OPTIMUM POST-FILTER ESTIMATION FOR NOISE REDUCTION IN MULTICHANNEL SPEECH PROCESSING

Stamatis Leukimmiatis and Petros Maragos

National Technical University of Athens, School of ECE, Zografou, Athens 15773, Greece.
Email:[sleukim, maragos]@cs.ntua.gr

ABSTRACT

This paper proposes a post-filtering estimation scheme for multichannel noise reduction. The proposed method is an extension and improvement of the existing Zelinski and McCowan post-filters which use the auto- and cross-spectral densities of the multichannel input signals to estimate the transfer function of the Wiener post-filter. A drawback in previous two post-filters is that the noise power spectrum at the beamformer's output is over-estimated and therefore the derived filters are sub-optimal in the Wiener sense. The proposed method overcomes this problem and can be used for the construction of an optimal post-filter which is also appropriate for a variety of different noise fields. In experiments with real noise multichannel recordings the proposed technique has shown to obtain a significant gain over the other studied methods in terms of signal-to-noise ratio, log area ratio distance and speech degradation measure. In particular the proposed post-filter presents a relative SNR enhancement of 17.3% and a relative decrease on signal degradation of 21.7% compared to the best of all the other studied methods.

1. INTRODUCTION

Nowdays the use of microphone arrays for speech enhancement seems very promising, with the main advantage being that a microphone array can simultaneously exploit the spatial diversity of speech and noise, so that both spectral and spatial characteristics of signals can be used [1]. In most cases the speech and noise sources are in different spatial locations, thus a multichannel system compared to a single channel system obtains a significant gain due to the ability of suppressing interfering signals and noise originating from undesired directions.

The spatial discrimination of the array is exploited by beamforming algorithms [1]. In many cases though the obtainable noise reduction is not sufficient and post-filtering techniques are applied to further enhance the output of the beamformer. The Minimum Mean Square Error (MMSE) estimation of a multichannel signal from its noisy observations is achieved using the multichannel Wiener filter. Simmer et al. [2] have shown that the optimal broadband multichannel MMSE filter can be factorized into a Minimum Variance Distortionless Response (MVDR) beamformer [3] followed by a single channel Wiener post-filter. In general, such a post-filter accomplishes higher noise reduction than the MVDR beamformer alone, therefore its integration in the beamformer output can lead to substantial SNR gain.

Despite its theoretically optimal results, Wiener post-filter can be difficult to realize in practice. This is due to the requirement for knowledge of second order statistics for both the signal and the corrupting noise that makes the Wiener filter signal-dependent. A variety of post-filtering techniques trying to address this issue have been proposed in the literature [4, 5, 6, 7]. A quite common method for the formulation of the post-filter transfer function is based on the use of the auto- and cross-spectral densities of the multichannel input signals [2, 4, 6].

One of the early methods for post-filter estimation is due to Zelinski [4] which was further studied by Marro et al. [8]. The generalized version of Zelinski's algorithm is based on the assumption of a spatially uncorrelated noise field. However this assumption is not realistic for most practical applications. If a more accurate noise field model was used instead, the overall performance of the noise reduction system would be improved . McCowan et al. [6] replaced this assumption by the most general assumption of a known noise field coherence function and extended the previous method to develop a more general post-filtering scheme. In [6] it is proved that Zelinski's post-filter is a special case of McCowan's post-filter for the case of spatially uncorrelated noise. However a drawback in both methods is that the noise power spectrum at the beamformer's output is over-estimated [6, 9] and therefore the derived filters are sub-optimal in the Wiener sense.

This paper deals with the problem of estimating the Wiener post-filter transfer function so that the estimated filter will be optimal in terms of MMSE, while still allowing for the development of a general post-filter appropriate for a variety of different noise fields. To accomplish these demands we preserve McCowan's general assumption of a known noise field coherence function [6] but also take into account the noise reduction performed by the MVDR beamformer. In this way we estimate the speech source's spectrum same as McCowan but we propose a new robust method for the estimation of the power spectrum at the beamformer's output which is consistent with the optimality in terms of MMSE.

