Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 657–666
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A generalized estimation approach for linear and nonlinear microphone array post-filters ☆

Stamatios Lefkimmiatis *, Petros Maragos

School of Electrical and Computer Engineering, National Technical University of Athens, Athens 15773, Greece

Received 30 June 2006; received in revised form 18 January 2007; accepted 4 February 2007

Abstract

This paper presents a robust and general method for estimating the transfer functions of microphone array post-filters, derived under various speech enhancement criteria. For the case of the mean square error (MSE) criterion, the proposed method is an improvement of the existing McCowan post-filter, which under the assumption of a known noise field coherence function uses the auto- and cross-spectral densities of the microphone array noisy inputs to estimate the Wiener post-filter transfer function. In contrast to McCowan post-filter, the proposed method takes into account the noise reduction performed by the minimum variance distortionless response (MVDR) beamformer and obtains a more accurate estimation of the noise spectral density. Furthermore, the proposed estimation approach is general and can be used for the derivation of both linear and nonlinear microphone array post-filters, according to the utilized enhancement criterion. 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 five different objective speech quality measures.

Keywords: Nonlinear; Noise reduction; Speech enhancement; Microphone array; Post-filter; Complex coherence

1. Introduction

The problem of multichannel speech enhancement has received much attention the last two decades. The main advantage of microphone arrays against single channel techniques is that they can simultaneously exploit the spatial diversity of speech and noise, so that both spectral and spatial characteristics of signals are considered. The spatial discrimination of an array is exploited by beamforming algorithms (Veen and Buckley, 1988). 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 most common-used criterion for speech enhancement is the mean-square error (MSE), leading to the Multichannel Wiener filter. This optimal multichannel MSE filter has been shown in Simmer et al. (2001) and Trees (2002) that can be factorized into a minimum variance distortionless response (MVDR) beamformer, followed by a single channel Wiener post-filter. However, the MSE distortion of the signal estimate is essentially not the optimum criterion for speech enhancement (Ephraim and Mallah, 1984; Ephraim and Mallah, 1985). More appropriate distortion measures for speech enhancement are based either on the MSE of the spectral amplitude or on the MSE of the log-spectral amplitude, leading to the short-time spectral amplitude (STSA) estimator (Ephraim and Mallah, 1984) and the log-spectral amplitude (log-STSA) estimator (Ephraim and Mallah, 1985), respectively. These estimators have also been proved to decompose into a MVDR beamformer followed by a single channel post-filter (Balan and Rosca, 2002). In general, all these post-filters accomplish higher

☆ This work was supported by the Greek GSRT under the research program ΠΕΝΕΔ space 2003-ΕΔ554 and in part by the European research program HIWIRE. Audiofiles available. See htpp://www.elsevier.com/locate/specom.
* Corresponding author.
E-mail addresses: sleukim@cs.ntua.gr (S. Lefkimmiatis), maragos@cs.ntua.gr (P. Maragos).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.02.004

noise reduction than the MVDR beamformer alone, therefore their integration in the beamformer output leads to substantial SNR gain.

Despite their theoretically optimal results, Wiener, STSA and log-STSA post-filters are 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 these filters signal-dependent. A variety of post-filtering techniques trying to address this issue have been proposed in the literature (Zelinski, 1988; Fischer and Simmer, 1996; Meyer and Simmer, 1997; Cohen and Berdugo, 2002; McCowan and Bourlard, 2003; Cohen, 2004). A quite common method for the formulation of the post-filter transfer function is based on the use of the auto- and cross-power spectral densities of the multichannel input signals (Simmer et al., 2001; Zelinski, 1988; McCowan and Bourlard, 2003). One of the early methods for post-filter estimation is due to (Zelinski, 1988), which was further studied by Marro et al. (1988). 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 of the practical applications, since the correlation of the noise between different channels can be significant, particularly at low frequencies. If a more accurate model of the noise field could be used, the overall performance of the noise reduction system would be improved. McCowan and Bourlard (2003) replaced this assumption by the more general assumption of a known noise field coherence function and extended the previous method (Zelinski, 1988) to develop a more efficient post-filtering scheme. However, a drawback in both methods is that the noise power spectrum at the beamformer's output is over-estimated (McCowan and Bourlard, 2003; Fischer and Kammeyer, 1997) and therefore the derived filters are sub-optimal. Moreover, these two estimation methods are not applicable for the cases of the STSA and log-STSA post-filters, a subject on which we will focus in detail.

