter results in significant improvement in terms of objective speech quality measures and speech recognition performance using a diffuse noise model.

1 Introduction

Much research has been done 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. Simmer *et al* [1] show that the optimal broadband multi-channel Minimum Mean Square Error (MMSE) enhancement solution can be decomposed into a Minimum Variance Distortionless Response (MVDR) beamformer followed by a single-channel Wiener filter, and so such a post-filter can result in significant improvement in the broadband Signal to Noise Ratio (SNR) over an MVDR beamformer used in isolation.

While the theory promises improved performance with such a Wiener post-filter, in practice it can be difficult to obtain reliable estimates of the signal and noise spectral densities required to formulate the post-filter transfer function. The most common method to estimate these spectral densities is based upon the use of the auto- and cross-spectral densities of the multi-channel input signals. Such a post-filter estimation is examined in detail by Marro *et al* [2], and is principally based on the work of Zelinski [3]. In this article, this technique is referred to as the *Zelinski post-filter*.

While the Zelinski post-filter shows reasonable performance, its formulation is based upon a number of assumptions. In particular, an 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.

This article demonstrates how the assumption of incoherent noise in the Zelinski post-filter estimator can be replaced by the more general assumption of a known noise field coherence function. A number of practical noise fields, such as those encountered in offices or cars, can be modeled by theoretical cases, such as spherically isotropic (diffuse) or cylindrically isotropic noise fields. The coherence functions for these theoretical noise fields already form the basis of several established beamforming techniques, in particular superdirective beamformers [4, 5, 6]. Here the use of the theoretical noise coherence models is extended to the post-filter estimation, allowing a more appropriate post-filter to be developed for different noise conditions. The proposed generalised post-filter estimation includes the Zelinski post-filter as a particular case, corresponding to a unity coherence matrix.

This paper is organised as follows. Section 2 gives a brief review of the theoretical basis of microphone array post-filtering and the Zelinski post-filter estimator on which the proposed technique is based. Section 3 reviews the use of the complex coherence function to characterise different noise fields, and uses this to derive the proposed generalised post-filter. A simple analysis is also given to highlight the advantage of the proposed technique over the conventional Zelinski post-filter. In Section 4, the performance of the new post-filter is assessed in speech enhancement and recognition experiments, using multi-channel office noise recordings. In these experiments, the new post-filter is shown to give significant performance improvement over the existing technique in terms of objective speech quality measures and speech recognition performance.

2 Microphone Array Post-filtering

2.1 Theoretical Framework

Common practice in microphone array processing is to model the received multi-channel input as the desired signal filtered by the acoustic path to each microphone, plus an additive noise component on each channel. That is (omitting the frequency dependence for clarity),

$$\mathbf{x}' = s\mathbf{d} + \mathbf{n}' \tag{1}$$

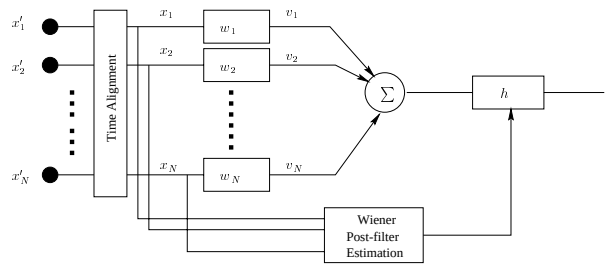

Figure 1: Filter-sum beamformer with post-filter

where s is the desired signal, $\mathbf{d}$ is the propagation vector of the signal source

$$\mathbf{d} = \begin{bmatrix} d_1 & d_2 & \cdots & d_N \end{bmatrix}^T \tag{2}$$

and $\mathbf{n}'$ is similarly the vector of additive noise signals

$$\mathbf{n}' = \begin{bmatrix} n'_1 & n'_2 & \cdots & n'_N \end{bmatrix}^T \tag{3}$$

where N is the number of microphones in the array.