2. PROBLEM STATEMENT

Let us consider an M-sensor linear microphone array where a speech source is located at a distance r and at an angle θ from the center of the array. The observed signal, $y_i(n), i = 0,\ldots,M-1$, at the ith sensor is a delayed and attenuated version of the original speech signal $s(n)$ with an additive noise component $v_i(n)$. Each microphone signal $y_i(n)$ can also be considered as a linearly filtered version of the source signal plus additive noise. Applying the short-time Fourier transform (STFT), the observed information in the joint time-frequency domain can be written as

$$\mathbf{Y}(k,\ell) = \mathbf{H}(k;\theta,r)S(k,\ell) + \mathbf{V}(k,\ell), \tag{1}$$

where k and ℓ are the frequency bin and the time frame index, respectively, and

$$\mathbf{Y}(k,\ell) = [Y_0(k,\ell)\,, Y_1(k,\ell), \ldots, Y_{M-1}(k,\ell)]^T \tag{2}$$

$$\mathbf{H}(k;\theta,r) = [H_0(k;\theta,r), \ldots, H_{M-1}(k;\theta,r)]^T \tag{3}$$

$$\mathbf{V}(k,\ell) = [V_0(k,\ell)\,, V_1(k,\ell), \ldots, V_{M-1}(k,\ell)]^T\,. \tag{4}$$

The ith element of the vector $\mathbf{H}(k;\theta,r)$ corresponds to the frequency response, $H_i(k;\theta,r) = \alpha_i(\theta,r)e^{-j\omega_k\tau_i(\theta,r)}$, of the acoustic path between the speech source and the ith sensor, where $a_i(\theta,r)$ is the attenuation factor, $\tau_i(\theta,r)$ is the time delay expressed in number of samples and ω_k is the discrete-time angular frequency corresponding to the kth frequency bin.

2.1 Noise Field

In microphone array applications, noise fields can be characterized by a measure known as *complex coherence function*. Coherence

This work was supported in part by the Greek GSRT research program PENED–2003 and in part by the European research project HIWIRE.

function measures the amount of correlation between noise signals at different spatial locations and is defined as [3]:

$$\Gamma_{V_pV_q}(\omega) = \frac{\Phi_{V_pV_q}(\omega)}{\sqrt{\Phi_{V_pV_p}(\omega)\Phi_{V_qV_q}(\omega)}}, \tag{5}$$

where $\Phi_{V_pV_q}(\omega)$ is the cross-spectral density between the noise arrived at sensors p and q and $\Phi_{V_pV_p}(\omega)$, $\Phi_{V_qV_q}(\omega)$ are the spectral densities of the noise at sensors p and q, respectively.

A *diffuse noise field* is defined as equally distributed uncorrelated white noise coming from all directions and is a widely-used model for many applications concerning noisy environments (e.g cars and offices [5],[6]). The complex coherence function for such a noise field can be approximated by

$$\Gamma_{V_pV_q}(\omega) = \frac{\sin(\omega f_s d/c)}{\omega f_s d/c}, \forall \omega, \tag{6}$$

where d is the distance between sensors p and q and ω is the discrete-time angular frequency.

For the case of a spatially uncorrelated noise field, the coherence function reduces to $\Gamma_{V_pV_q}(\omega) = 1$, for $p = q$ and $\Gamma_{V_pV_q}(\omega) = 0$, for $p \neq q, \forall \omega$. Such a noise field can be generated by thermal noise in the microphones and is randomly distributed, in general.

2.2 Multichannel Wiener Filter

The optimum, in terms of MMSE, weight vector $\mathbf{W}_{\text{opt}}(k,\ell)$ that transforms the corrupted input signal vector, $\mathbf{H}(k;\theta,r)S(k,\ell)$, by additive noise $\mathbf{V}(k,\ell)$, into the best MMSE approximation of the source signal $S(k,\ell)$ is known as *multichannel Wiener filter*. To find this optimum weight vector we have to minimize the mean square error at the beamformer's output. In time-frequency domain the error at the beamformer's output is defined as $\mathscr{E}(k,\ell) = S(k,\ell) - \mathbf{W}^H(k,\ell)\mathbf{Y}(k,\ell)$ and the optimum solution $\mathbf{W}_{\text{opt}}(k,\ell)$, assuming that the matrix $\mathbf{\Phi}_{\text{YY}}(k,\ell)$ is invertible, is given by

$$\mathbf{W}_{opt}(k,\ell) = \mathbf{\Phi}_{\text{YY}}^{-1}(k,\ell)\mathbf{\Phi}_{\text{YS}}(k,\ell), \tag{7}$$

where $\mathbf{\Phi}_{\text{YS}}(k,\ell)$ is the cross-spectral density vector between the source signal and the sensors' inputs and $\mathbf{\Phi}_{\text{YY}}(k,\ell)$ is the spectral density matrix of the sensors' inputs.

