An Optimum Microphone Array Post-Filter for Speech Applications

Stamatis Leukimmiatis, Dimitrios Dimitriadis and Petros Maragos

National Technical University of Athens, School of ECE, Zografou, Athens 15773, Greece.

Email:[sleukim, ddim, maragos]@cs.ntua.gr

Abstract

This paper proposes a post-filtering estimation scheme for multichannel noise reduction. The proposed method extends and improves the existing Zelinski's and, the most general and prominent, McCowan's post-filtering methods that use the auto- and cross-spectral densities of the multichannel input signals to estimate the transfer function of the Wiener post-filter. A major drawback of these two speech enhancement algorithms 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 deals with this problem and can be considered as an optimal post-filter that is appropriate for a wide variety of different noise fields. In experiments over real-noise multichannel recordings, the proposed technique is shown to obtain a significant headstart over the other methods in terms of signal-to-noise ratio and speech degradation measures. In addition it is used for ASR experiments where promising preliminary results are presented.

Index Terms: Speech enhancement, microphone array, post-filter, complex coherence, speech recognition.

1. Introduction

Nowadays, the use of microphone arrays for speech enhancement and robust speech recognition seems very promising. The main advantage against single channel techniques is 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 considered. The spatial discrimination of the array is exploited by beamforming algorithms [1]. In many cases though, the obtainable noise reduction performance 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 taken from its noisy observations is obtained using the multichannel Wiener filter. Simmer et al. [2] have recently 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 levels than the MVDR beamformer alone, and therefore its integration in the beamformer output can lead to substantial SNR gain.

Despite its theoretical optimality, the Wiener post-filter is difficult to be realized. This is due to the requirement for the knowledge of second order statistics for both speech and the corrupting noise signals that make Wiener filter signal-dependent. A variety of post-filtering techniques trying to address this issue have been proposed in the literature [4, 5, 6]. A quite common method for the formulation of the post-filter's 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] that was further studied by Marro et al. [7]. A more generalized approach of the Zelinski's algorithm is based on the assumption of a spatially uncorrelated noise field. However, this assumption is not realistic for most of the practical applications. If a more accurate noise field model could be used, the overall performance of the noise reduction system would be improved. McCowan and Bourlard [6] assume a known noise field coherence function and propose a more general post-filtering scheme improving the overall performance. A certain drawback though in both methods is that the noise power spectrum at the beamformer's output is over-estimated [6, 8] 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 is optimal in terms of MMSE, allowing though, the development of a general post-filter appropriate for different noise fields. To meet with these demands we preserve the general assumption of a known noise field coherence function [6] but in addition, we take into account the noise reduction performed by the MVDR beamformer. This way, we estimate the speech source's spectrum like in [6] but we propose a new robust method for estimating the power spectrum at the beamformer's output, being consistent with the optimality in the MMSE sense. The enhanced speech signals are used in noisy ASR tasks where promising preliminary results are presented.

2. Microphone Array Post-Filtering

Considering an M-sensor linear microphone array, the observed signal $y_i(n)$, $i = 0, \ldots, M-1$, at the i^{th}-sensor is a delayed and attenuated version of the original speech signal $s(n)$ with an additive noise component $v_i(n)$. Applying the short-time Fourier transform (STFT), the observed information in the joint time-frequency domain can be written as

$$\mathbf{Y}(k,\ell) = \mathbf{H}(k)S(k,\ell) + \mathbf{V}(k,\ell)\,, \qquad (1)$$

where k and ℓ are the frequency bin and the time frame index, respectively. $\mathbf{Y}$, $\mathbf{H}$ and $\mathbf{V}$ are $M \times 1$ column vectors with $\mathbf{H}$ being the propagation vector of the signal source.