Using this model, Simmer *et al* [1] demonstrate how the optimal broadband Minimum Mean Square Error (MMSE) filter solution (that is, the multi-channel Wiener filter) can be expressed as a single-channel Wiener filter operating on the output of a classical Minimum Variance Distortionless Response (MVDR) beamformer, that is,

$$\mathbf{w}_{opt} = \left[\frac{\phi_{ss}}{\phi_{ss} + \phi_{nn}}\right] \frac{\mathbf{\Phi}_{nn}^{-1}\mathbf{d}}{\mathbf{d}^H \mathbf{\Phi}_{nn}^{-1}\mathbf{d}} \tag{4}$$

where $\mathbf{w}_{opt}$ is the vector of optimal filter coefficients, ϕ_{ss} and ϕ_{nn} are respectively the (single-channel) signal and noise auto-spectral density vectors, and $\mathbf{\Phi}_{nn}$ is the (multi-channel) noise cross-spectral matrix. The bracketed factor in the above expression corresponds to a single-channel Wiener filter, while the remaining factor forms the well known solution for the filters of a MVDR beamformer [5]. The above equation suggests an optimal array processing structure like that shown in Figure 1, in which the transfer function of the single-channel Wiener post-filter is typically estimated from the aligned multi-channel input. The beamformer first maximises the directivity of the array response, and then the post-filter further enhances the output broadband Signal to Noise Ratio (SNR). This article focuses on the problem of estimating the post-filter term in the above equation, that is

$$h = \frac{\phi_{ss}}{\phi_{ss} + \phi_{nn}} \tag{5}$$

2.2 Zelinski Post-filter

This section presents a brief review of the Zelinski post-filter on which the proposed technique is based [3]. This post-filter was further developed by Marro *et al* [2], who thoroughly investigated the interaction of the beamformer and post-filter in the structure of Figure 1.

It is assumed that at the output of the time alignment module, the inputs have been scaled and aligned to account for the effect of the propagation vector $\mathbf{d}$. The signals at the output of the time alignment can thus be modeled as

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

where $\mathbf{n}$ is the noise signal vector 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

$$\phi_{x_i x_i} = \phi_{ss} + \phi_{n_i n_i} + 2\Re\{\phi_{s n_i}\} \tag{7}$$

and

$$\phi_{x_i x_j} = \phi_{ss} + \phi_{n_i n_j} + \phi_{s n_j} + \phi_{n_i s} \tag{8}$$

Under the assumptions that:

1. the signal and noise are uncorrelated ($\phi_{n_i s} = 0, \forall i$),
2. the noise power spectrum is the same on all sensors ($\phi_{n_i n_i} = \phi_{nn}, \forall i$), and
3. the noise is uncorrelated between sensors ($\phi_{n_i n_j} = 0, \forall i \neq j$),

these reduce to

$$\phi_{x_i x_i} = \phi_{ss} + \phi_{nn} \tag{9}$$

$$\phi_{x_i x_j} = \phi_{ss} \tag{10}$$

The auto- and cross-spectral densities of the time-aligned inputs, $\phi_{x_i x_i}$ and $\phi_{x_i x_j}$, can be estimated using a standard recursive update formula [7]

$$\hat{\phi}_{x_i x_j} = \alpha \hat{\phi}'_{x_i x_j} + (1-\alpha) x_i x_j^* \tag{11}$$

where $\hat{\phi}'_{x_i x_j}$ and $\hat{\phi}_{x_i x_j}$ are the spectral estimates for the previous and current frames respectively, and $(\cdot)^*$ is the complex conjugate operator. The term α is a number close to unity, and is given by $\alpha = \exp(-D/\tau f_s)$, where D is the filter-bank decimation factor, f_s is the sampling frequency, and τ is the decay time constant.

From the above equations, it is evident that the numerator and denominator of the Wiener filter transfer function in Equation 5 can be estimated from the cross- and auto-spectral densities of the input channels, respectively. This estimate can be made more robust by averaging the spectral densities over all possible sensor combinations, resulting in the post-filter estimator

$$\hat{h}_z = \frac{\frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Re\{\hat{\phi}_{x_i x_j}\}}{\frac{1}{N} \sum_{i=1}^{N} \hat{\phi}_{x_i x_i}} \tag{12}$$

The real operator $\Re\{\cdot\}$ is used because the term being estimated in the numerator, ϕ_{ss}, is necessarily real. It should be noted that the denominator in fact provides an over-estimate of the noise power at the beamformer output, as it is calculated using the input signals. An alternative is to use the beamformer output power in the denominator, however the above post-filter has been found to give better noise reduction in practice [2].