Under the assumption that the source signal $S(k,\ell)$ and the noise are uncorrelated, it has been shown in [2] that (7) can be further decomposed into a MVDR beamformer followed by a single channel Wiener filter, which operates at the output of the beamformer:

$$\mathbf{W}_{opt}(k,\ell) = \underbrace{\frac{\mathbf{\Phi}_{\text{VV}}^{-1}(k,\ell)\mathbf{H}(k;\theta,r)}{\mathbf{H}^H(k;\theta,r)\mathbf{\Phi}_{\text{VV}}^{-1}(k,\ell)\mathbf{H}(k;\theta,r)}}_{\mathbf{W}_{\text{mvdr}}(k,\ell)} \cdot H_{\text{post}}(k,\ell), \tag{8}$$

where
$$H_{\text{post}}(k,\ell) = \frac{\Phi_{\text{SS}}(k,\ell)}{\Phi_{\text{SS}}(k,\ell) + \Phi_{\text{nn}}(k,\ell)}. \tag{9}$$

With $\Phi_{\text{SS}}(k,\ell)$ we denote the power spectral density of the source signal whereas with $\Phi_{\text{nn}}(k,\ell)$ the power spectrum of the noise at the output of the beamformer which equals to

$$\Phi_{\text{nn}}(k,\ell) = \Phi_{nf}(k,\ell)\mathbf{W}_{\text{mvdr}}^H(k,\ell)\mathbf{\Phi}_{\text{VV}}(k,\ell)\mathbf{W}_{\text{mvdr}}(k,\ell). \tag{10}$$

The quantity $\Phi_{nf}(k,\ell)$ is the normalization factor of the noise cross-power spectral matrix defined as

$$\Phi_{nf}(k,\ell) = \frac{1}{M}\sum_{p=0}^{M-1}\Phi_{V_pV_p}(k,\ell). \tag{11}$$

In the case of the MVDR beamformer the weight vector $\mathbf{W}_{\text{mvdr}}(k,\ell)$ can be evaluated since it is data independent, though this is not possible for the Wiener post-filter. As can be seen by Eq. (9), the solution depends on the knowledge of $\Phi_{\text{SS}}(k,\ell)$. Since the original values of $\Phi_{\text{SS}}(k,\ell)$ are not available, estimation is necessary. In the next sections this paper focuses on addressing the problem of estimating the Wiener post-filter transfer function.

3. POST-FILTER ESTIMATION

In the current section we first provide a short review of McCowan's post-filter estimation method [6] and then we propose a new estimation scheme that succeeds to provide a general post-filter as McCowan's, appropriate for a variety of different noise fields, and also be optimal in the Wiener sense. In addition we point out the similarities and differences of the discussed methods.

An overview of the overall multichannel noise reduction system is provided in Fig. 1. At the output of the sensors the multichannel input signals are time aligned and scaled to compensate for the time delay and attenuation, caused by the propagation of the source signal on the acoustic paths. According to this, $\mathbf{H}(k;\theta,r)$ will be equal to a M column vector of ones, $\mathbf{I}$. The signals at the delay compensation output can be denoted in matrix notation as

$$\mathbf{Y}(k,\ell) = \mathbf{I}\cdot S(k,\ell) + \mathbf{V}(k,\ell). \tag{12}$$

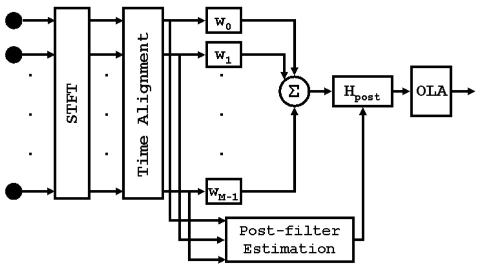


Figure 1: Block diagram of the noise reduction system.

3.1 McCowan's Post-Filter

Computing the auto and cross power spectral densities of the time aligned input signals on sensors p and q, leads to

$$\Phi_{Y_pY_q} = \Phi_{\text{SS}} + \Phi_{V_pV_q} + \Phi_{SV_p} + \Phi_{SV_q} \tag{13a}$$
$$\Phi_{Y_pY_p} = \Phi_{\text{SS}} + \Phi_{V_pV_q} + 2\Re\left\{\Phi_{SV_p}\right\}. \tag{13b}$$

The formulation of McCowan's post-filter is based on the following assumptions:

1. The speech and noise signals are uncorrelated, $\Phi_{SV_p} = 0\,\forall p$.
2. The noise field is homogeneous, meaning that the noise power spectrum is the same on all sensors, $\Phi_{V_pV_p} = \Phi_{\text{VV}}$.
3. An estimation of the coherence function $\Gamma_{V_pV_q}(\omega)$ is given.