A standard method to estimate the desired signal, based on the observed noisy signals, is to compute the weight vector that transforms the corrupted input signal vector into the best MMSE approximation of the source signal. This weight vector is known as *multichannel Wiener filter*, and it can be further decomposed into a MVDR beamformer followed by a single-channel Wiener filter [2],

$$\mathbf{W}_{opt}(k,\ell) = \underbrace{\frac{\mathbf{\Phi}_{\text{VV}}^{-1}(k,\ell)\mathbf{H}(k)}{\mathbf{H}^H(k)\mathbf{\Phi}_{\text{VV}}^{-1}(k,\ell)\mathbf{H}(k)}}_{\mathbf{W}_{\text{mvdr}}(k,\ell)} \cdot H_{\text{post}}(k,\ell)\,, \qquad (2)$$

$$H_{\text{post}}(k,\ell) = \frac{\Phi_{\text{SS}}(k,\ell)}{\Phi_{\text{SS}}(k,\ell) + \Phi_{\text{nn}}(k,\ell)}\,. \quad (3)$$

In Eq. (2), $\mathbf{\Phi}_{\text{VV}}(k,\ell)$ is the normalized cross-power spectral density matrix of the noise, while in Eq. (3) $\Phi_{\text{SS}}(k,\ell)$ is the power spectral density of the source signal and $\Phi_{\text{nn}}(k,\ell)$ is the power spectrum of the noise at the output of the beamformer,

$$\Phi_{\text{nn}}(k,\ell) = \Phi_{nf}(k,\ell)\mathbf{W}^{H}_{\text{mvdr}}(k,\ell)\mathbf{\Phi}_{\text{VV}}(k,\ell)\mathbf{W}_{\text{mvdr}}(k,\ell)\,. \quad (4)$$

The quantity $\Phi_{nf}(k,\ell)$ is the normalization factor of the noise cross-power spectral matrix defined as the ratio of the matrix trace to the number of its diagonal elements. In the case of the MVDR beamformer, the weight vector $\mathbf{W}_{\text{mvdr}}(k,\ell)$ can be evaluated since it is data independent. However, for the Wiener post-filter the solution depends on $\Phi_{\text{SS}}(k,\ell)$.

3. Post-Filter Estimation

At first, we introduce the coherence function which provides us with a model for the noise field. Then, we propose a new estimation scheme that extends McCowan's post-filter estimation method and succeeds to provide a general post-filter, appropriate for different noise fields that is optimal in the Wiener sense. In addition, we point out the similarities and differences of these two methods.

3.1. Noise Field

In microphone array applications, noise fields can be characterized by a measure known as the *complex coherence function*. Coherence function measures the amount of correlation between noise signals at different spatial locations and is defined [3] as

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

where ω is the discrete-time angular frequency, $\Phi_{V_pV_q}(\omega)$ 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 an equally distributed uncorrelated white noise signal 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 is approximated by

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

where d is the distance between sensors p and q and f_s is the sampling frequency.

3.2. Proposed Generalized Post-Filter

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, thus $\mathbf{H}(k)$ equals to a $M \times 1$ column vector of ones, $\mathbf{I}$. The signals at the delay compensation output are denoted as

$$\mathbf{Y}(k,\ell) = \mathbf{I}\cdot S(k,\ell) + \mathbf{V}(k,\ell) \quad (7)$$

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

$$\Phi_{Y_pY_q} = \Phi_{\text{SS}} + \Phi_{V_pV_q} + \Phi_{SV_p} + \Phi_{SV_q} \quad (8a)$$

$$\Phi_{Y_pY_p} = \Phi_{\text{SS}} + \Phi_{V_pV_q} + 2\Re\ \ \Phi_{SV_p} \quad (8b)$$

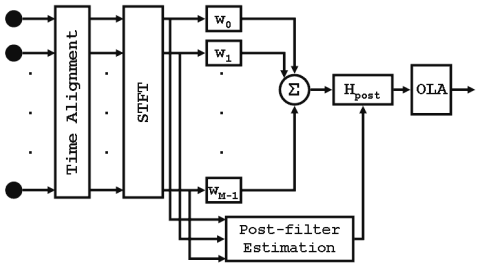

Figure 1: Block diagram of the noise reduction system.