3 Generalised Post-filter based on Noise Field Coherence

This section describes the complex coherence function and its application in modelling different noise fields. The coherence function is then used in the formulation of the proposed generalised post-filter transfer function, and an analysis highlighting the advantage of the new post-filter is presented.

3.1 Complex Coherence Function of Noise Fields

A common measure used to characterise noise fields is the *complex coherence function*. The coherence between two signals at points i and j is defined as

$$\Gamma_{ij} = \frac{\phi_{ij}}{\sqrt{\phi_{ii}\phi_{jj}}} \tag{13}$$

where ϕ_{ij} is the cross-spectral density between the signals at i and j. The coherence has the range $|\Gamma_{ij}| \leq 1$, and is essentially a normalised measure of the correlation that exists between the signals at two discrete points in a noise field. The complex coherence is a powerful tool in array processing, both in the development of algorithms (such as superdirective beamforming), and for analysing the array performance in theoretical noise fields.

A spherically isotropic or diffuse noise field is a good model for a number of practical reverberant noise environments encountered in speech enhancement applications, such as offices and cars [8, 9, 10]. A diffuse noise field is characterised by uncorrelated noise signals of equal power propagating in all directions simultaneously. It can be shown that the coherence of a diffuse noise field is real-valued and is given by [11]

$$\Gamma_{ij} = \text{sinc}\left(\frac{2\pi f d_{ij}}{c}\right) \tag{14}$$

A diffuse noise field assumption is often used in array processing, primarily with superdirective beamformers, whose channel filters are calculated to optimise the array gain in such conditions. Such a diffuse noise field model will be assumed in the experiments in this paper. For a detailed discussion of the coherence functions of different noise fields, the reader is referred to [12].

3.2 Derivation of Proposed Post-filter

The Zelinski post-filter formulation described in Section 2.2 makes the assumption that the noise between sensors is uncorrelated, corresponding to a perfectly incoherent noise field ($\mathbf{\Gamma}_{nn} = \mathbf{I}$). Such a noise field will seldom occur in practice, although it can be a reasonable approximation if the spacing between sensors is sufficiently large. Electrical noise on microphones in an array is also typically random, and can be characterised as a source of incoherent noise.

While the Zelinski post-filter approximation has been shown to give reasonable performance in a variety of conditions [2, 13], the performance would be improved if a more accurate model of the noise field were used. In the following, the complex coherence function of the noise field is used to formulate a more appropriate estimation of the array post-filter transfer function.

With the assumption of aligned signal on all sensors, and zero correlation between the desired signal and the noise,

$$\phi_{x_i x_i} = \phi_{ss} + \phi_{n_i n_i} \tag{15}$$

$$\phi_{x_j x_j} = \phi_{ss} + \phi_{n_j n_j} \tag{16}$$

$$\phi_{x_i x_j} = \phi_{ss} + \phi_{n_i n_j} \tag{17}$$

$$\Gamma_{n_i n_j} = \frac{\phi_{n_i n_j}}{\sqrt{\phi_{n_i n_i}\phi_{n_j n_j}}} \tag{18}$$

If a model of the coherence ($\hat{\mathbf{\Gamma}}_{nn}$) is available, then the above form a set of four equations with four unknown variables, and thus can be solved for ϕ_{ss}. Such a solution was examined in [14], however

due to the square root term in Equation 18 it has the disadvantages of dual solutions and significantly higher computational expense than the Zelinski post-filter. If the assumption of the same noise power spectrum across sensors ($\phi_{n_i n_i} = \phi_{nn}, \forall i$) is also made (as is the case for isotropic noise fields), then $\Gamma_{n_i n_j} = \frac{\phi_{n_i n_j}}{\phi_{nn}}$, and so Equations 15-18 simplify to