Under these assumptions and by Eqs. (5) and (13) it follows that:

$$\Phi_{Y_pY_q} = \Phi_{\text{SS}} + \Phi_{V_pV_q} \tag{14a}$$
$$\Phi_{Y_pY_p} = \Phi_{\text{SS}} + \Phi_{\text{VV}} \tag{14b}$$
$$\Phi_{V_pV_q} = \Phi_{\text{VV}}\Gamma_{V_pV_q}. \tag{14c}$$

Equation set (14) forms a 3×3 linear system. Noting that under the adopted assumptions it holds $\Phi_{Y_pY_p}(k,\ell) = \Phi_{Y_qY_q}(k,\ell)$ and solving for Φ_{SS} we obtain:

$$\hat{\Phi}_{\text{SS}}^{(pq)} = \frac{\Re\left\{\hat{\Phi}_{Y_pY_q}\right\} - \frac{1}{2}\left(\hat{\Phi}_{Y_pY_p} + \hat{\Phi}_{Y_qY_q}\right)\Re\left\{\hat{\Gamma}_{V_pV_q}\right\}}{1 - \Re\left\{\hat{\Gamma}_{V_pV_q}\right\}}. \tag{15}$$

which is the derived estimation of $\Phi_{\text{SS}}(k,\ell)$ using the auto- and cross-spectral densities between sensors p and q. The notation $\hat{(\cdot)}$ stands for the estimated quantity. The average between the auto-spectral densities of channels p and q is taken to improve robustness. In $\Phi_{Y_pY_q}$ the real operator $\Re\{\cdot\}$ is used according to the definition that the power spectrum must always be real. Robustness can be further improved by taking the average over all $\binom{M}{2}$ possible combinations of channels p and q, resulting in

$$\hat{\Phi}_{\rm SS} = \frac{2}{M(M-1)} \sum_{p=0}^{M-2} \sum_{q=p+1}^{M-1} \hat{\Phi}_{\rm SS}^{(pq)} . \tag{16}$$

The post-filter denominator is estimated by $\hat{\Phi}_{Y_pY_p}$, as for the Zelinski technique and the transfer function of the post-filter is expressed as

$$\hat{H}_M = \frac{\hat{\Phi}_{\rm SS}}{\frac{1}{M}\sum\limits_{p=0}^{M-1} \hat{\Phi}_{Y_pY_p}} . \tag{17}$$

As it has already been mentioned, Zelinski's post-filter is a special case of McCowan's general expression. This can be verified by Eq. (15): For a spatially uncorrelated noise field the coherence function will equal to $\hat{\Gamma}_{V_pV_q} = 0$. Thus $\hat{\Phi}_{\rm SS}^{(pq)}(k,\ell) = \Re\left\{\hat{\Phi}_{Y_pY_q}(k,\ell)\right\}$, i.e the spectral density estimation of the speech source in Zelinski's post-filter [4].

3.2 Proposed Generalized Post-Filter

In our proposed post-filter estimation scheme we adopt the same assumptions as McCowan et al. and estimate the power spectral density of the speech source, the numerator of the Wiener post-filter transfer function (9), as proposed in [6]. The difference between the two methods lies in the estimation of the post-filter's denominator. The denominator of Eq. (9) denotes the power spectrum of the MVDR beamformer's output. Denoting with Z the output of the beamformer, we can write

$$\Phi_{\rm ZZ} = \Phi_{\rm SS} + \Phi_{\rm nn} . \tag{18}$$

With the assumption of a homogeneous noise field, $\Phi_{\rm nn}$ can then be written from Eq. (10) as

$$\Phi_{\rm nn} = \Phi_{\rm VV} \mathbf{W}_{\rm mvdr}^{H} \mathbf{\Gamma}_{\rm VV} \mathbf{W}_{\rm mvdr} , \tag{19}$$

where $\mathbf{\Gamma}_{\rm VV}$[1] is the coherence matrix of the noise field:

$$\mathbf{\Gamma}_{\rm VV} = \begin{pmatrix} 1 & \Gamma_{V_0V_1} & \cdots & \Gamma_{V_0V_{M-1}} \\ \Gamma_{V_1V_0} & 1 & & \\ \vdots & & \ddots & \\ \Gamma_{V_{M-1}V_0} & \cdots & & 1 \end{pmatrix} \tag{20}$$

Solving the system (14) for $\Phi_{\rm VV}$ instead of $\Phi_{\rm SS}$, results in

$$\hat{\Phi}_{\rm VV}^{(pq)} = \frac{\frac{1}{2}\left(\hat{\Phi}_{Y_pY_p} + \hat{\Phi}_{Y_qY_q}\right) - \Re\left\{\hat{\Phi}_{Y_pY_q}\right\}}{1 - \Re\left\{\hat{\Gamma}_{V_pV_q}\right\}} , \tag{21}$$

which is the estimation of $\Phi_{\rm VV}$ using the auto- and cross-spectral densities between sensors p and q. The average between the auto-spectral densities of channels p and q is used to improve robustness. Further robustness on the solution can be established by taking the average between all combinations of channels p and q, resulting finally in

$$\hat{\Phi}_{\rm VV} = \frac{2}{M(M-1)} \sum_{p=0}^{M-2} \sum_{q=p+1}^{M-1} \hat{\Phi}_{\rm VV}^{(pq)} . \tag{22}$$