The formulation of the proposed post-filter is based on the same assumptions adopted in [6]: (i) The speech and noise signals are uncorrelated, (ii) The noise field is homogeneous, i.e $\Phi_{V_pV_p} = \Phi_{\text{VV}}$, (iii) An estimation of the coherence function $\Gamma_{V_pV_q}(\omega)$ is given. Under these assumptions and by combining Eqs. (5) and (8) it follows that,

$$\Phi_{Y_pY_q} = \Phi_{\text{SS}} + \Phi_{V_pV_q} \quad (9a)$$

$$\Phi_{Y_pY_p} = \Phi_{\text{SS}} + \Phi_{\text{VV}} \quad (9b)$$

$$\Phi_{V_pV_q} = \Phi_{\text{VV}}\Gamma_{V_pV_q}\,. \quad (9c)$$

Equation set (9) forms a 3×3 linear system. Noting that it holds $\Phi_{Y_pY_p}(k,\ell) = \Phi_{Y_qY_q}(k,\ell)$ and solving for Φ_{SS}, the input signal power spectral density is computed as [6],

$$\hat{\Phi}^{(pq)}_{\text{SS}} = \frac{\Re\left\{\hat{\Phi}_{Y_pY_q}\right\} - \frac{1}{2}\ \ \hat{\Phi}_{Y_pY_p} + \hat{\Phi}_{Y_qY_q}\ \ \Re\left\{\hat{\Gamma}_{V_pV_q}\right\}}{1 - \Re\left\{\hat{\Gamma}_{V_pV_q}\right\}}\,. \quad (10)$$

The average sum of the auto-spectral densities for channels p and q is considered to improve robustness. Robustness can be further improved by taking the average sum over all $\binom{M}{2}$ possible combinations of channels p and q, resulting to

$$\hat{\Phi}_{\text{SS}} = \frac{2}{M(M-1)}\sum_{p=0}^{M-2}\sum_{q=p+1}^{M-1}\hat{\Phi}^{(pq)}_{\text{SS}}\,. \quad (11)$$

The denominator of Eq. (3) denotes the power spectrum of the MVDR's output. If Z is the output of the beamformer, then

$$\Phi_{\text{ZZ}} = \Phi_{\text{SS}} + \Phi_{\text{nn}}\,. \quad (12)$$

We propose a more robust and accurate way of estimating Φ_{nn}. Assuming a homogeneous noise field, it holds $\mathbf{\Phi}_{\text{VV}} = \mathbf{\Gamma}_{\text{VV}}$ and $\Phi_{nf} = \Phi_{\text{VV}}$. Thus Φ_{nn} can be written from Eq. (4) as

$$\Phi_{\text{nn}} = \Phi_{\text{VV}}\mathbf{W}^{H}_{\text{mvdr}}\mathbf{\Gamma}_{\text{VV}}\mathbf{W}_{\text{mvdr}}\,. \quad (13)$$

where $\mathbf{\Gamma}_{\text{VV}}$ is the coherence matrix of the noise field.
Solving the system in (9) for Φ_{VV} instead of Φ_{SS}, the noise power spectrum is estimated as:

$$\hat{\Phi}^{(pq)}_{\text{VV}} = \frac{\frac{1}{2}\ \ \hat{\Phi}_{Y_pY_p} + \hat{\Phi}_{Y_qY_q}\ \ -\Re\left\{\hat{\Phi}_{Y_pY_q}\right\}}{1 - \Re\left\{\hat{\Gamma}_{V_pV_q}\right\}}\,. \quad (14)$$

Following the previous clues, additional robustness can be established by averaging all combinations of channels p and q, resulting in

$$\hat{\Phi}_{\text{VV}} = \frac{2}{M(M-1)}\sum_{p=0}^{M-2}\sum_{q=p+1}^{M-1}\hat{\Phi}^{(pq)}_{\text{VV}}\,. \quad (15)$$