$$\phi_{x_i x_i} = \phi_{ss} + \phi_{nn} \tag{19}$$

$$\phi_{x_j x_j} = \phi_{ss} + \phi_{nn} \tag{20}$$

$$\phi_{x_i x_j} = \phi_{ss} + \Gamma_{n_i n_j} \phi_{nn} \tag{21}$$

The signal power spectral density can thus be estimated as

$$\hat{\phi}_{ss}^{(ij)} = \frac{\Re\left\{\hat{\phi}_{x_i x_j}\right\} - \frac{1}{2}\Re\left\{\hat{\Gamma}_{n_i n_j}\right\}\left(\hat{\phi}_{x_i x_i} + \hat{\phi}_{x_j x_j}\right)}{\left(1 - \Re\left\{\hat{\Gamma}_{n_i n_j}\right\}\right)} \tag{22}$$

where the average of $\phi_{x_i x_i}$ and $\phi_{x_j x_j}$ is taken to improve robustness. The post-filter denominator $(\phi_{ss} + \phi_{nn})$ can still be estimated by $\hat{\phi}_{x_i x_i}$, as for the Zelinski technique.

As for the Zelinski technique, the estimate is improved by averaging the solution over all unique sensor combinations, resulting in the post-filter

$$\hat{h}_p = \frac{\frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\phi}_{ss}^{(ij)}}{\frac{1}{N} \sum_{i=1}^{N} \hat{\phi}_{x_i x_i}} \tag{23}$$

It is evident that the Zelinski post-filter is a particular instance of the above general expression, in which $\hat{\mathbf{\Gamma}}_{nn} = \mathbf{I}$, leading to the estimate $\hat{\phi}_{ss}^{(ij)} = \hat{\phi}_{x_i x_j}$. In terms of computational expense, the proposed procedure requires an additional 4 vector operations for each unique (i,j) pair compared to the Zelinski technique.

A problem arises if $\hat{\Gamma}_{n_i n_j} = 1$, $i \neq j$, as this leads to an indeterminite solution. In practice, this can be resoved by applying a maximum threshold on the coherence model, limiting the assumed coherence function to $\left|\hat{\Gamma}_{ij}\right| \leq \xi$, $\forall i \neq j$ where $\xi < 1$.

3.3 Analysis of Proposed Post-filter

This section presents a simple analysis comparing the error terms of the proposed post-filter and the Zelinski post-filter. Under the assumptions of uncorrelated signal and noise, equal noise power on sensors, and given the *actual* noise field coherence $\Gamma_{n_i n_j}$, the auto- and cross-spectral densities of the inputs can be expressed as

$$\phi_{x_i x_i} = \phi_{ss} + \phi_{nn} \tag{24}$$

$$\phi_{x_j x_j} = \phi_{ss} + \phi_{nn} \tag{25}$$

$$\phi_{x_i x_j} = \phi_{ss} + \Gamma_{n_i n_j} \phi_{nn} \tag{26}$$

Substituting into the Zelinski post-filter transfer function (Equation 12) leads to the expression

$$\hat{h}_z = \frac{\phi_{ss}}{\phi_{ss} + \phi_{nn}} + \underbrace{\frac{\phi_{nn}}{\phi_{ss} + \phi_{nn}} \left[\frac{2}{N(N-1)} \Re\left\{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Gamma_{n_i n_j}\right\}\right]}_{\epsilon_z} \tag{27}$$

From this expression, it can be seen that the error in the post-filter estimator, ϵ_z, is determined by the average of the non-diagonal elements in the actual noise coherence matrix. This is, in effect, the average error between the actual noise coherence matrix and a unity coherence matrix. Thus, ϵ_z becomes zero in an incoherent noise field.

Similarly, the signal spectral density estimate for the proposed technique (Equation 22) becomes

$$\begin{aligned}\hat{\phi}_{ss} &= \frac{\phi_{ss} + \Gamma_{n_i n_j}\phi_{nn} - \hat{\Gamma}_{n_i n_j}\phi_{ss} - \hat{\Gamma}_{n_i n_j}\phi_{nn}}{1 - \hat{\Gamma}_{n_i n_j}} \\ &= \phi_{ss} + \phi_{nn}\frac{\Gamma_{n_i n_j} - \hat{\Gamma}_{n_i n_j}}{1 - \hat{\Gamma}_{n_i n_j}}\end{aligned} \tag{28}$$