We must note that a problem may arise in the estimation of $\hat{\Phi}_{\rm SS}^{(pq)}$ (15) and $\hat{\Phi}_{\rm VV}^{(pq)}$ (21) in the case that $\hat{\Gamma}_{V_pV_q} = 1$, for all $p \neq q$. A possible solution proposed in [6] to deal with this problem would be to bound the model of the coherence function so as $\hat{\Gamma}_{V_pV_q} < 1$, for all $p \neq q$.

To estimate the power spectrum at the beamformer's output, with no prior knowledge of the $\Phi_{\rm SS}$ values, we use the existing estimations. The post-filter's denominator will then be

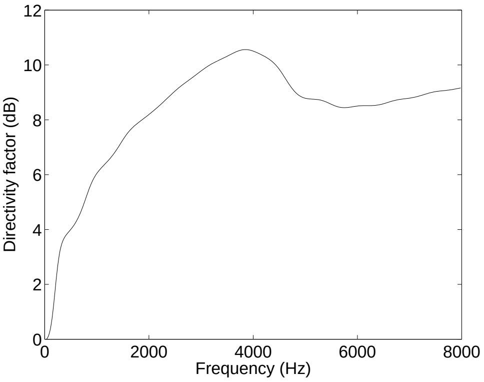


Figure 2: MVDR beamformer directivity factor.

$$\hat{\Phi}_{\rm ZZ} = \hat{\Phi}_{\rm SS} + \hat{\Phi}_{\rm VV} \mathbf{W}_{\rm mvdr}^{H} \hat{\mathbf{\Gamma}}_{\rm VV} \mathbf{W}_{\rm mvdr} . \tag{23}$$

An alternative approach would be to estimate the spectral density $\Phi_{\rm ZZ}$ directly from the output of the MVDR beamformer. However in such case the estimation would lack robustness since we would have available only one output signal to make the estimation, instead of N signals.

From Eqs. (9), (16) and (23) we obtain the transfer function of the Wiener post-filter

$$\hat{H}_{\rm prop} = \frac{\hat{\Phi}_{\rm SS}}{\hat{\Phi}_{\rm SS} + \hat{\Phi}_{\rm VV} \mathbf{W}_{\rm mvdr}^{H} \hat{\mathbf{\Gamma}}_{\rm VV} \mathbf{W}_{\rm mvdr}} . \tag{24}$$

At this point we have to note that in both methods of Zelinski [4] and McCowan [6], the estimated denominator given in (17), is an over-estimation of the noise power spectrum at the beamformer's output. This is attributed to the fact that the noise attenuation, already provided by the MVDR beamformer, is not taken into account. Therefore the derived filters are sub-optimal in the Wiener sense [6, 9].

4. EXPERIMENTS AND RESULTS

To validate the effectiveness of the proposed post-filter we compared its performance to other multi-channel noise reduction techniques, including the MVDR beamformer [3], the generalized Zelinski post-filter [4] and the McCowan post-filter [6], under the assumption of a diffuse noise field.

4.1 Speech Corpus and System Realization

The microphone data set used for the experiments is from CMU Microphone Array Database[10], recorded in a noisy computer lab at Carnegie Mellon University with many computer and disk-drive fans. The data set contains recordings by 10 male speakers of 13 utterances each. The recordings were collected by a linear microphone array. It consisted of 8 sensors with a spacing of 7 cm between adjacent sensors. The desired speech source was positioned directly in front of the array at a distance of 1 m from the center. All the recordings were sampled at 16 kHz with 16-bit linear sampling.

We window the sampled input signals into frames of 640 samples (40 ms) and apply to each frame a Hamming window. The overlap between adjacent frames is 480 samples (30 ms). Each data block is then Fourier transformed with a FFT of size 1024.