We must note that a problem may arise in the estimation of $\hat{\Phi}_{\text{SS}}^{(pq)}$ (10) and $\hat{\Phi}_{\text{VV}}^{(pq)}$ (14) when $\hat{\Gamma}_{V_pV_q} = 1$, for all $p \neq q$. A possible solution proposed in [6] is 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 Φ_{SS} values, we use the existing estimations. The post-filter's denominator becomes

$$\hat{\Phi}_{\text{ZZ}} = \hat{\Phi}_{\text{SS}} + \hat{\Phi}_{\text{VV}} \mathbf{W}_{\text{mvdr}}^H \hat{\mathbf{\Gamma}}_{\text{VV}} \mathbf{W}_{\text{mvdr}} \,. \tag{16}$$

An alternative approach would be to estimate the spectral density Φ_{ZZ} directly from the output of the MVDR beamformer. However, in such case the estimation would lack robustness since only one output signal would be available for the estimation process, instead of N signals.

From Eqs. (3), (11) and (16) we finally obtain the transfer function of the Wiener post-filter

$$\hat{H}_{\text{prop}} = \frac{\hat{\Phi}_{\text{SS}}}{\hat{\Phi}_{\text{SS}} + \hat{\Phi}_{\text{VV}} \mathbf{W}_{\text{mvdr}}^H \hat{\mathbf{\Gamma}}_{\text{VV}} \mathbf{W}_{\text{mvdr}}} \,. \tag{17}$$

At this point we have to note that in both methods proposed in [4] and [6], the post-filter's denominator is

$$\hat{\Phi}_{\text{ZZ}} = \frac{1}{M} \sum_{p=0}^{M-1} \hat{\Phi}_{Y_pY_p} \,. \tag{18}$$

This is an over-estimation of the noise power spectrum at the beamformer's output due to the fact that the noise attenuation, already provided by the MVDR beamformer, is not taken into account in the post-filtering process. Therefore the derived filters are sub-optimal in the Wiener sense [6, 8].

4. Speech Experiments and Results

The effectiveness of the proposed post-filter is examined by comparing it with the other multi-channel, noise-reduction techniques, including the MVDR beamformer [3], the generalized Zelinski [4] and the McCowan post-filters [6], under the assumption of a diffuse noise field.

4.1. Speech Corpus and System Realization

The microphone data set (source signals) used for the experiments is taken from the TIDIGITS database and recorded in a room with diffuse noise. The recordings were collected by a linear microphone array consisting of 16 sensors with a spacing of 2cm between the adjacent sensors. The desired speech source was positioned directly in front of the array at a distance of 1.3m from the center. All the recordings were sampled at 16 kHz. The data set contains recordings from 52 male and 52 female adult speakers.

We window the sampled input signals into frames of length 400 samples (25 ms) and apply to each frame a Hamming window. The overlap between adjacent frames equals to 300 samples ($\approx$ 19 ms). Each data block is then transformed in the frequency domain with a FFT of size 512 samples.

The MVDR weight vector is estimated under a White Noise Gain (WNG) constraint [9]. The multichannel noisy signals are first applied to the MVDR beamformer. The beamformer's output is further processed by the above-mentioned post-filters. To calculate the Wiener post-filters' transfer functions, the auto- and cross-spectral densities have to be estimated. The power spectra are estimated using the short-time spectral estimation method proposed in [10]. This method smoothes the spectra in both time and frequency domains yielding 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.

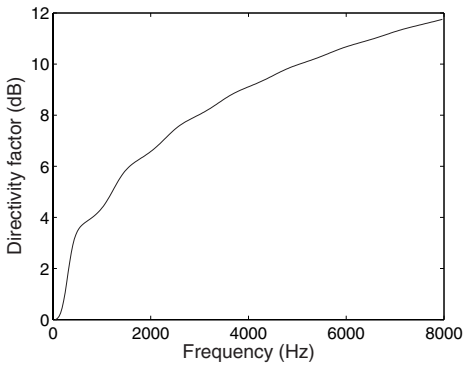

Figure 2: MVDR beamformer directivity factor.

4.2. Speech Enhancement Experiments

The advantages of estimating the post-filter transfer function with the proposed method are examined by two different objective speech quality measures.