Substituting this expression into the post-filter estimator of Equation 23 leads to

$$\begin{aligned}\hat{h}_p = &\frac{\phi_{ss}}{\phi_{ss} + \phi_{nn}} + \\ &\underbrace{\frac{\phi_{nn}}{\phi_{ss} + \phi_{nn}}\left[\frac{2}{N(N-1)}\Re\left\{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{\Gamma_{n_i n_j} - \hat{\Gamma}_{n_i n_j}}{1 - \hat{\Gamma}_{n_i n_j}}\right\}\right]}_{\epsilon_p}\end{aligned} \tag{29}$$

In this case, the error term ϵ_p is determined by the averaged difference between the actual and assumed noise coherence matrices, and reduces to zero when these are equivalent. This error term simplifies to the Zelinski post-filter error if an incoherent noise field is assumed. This analysis demonstrates the advantage of the proposed technique : it provides a generalised expression that can be optimised for an arbitrary noise field model, while the Zelinski post-filter is explicitly optimised for an incoherent noise field.

Both the Zelinski and the proposed post-filter will also contain an error term due to any correlation between the desired signal and the noise on each channel (assumed zero in the formulations). It can be shown that the error term is the same for both post-filters, and is given as

$$\epsilon_{sn} = \frac{\frac{2}{N}\sum_{i=1}^{N}\Re\{\phi_{sn_i}\}}{\phi_{ss} + \phi_{nn} + \frac{2}{N}\sum_{i=1}^{N}\Re\{\phi_{sn_i}\}} \tag{30}$$

Assuming that $\phi_{ss} + \phi_{nn} \gg \frac{2}{N}\sum_{i=1}^{N}\Re\{\phi_{sn_i}\}$ the effect of this error term on the post-filter will be negligible.

4 Experiments and Results

This section details speech enhancement and recognition experiments performed in an office of computer workstations. The performance of the proposed technique was investigated using the assumption of a diffuse noise field. Thus the expression

$$\hat{\Gamma}_{n_i n_j} = \text{sinc}\left(\frac{2\pi f d_{ij}}{c}\right) \tag{31}$$

was used in the solution of Equation 22.

Figure 2 shows the geometry of the microphone array used in the experiments. It consists of 5 microphones, equi-spaced in a linear broadside configuration with inter-element spacing of 5 cm. The desired speaker was located directly in front of the centre microphone, at a distance of 70 cm. The beamforming filters were calculated using the fixed superdirective technique, detailed in [5], which gives

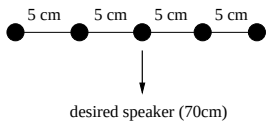

Figure 2: Microphone array geometry

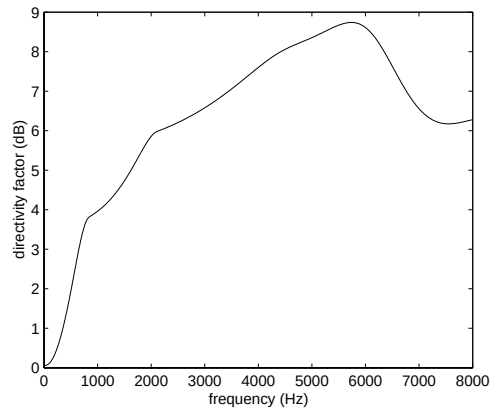

Figure 3: Beamformer directivity factor

the MVDR solution under the assumption of a diffuse noise field by using $\Phi_{nn} = \mathbf{\Gamma}_{nn}$ in Equation 4. The directivity factor for this beamformer is given by

$$DF = \frac{\left|\mathbf{w}_{opt}^H \mathbf{d}\right|^2}{\mathbf{w}_{opt}^H \mathbf{\Gamma}_{nn} \mathbf{w}_{opt}} \tag{32}$$

and is shown in Figure 3.

The office room had a measured reverberation time of approximately RT60 $\approx 420ms$. 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 using a maximum-length sequence (MLS) technique [15]. These impulse responses were used to generate the multi-channel desired speech input for the test database.