We first apply the MVDR beamformer to the multichannel noisy signals. Superdirective beamformers are known to be very sensitive to microphone mismatch and boost uncorrelated noise at lower frequencies. In order to overcome this problem of self-noise amplification we compute the MVDR weight vector under a White Noise Gain (WNG) constraint [11]. Under the assumption of a diffuse noise field the directivity factor of the beamformer is given by

[1] For the case of a homogeneous noise field $\mathbf{\Gamma}_{\rm VV} = \mathbf{\Phi}_{\rm VV}$.

$$Df = \frac{\left|\mathbf{W}^{H}_{\mathrm{mvdr}}\mathbf{H}\right|^{2}}{\mathbf{W}^{H}_{\mathrm{mvdr}}\boldsymbol{\Gamma}_{\mathrm{VV}}\mathbf{W}_{\mathrm{mvdr}}}\,. \tag{25}$$

The beamformer's output is further processed by the studied post-filters. To calculate the Wiener post-filters' transfer functions, the auto- and cross-spectral densities $\Phi_{Y_pY_p}$ and $\Phi_{Y_pY_q}$ have to be estimated. Due to the non-stationarity of the speech signals, only short data blocks are available for spectrum estimation. The power spectra are estimated using the short-time spectral estimation method proposed in [12], which can be viewed as a recursive Welch periodogram. This method smoothes the spectra in time and frequency and yields improved estimates. Finally, the output of the noise reduction system, Fig. 1, is transformed to the time-domain using the Overlap and Add synthesis (OLA) method .

4.2 Speech Enhancement Experiments

To demonstrate the benefits of estimating the post-filter transfer function with the proposed method, we use three different objective speech quality measures for the algorithms under test.

To assess the noise reduction, the segmental signal-to-noise ratio enhancement (SNRE) is used. The SNRE is defined as the difference in segmental SNR between the enhanced output and the noisy input of the noise reduction system, Fig. 1. The post-filter transfer function of each studied technique is derived by applying as inputs in the noise reduction system, the noisy speech signals. To calculate the SNRE, we compute the output of the noise reduction system using the clean speech and the noisy speech signals as inputs. In this way, we have available two signals at the output; the processed clean speech signal and the enhanced output signal. The segmental SNR is computed from consecutive samples with block size of $bs = 512$ samples. The quantities SNR$_{in}$, SNR$_{out}$ and SNRE are defined as follows:

$$SNR_{in}(\ell,i) = 10\log_{10}\left(\frac{\sum\limits_{k=1}^{bs}|S(k,\ell)|^2}{\sum\limits_{k=1}^{bs}\left||Y_i(k,\ell)|^2-|S(k,\ell)|^2\right|}\right) \tag{26}$$

$$SNR_{out}(\ell) = 10\log_{10}\left(\frac{\sum\limits_{k=1}^{bs}|F_s(k,\ell)|^2}{\sum\limits_{k=1}^{bs}\left||F(k,\ell)|^2-|F_s(k,\ell)|^2\right|}\right) \tag{27}$$

$$SNRE(\ell) = SNR_{out}(\ell) - \frac{1}{M}\left(\sum_{i=0}^{M-1} SNR_{in}(\ell,i)\right), \tag{28}$$

where $F(k,\ell)$ and $F_s(k,\ell)$ are the short-time Fourier transforms of the enhanced noisy signal and the processed speech signal respectively.

To assess the speech quality of the enhanced output signal, the Log-Area-Ratio distance (LAR) and the speech degradation (SD) measure are used. These measures are found to have a high correlation with the human perception [13]. Low LAR and SD values denote high speech quality. The LAR distance and the SD measure are defined according to the following formulas:

$$LAR(\ell) = \frac{1}{P}\sum_{p=1}^{P}\left|20\log 10\left|\frac{g_{\mathrm{s}}(p,\ell)}{g_{\mathrm{f}}(p,\ell)}\right|\right| \tag{29}$$

$$SD(\ell) = \frac{1}{P}\sum_{p=1}^{P}\left|20\log 10\left|\frac{g_{\mathrm{s}}(p,\ell)}{g_{f_{\mathrm{s}}}(p,\ell)}\right|\right|, \tag{30}$$

where $g_{\mathrm{s}}(p,\ell)$, $g_{\mathrm{f}}(p,\ell)$ and $g_{f_{\mathrm{s}}}(p,\ell)$ represent the pth area ratio function of the desired signal, the enhanced signal and the processed clean signal respectively, computed over the ℓth frame.

For every speaker of the test set, the SNRE, LAR and SD results are averaged across all the 13 utterances and are shown in Tables 1–3. In addition, Fig. 3 shows the spectrograms of the clean and the noisy input signal along with the output signals of the studied methods, for an utterance corresponding to the word "thomas".