At first, 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 for the examined algorithms is derived when applying the noisy speech signals to the noise reduction system. For the calculation of the SNRE, we estimate the output of the noise reduction system using the clean, noisy speech and the noise signals as inputs. This way, we have available three signals as output; the processed clean speech signal, the enhanced output signal and the processed noise. The segmental SNR is estimated over consecutive samples with block size of 512 samples according to the definition given in [2]. Then, the speech degradation (SD) measure is used to assess the speech quality of the enhanced output signal. This measure is found to be highly correlated with the human perception and provides a quantitative measure of the speech distortion caused by the overall system. Low SD values denote high speech quality. The SD measure is defined according to the formula provided in [2].

The SNRE and SD results are averaged across the male and the female data set and are presented in Table 1. Examining these results we note that neither the beamformer alone nor the Zelinski post-filter can remove sufficiently the noise. The low SNRE results 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). The poor performance of Zelinski's post-filter is expected since it assumes a noise field model that is inappropriate. An additional explanation can be found in [11] where it is shown that this method, works well only for reverberation times above 300 ms. On the other hand, McCowan's method performs better than the previous two. However, we note 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 great-

est noise reduction while still providing the highest speech quality signal.

Method	Male		Female		Total	
	SNRE	SD	SNRE	SD	SNRE	SD
MVDR	0.47	3.66	0.37	3.67	0.42	3.66
Zelinski	0.16	5.10	0.01	5.12	0.09	5.11
McCowan	1.40	3.90	3.21	4.09	2.31	4.00
Proposed	3.73	3.37	5.14	3.42	4.44	3.40

Table 1: Speech Enhancement Results (dB)

4.3. Speech Recognition Experiments

The examined algorithms are used for the feature extraction process of an HMM-based recognizer like the HTK Toolkit. We extract features from the enhanced signals and apply them to the HTK Toolkit to examine their impact on ASR tasks. For the ASR experiments, we have used the same database as mentioned in Section 4, dividing it into 2 separate sets though. In detail, 700 of the sentences are used as the training corpus and the other 300 are used as the testing sentences. Context-independent, 12-state, left-right word HMMs with 3 gaussian mixtures are used. The grammar used is the all-pair, unweighted grammar. We have examined all seven different versions of the speech sentences, the source, clean and noisy data and the four enhanced versions.

Computational Architecture: For the ASR experiments we are using an attractive computational architecture for the feature extraction process taking advantage of the STFT. The extracted features are the widely-used MFCC features plus their first and second-order time derivatives. The features are extracted directly from the frequency versions of the enhanced signals before resynthesizing them with the OLA algorithm, to avoid inserting additional modeling errors. For the noisy speech data, the signals recorder by the central microphone are considered.

Herein, in Table 2, the ASR results only for the most prominent algorithms are presented due to lack of space. Note that the speech database used was not originally designed for ASR tasks so it lacks of training and testing variability in speakers and sentences. This is the main reason for the very high correct accuracy percentages.

Correct Word Accuracies (%)				
Input Signals for ASR task	Original	Noisy	McCowan	Proposed
MFCC+D+DD	96.37	94.98	93.83	95.23

Table 2: ASR Results For the Matched Training/Testing Scenario.

5. Conclusions

In this paper a multichannel noise reduction system with an additional post-filtering process has been presented. The proposed post-filter estimation scheme is an extension of the existing McCowan post-filter. This method and its special case, the Zelinski post-filter, use an over-estimation of the spectral density in the output of the beamformer, which constitutes them sub-optimal in terms of MMSE. On the other hand, 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 room, the proposed technique has been shown to obtain a significant gain over McCowan post-filter in terms of signal-to-noise ratio, speech degradation measure and speech recognition performance. The ASR results yielded by the proposed algorithm are close to the clean-speech case. These results seem promising and further research in feature extraction and estimation is in our near future research plans.

6. Acknowledgments

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

7. 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*, 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*, 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] 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.

[8] 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.

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

[10] 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.

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