To verify that the diffuse noise assumption is valid for this office noise recording, the measured coherence function was compared with the theoretical sinc function. While significant differences exist in the instantaneous values, in general the actual coherence follows the trend of the theoretical values quite closely, showing that a diffuse noise assumption is much more appropriate than an 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 representative frame in Figure 4 ($d_{ij} = 0.1m$).

All experiments were conducted using utterances from the male adult portion of the TIDIGITS database. The TIDIGITS database consists of connected strings of the ten digits (0-9) and the word 'oh'. The male adult portion has 111 speakers, of which 55 are used for training and 56 for testing. This test set comprises 4311 digit strings.

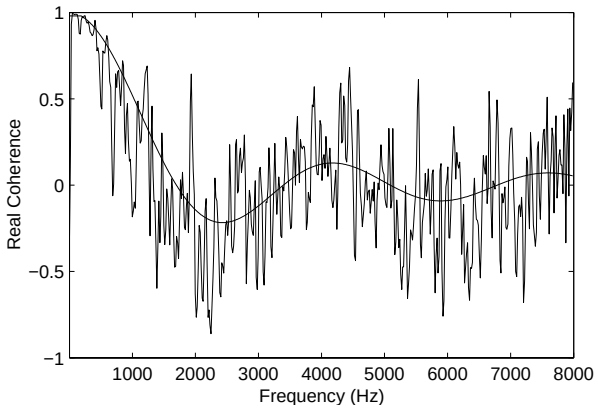

Figure 4: Example of actual and theoretical coherence functions (office noise, $d_{ij} = 0.1m$)

Table 1: SNRE results

	SNRE (dB)		
signal	10 dB	5 dB	0 dB
beamformer output	0.7	0.7	0.7
Zelinski post-filter	0.9	1.0	1.0
proposed post-filter	4.3	6.4	6.9

4.1 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 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 the entire test set. Each speech signal was corrupted with randomly selected portions of the noise recordings. Figure 5 plots these outputs as well as the clean input signal for an utterance corresponding to the digit string '111' (input SNR level of 5 dB). Similarly, spectrograms for the same outputs are shown in Figure 6.

In order to assess the speech enhancement, in addition to the noisy signal, the separate multi-channel speech and noise signals were also both processed for each technique. For the post-filtering techniques, the post-filter was calculated on the noisy signal and then also applied to both the clean speech and pure noise signals. In this way, three signals were available at the output - the enhanced noisy signal, along with the processed clean speech and noise signals.

To assess the noise reduction, the SNR Enhancement (SNRE) was used. The SNRE is defined as the difference in segmental SNR between the noisy input and the enhanced output, and was calculated using the separate clean speech and noise signals at both the input and output, giving the true SNR. In addition, to assess the distortion to the desired signal, the Log Area Ratio (LAR) was used. 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 [16]. The LAR was calculated between the original clean speech input and the processed output of the clean speech signal.

The SNRE and the LAR results were averaged across the entire TIDIGITS male adult test set, and are shown in Tables 1 and 2 respectively. The low SNRE of the beamformer is due to the fact that the majority of the noise energy is in the region below 1000 Hz, where the beamformer offers little directivity (Figure 3). The SNRE of the Zelinski post-filter shows only a small improvement for this

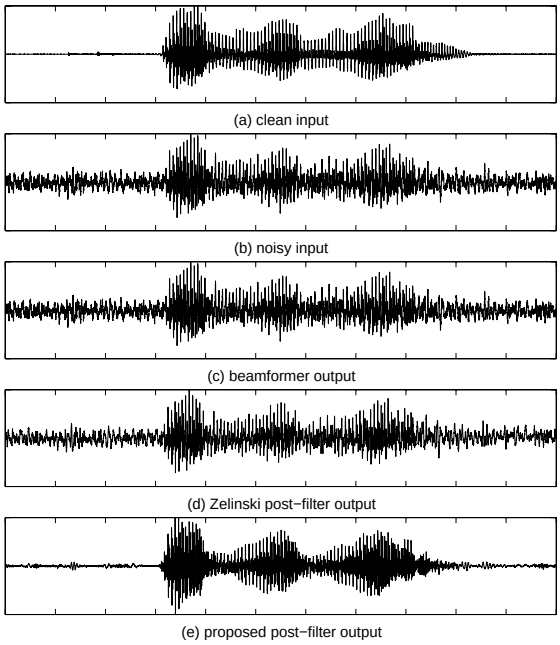

Figure 5: Signal plots (5 dB SNR)