From Figs. 3(c) and 3(d) we note that neither the beamformer alone nor the Zelinski post-filter can remove sufficiently the noise in the low frequency region. This inadequacy is also illustrated in Table 3, where the SNR enhancement of the above two methods is quite poor compared to the SNR enhancement provided by McCowan's and by the proposed post-filter. What is also noteworthy from the results in Table 3, is that Zelinski's post-filter not only gives the lowest SNRE of all the studied methods, but in addition in some cases the output SNR is smaller than the input SNR (negative SNRE). An explanation can be found in [13], where it has been shown that Zelinski's method, works well only for reverberation times above 300 ms. For very low reverberation times, the output speech quality is poorer than the input speech quality. The low SNRE of the MVDR beamformer, can be attributed to the fact that the greatest portion of the noise energy is concentrated in the low frequency region, where the beamformer has a low directivity factor (Fig. 2). Comparing the spectrograms of Figs. 3(e), 3(f) derived by applying McCowan's and the proposed post-filter, respectively, at the output of the beamformer, we can note that even though McCowan's post-filter performs sufficient noise reduction at low frequencies, its behavior at mid and high frequencies is not as efficient as the proposed post-filter. From Fig. 3 it can also be seen that the spectrogram closest to the clean speech is the one derived by applying the proposed post-filter. This is due to the fact that the proposed post-filter performs a sufficient noise reduction on every frequency region (low-mid-high).

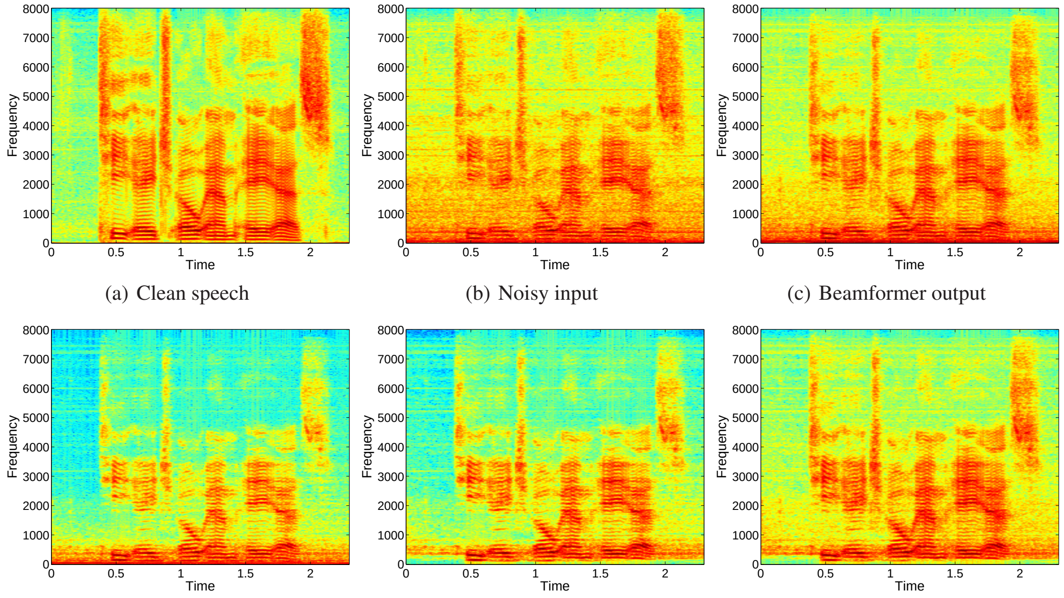
From the results in Tables 1, 2 and 3 it is clearly evident that the proposed post-filter consistently outperforms all the other methods as it produces the best results for all the objective measures. It gives the greater noise reduction while still providing the highest speech quality signal. In particular the proposed post-filter estimation scheme presents a relative SNR enhancement of 17.3% and a relative decrease on signal degradation of 21.7% compared to the best of all the other studied methods (McCowan's Post-filter).

Table 1: LAR Results

Speaker	Noisy Input	LAR (dB)			
		MVDR	Zel.	Mc.	Prop.
sp1	2.83	3.91	5.23	3.94	2.96
sp2	3.00	4.00	4.98	3.33	2.66
sp3	2.98	3.79	4.70	3.11	2.56
sp4	3.03	3.96	5.15	3.30	2.60
sp5	3.01	3.85	5.00	3.28	2.52
sp6	3.26	3.85	5.12	3.44	2.77
sp7	3.12	3.74	4.65	3.18	2.59
sp8	3.23	3.90	4.99	3.39	2.70
sp9	3.05	3.89	4.80	3.06	2.41
sp10	3.06	3.75	5.14	3.55	2.64
mean	3.05	3.86	4.98	3.36	2.64

Table 2: SD Results

Speaker	SD (dB)			
	MVDR	Zel.	Mc.	Prop.
sp1	3.91	5.35	4.07	3.03
sp2	4.00	5.09	3.42	2.72
sp3	3.79	4.80	3.22	2.63
sp4	3.96	5.25	3.41	2.67
sp5	3.85	5.11	3.38	2.59
sp6	3.85	5.24	3.56	2.85
sp7	3.74	4.74	3.27	2.65
sp8	3.90	5.10	3.48	2.76
sp9	3.89	4.89	3.15	2.47
sp10	3.75	5.25	3.65	2.70
mean	3.86	5.08	3.46	2.71