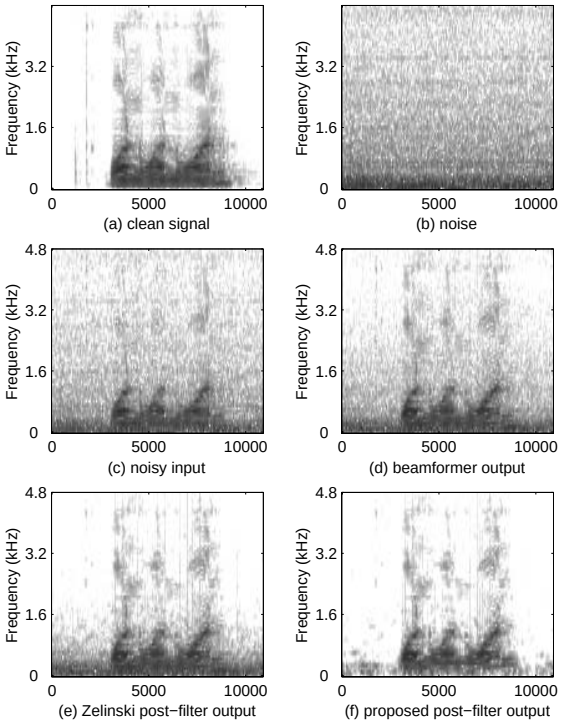

Figure 6: Signal spectrograms (5 dB SNR)

Table 2: LAR results

	LAR			
signal	clean	10 dB	5 dB	0 dB
input (centre mic)	3.7	3.7	3.7	3.7
beamformer output	3.6	3.6	3.6	3.6
Zelinski post-filter	3.9	4.7	5.1	5.6
proposed post-filter	4.0	4.7	4.9	5.3

Table 3: Speech recognition results

	WER (%)			
signal	clean	10 dB	5 dB	0 dB
noisy input	2.3	12.9	43.9	85.8
beamformer output	1.6	5.9	15.7	40.6
Zelinski post-filter	1.9	7.6	15.7	37.4
proposed post-filter	3.3	5.2	9.5	20.7

scenario as the error term in the post-filter ϵ_z (Equation 27) means that h_z tends to one for high values of the coherence, thus little additional noise reduction will occur at low frequencies. The spectrograms of the beamformer (Figure 6(d)) and the Zelinski post-filter (Figure 6(e)) illustrate this behaviour : the Zelinski post-filter improves the noise reduction for mid to high frequencies, but has no effect at low frequencies. Conversely, due to the use of an appropriate coherence model in the formulation, the proposed post-filter is able to provide significant broadband noise reduction (3.6 - 6.2 dB) over the beamformer output (Figure 6(f)).

The LAR values between the original desired signal and the processed outputs (desired signal only) show some interesting trends. First, it is seen that that the beamformer in isolation does not further distort to the desired signal compared to the input, due to the unity constraint in the filter optimisation. As expected, both post-filtering techniques introduce some distortion to the desired signal, with more distortion as the noise level increases. However, it is noted that the proposed post-filter gives greater noise reduction than the Zelinski post-filter, without introducing additional signal distortion.

These results clearly show that incorporating a more accurate noise field coherence model (diffuse noise in this case) into the post-filter estimation process results in improved speech enhancement. This is particularly observed at low frequencies, which are generally problematic for traditional array processing techniques. While a better array geometry (with e.g. more microphones, greater aperture, frequency dependent spacings) would result in higher directivity of the beamformer, it is evident that by using an appropriate post-filter even small, simple array designs can yield acceptable speech enhancement.

4.2 Speech Recognition Experiments

Speech recognition experiments were conducted on the same test files (TIDIGITS adult male test set) used in the enhancement experiments described above. Triphone Hidden Markov Models (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 3, and shown graphically in Figure 7, in terms of the percentage word error rate.

These results underline the success of the new technique in enhancing the speech signal in the

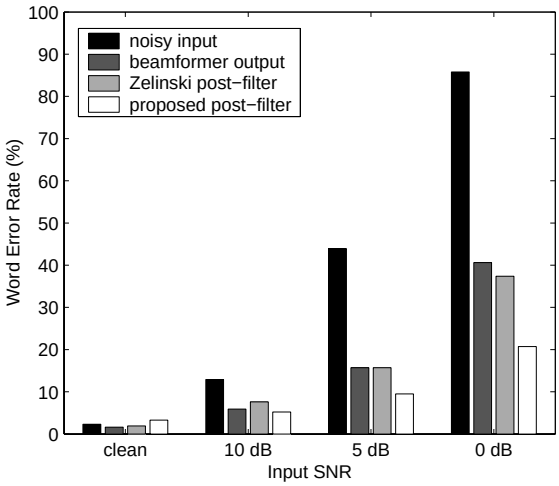

Figure 7: Speech recognition results

presence of a high level of diffuse noise. While the beamformer gives improved speech recognition performance compared to the noisy input, the standard Zelinski post-filter is not effective in further improving results. In contrast, the proposed post-filter gives significantly better recognition performance, particularly as the noise level increases. As expected, in clean and low noise conditions the beamformer in isolation works best due to the lack of signal distortion. Naturally, better speech recognition performance could be attained by adapting the recognition models to the noise conditions, the distant microphone, and also the enhanced output, and this would be done in any practical system. However, the purpose of these experiments is to simply show the degree of noise robustness that can be achieved by the enhancement technique in isolation. In this respect, these recognition results, coupled with those above for speech enhancement, clearly demonstrate that the proposed post-filter is successful in approaching the quality of the clean speech signal in reasonable levels of diffuse noise.

5 Conclusions

This article has presented a microphone array post-filter formulated to handle a variety of noise fields. The technique builds upon the existing Zelinski array post-filter by replacing the assumption of incoherent noise with the assumption of a known noise field coherence function. Knowledge of the noise field coherence function is 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 a reverberant 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 the experiments in this paper have focussed on a diffuse noise field, the technique may equally be applied to any noise field where the complex coherence function can be adequately modeled.

Acknowledgements

The author would like to thank Darren Moore and Claude Marro for their comments and suggestions at various stages of this work.

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] 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.

[3] 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.

[4] J. Bitzer and K. Uwe Simmer. Superdirective microphone arrays. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 2, pages 19–38. Springer, 2001.

[5] H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(10):1365–1376, October 1987.

[6] H. Cox, R. Zeskind, and T. Kooij. Practical supergain. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34(3):393–397, June 1986.

[7] 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.

[8] J. Bitzer, K. Uwe Simmer, and K. Kammeyer. Theoretical noise reduction limits of the generalized sidelobe canceller (gsc) for speech enhancement. In *Proceedings of ICASSP 99*, volume 5, pages 2965–2968, 1999.

[9] J. Meyer and K. Uwe Simmer. Multi-channel speech enhancment in a car environment using wiener filtering and spectral subtraction. In *Proceedings of ICASSP 97*, volume 2, pages 1167–1170, 1997.

[10] G. W. Elko. Superdirectional microphone arrays. In S.L Gay and J. Benesty, editors, *Acoustic Signal Processing for Telecommunication*, chapter 10, pages 181–237. Kluwer Academic Publishers, 2000.

[11] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson Jr. Measurement of correlation coefficients in reverberant sound fields. *Journal of the Acoustic Society of America*, 27:1072–1077, 1955.

[12] G. Elko. Spatial coherence functions for differential microphones in isotropic noise fields. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 4, pages 61–85. Springer, 2001.

[13] I. McCowan, C. Marro, and L. Mauuary. Robust speech recognition using near-field superdirective beamforming with post-filtering. In *Proceedings of ICASSP 2000*, volume 3, pages 1723–1726, 2000.

[14] I. McCowan and H. Bourlard. Microphone array post-filter for diffuse noise field. In *Proceedings of ICASSP-02*, volume 1, pages 905–908, 2002.

[15] D. Rife and J. Vanderkooy. Transfer-function measurement with maximum-length sequences. *Journal of the Audio Engineering Society*, 37:419–444, June 1989.

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