(a) Clean speech

(b) Noisy input

(c) Beamformer output

(d) Zelinski post-filter

(e) McCowan post-filter

(f) Proposed post-filter

Figure 3: Speech Spectrograms. (a)Original clean speech signal:"thomas". (b)Noisy signal at sensor #4. (c) Beamformer output (SNRE=2.23 dB, SD=3.65 dB, LAR=3.65 dB). (d)Zelinski post-filter (SNRE=1.44 dB, SD=5.15 dB, LAR=5.02 dB). (e)McCowan post-filter (SNRE=5.97 dB, SD=4.31 dB, LAR=4.16 dB). (f)Proposed post-filter (SNRE=7.07 dB, SD=3.15 dB, LAR=3.09 dB).

Table 3: SNRE Results

	SNRE (dB)			
Speaker	MVDR	Zel.	Mc.	Prop.
sp1	1.95	0.51	8.35	9.98
sp2	2.08	0.71	12.55	14.49
sp3	2.01	0.99	12.49	13.97
sp4	1.76	-0.26	11.12	13.57
sp5	1.86	0.11	10.94	12.90
sp6	2.22	0.70	11.34	12.89
sp7	2.38	0.46	11.01	12.82
sp8	1.62	0.03	9.03	10.91
sp9	1.84	-0.48	11.25	13.60
sp10	3.00	0.76	11.19	13.11
mean	2.07	0.35	10.93	12.82

5. CONCLUSIONS

In this paper a multichannel noise reduction system with additional post-filtering has been presented. The proposed post-filter estimation scheme is an extension of the existing Zelinski's and McCowan's post-filters. While in these two methods an overestimation of the spectral density in the output of the beamformer has been used, which constitutes these methods sub-optimal in terms of MMSE, the proposed post-filter takes into account the noise reduction performed by the beamformer and produces a robust spectral estimation that satisfies the MMSE optimality of the Wiener filter. In experiments with real noise multichannel recordings from a noisy computer lab, the proposed technique has shown to obtain a significant gain over the other studied methods in terms of signal-to-noise ratio, log area ratio distance and speech degradation measure. In particular the proposed post-filter presents a relative SNR enhancement of 17.3% and a relative decrease on signal degradation of 21.7% compared to the best of all the other studied methods.

REFERENCES

[1] B. D. Van Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine*, vol. 5, pp. 4–24, 1988.

[2] 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., chapter 3, pp. 39–60. Springer Verlag, 2001.

[3] J. Bitzer and K. U. Simmer, "Superdirective Microphone Arrays," in *Microphone Arrays: Signal Processing Techniques and Applications*, M. Brandstein and D. Ward, Eds., chapter 2, pp. 19–38. Springer Verlag, 2001.

[4] R. Zelinski, "A Microphone Array With Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms," in *ICASSP*, 1988, vol. 5, pp. 2578–2581.

[5] J. Meyer and K. U. Simmer, "Multi-Channel Speech Enhancement in a Car Environment Using Wiener Filtering and Spectral Subtraction," in *ICASSP*, 1997, vol. 2, pp. 1167–1170.

[6] I. A. McCowan and H. Bourlard, "Microphone Array Post-Filter Based on Noise Field Coherence," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, 2003.

[7] J. Li and M. Akagi, "A Hybrid Microphone Array Post-Filter in a Diffuse Noise Field," in *Proc. Interspeech-Eurospeech*, 2005, pp. 2313–2316.

[8] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of Noise Reduction Techniques Based on Microphone Arrays with Postfiltering," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, 1988.

[9] S. Fischer and K. D. Kammeyer, "Broadband Beamforming With Adaptive Postfiltering for Speech Acquisition in Noisy Environments," in *ICASSP*, 1997, vol. 1, pp. 359–362.

[10] Tom Sullivan, "Cmu microphone array database," 1996, `http://www.speech.cs.cmu.edu/databases/micarray`.

[11] H. Cox, R. M. Zeskind, and T. Kooij, "Practical Supergain," *IEEE Trans. Speech and Audio Processing*, vol. 34, no. 3, pp. 393–398, 1986.

[12] J. B. Allen, D. A. Berkley, and J. Blauert, "Multimicrophone Signal-Processing Technique to Remove Room Reverberation from Speech Signals," *Journ. Acoustical Society of America*, vol. 62, no. 4, pp. 912–915, 1977.

[13] S. Fischer and K. U. Simmer, "Beamforming Microphone Arrays For Speech Acquisition in Noisy Environments," *Speech Communication*, vol. 20, pp. 215–227, 1996.