In this paper, we deal with the problem of estimating the transfer functions of microphone array post-filters, derived under the three most commonly used speech enhancement criteria (MSE, MSE-STSA, MSE log-STSA). Specifically, we present a robust method for estimating the speech and noise power spectral densities to be used in the transfer functions. This method is general, appropriate for a variety of different noise conditions, as it preserves the general assumption of a known model for the coherence function of the noise field; and can be applied to both linear and non-linear post-filters. The noise power spectrum is estimated by taking into account the noise reduction performed already by the MVDR beamformer. This approach is different from the one followed by McCowan and Bourlard (2003) who ignored this noise reduction in their method. In this way it is shown that the obtainable estimation of the noise spectral density is more accurate and leads to better results. This is confirmed with experiments on the CMU multichannel database (Sullivan, 1996), by using five different objective speech quality measures.

The rest of this paper is organized as follows: Section 2 contains mainly background material. It describes the recording procedure for speech signals in a noisy acoustic environment and establishes the statistical model for multichannel speech enhancement in the joint time–frequency domain. In addition discusses the derivation of the MVDR beamformer along with the Wiener, STSA and log-STSA post-filters. The main contributions of this paper are in Sections 3 and 4. In Section 3 the coherence function, a popular measure for characterizing different noise fields, is presented and a novel post-filter estimation scheme is proposed. Finally, in Section 4 the performance of the proposed method is evaluated in speech enhancement experiments, using multichannel noisy office recordings.

2. Multichannel speech enhancement

Let us consider a N-sensor linear microphone array in a noisy environment where a desired source signal is located at a distance r and at an angle θ from the center of the array. The observed signal, $x_i(n)$, $i = 0, \ldots, N-1$, at the ith sensor corresponds to a linearly filtered version of the source signal $s(n)$, plus an additive noise component $v_i(n)$:

$$x_i(n) = d_i(n;\theta,r) * s(n) + v_i(n), \tag{1}$$

where $d_i(n;\theta,r)$ is the impulse response of the acoustic path from the desired source to the ith sensor and $*$ denotes convolution. Due to the non-stationary nature of the speech and the noise components, a short-time analysis must follow. The observed signals are divided in time into overlapping frames and in every frame a window function is applied. Then, each frame is analyzed by means of the short-time Fourier transform (STFT). Assuming time-invariant transfer functions we can express the observed information in the joint time–frequency domain as

$$\boldsymbol{X}(k,\ell) = \boldsymbol{D}(k;\theta,r)S(k,\ell) + \boldsymbol{V}(k,\ell), \tag{2}$$

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

$$\boldsymbol{X}(k,\ell) = [X_0(k,\ell)\ X_1(k,\ell)\ \cdots\ X_{N-1}(k,\ell)]^{\mathrm{T}},$$
$$\boldsymbol{D}(k;\theta,r) = [D_0(k;\theta,r)\ D_1(k;\theta,r)\ \cdots\ D_{N-1}(k;\theta,r)]^{\mathrm{T}},$$
$$\boldsymbol{V}(k,\ell) = [V_0(k,\ell)\ V_1(k,\ell)\ \ldots\ V_{N-1}(k,\ell)]^{\mathrm{T}}.$$

The complex vector $\boldsymbol{D}(k;\theta,r)$ is called the array steering vector or the array manifold (Trees, 2002) and incorporates all the spatial characteristics of the array. The impulse response of every acoustic path, in a non-reverberant environment, can be modeled as an attenuated and delayed Kronecker delta function $d_i(n;\theta,r) = \alpha_i(\theta,r)\delta(n - \tau_i(\theta,r))$, where α_i is the attenuation factor and τ_i is the time delay expressed in number of samples. This delay represents the additional time needed by the source signal to travel to the ith sensor after it has reached the center of the array. In the non-reverberant case the ith element of the array steering vector can be written as $D_i(k;\theta,r) = \alpha_i(\theta,r)\mathrm{e}^{-\mathrm{j}\omega_k\tau_i(\theta,r)}$ (Doclo and Moonen, 2003) with ω_k the

discrete-time angular frequency corresponding to the kth frequency bin.

By using this model our goal is to estimate the source signal $s(n)$ in an optimal sense, given the noisy observations at the microphones' outputs. In this paper we are going to focus on three optimization criteria for speech enhancement. These are the most commonly used and have been proved to lead to estimators that can be decomposed into a MVDR beamformer followed by a single channel post-filter. The examined estimators are the minimum mean square error (MMSE) estimator, the MMSE short-time spectral amplitude (MMSE–STSA) estimator and the MMSE short-time log-spectral amplitude estimator (MMSE log-STSA).

To derive the above estimators the *a priori* probability density function (pdf) of the speech and the noise Fourier coefficients should be known. Since in practice this is not the case and furthermore their measurement is a complicated and cumbersome task, the following assumptions (Ephraim and Mallah, 1984), motivated by the central limit theorem, are adopted:

(1) The source signal is a gaussian random process with zero mean and power spectrum ϕ_{ss}.
(2) The noise signals are gaussian random processes with zero mean and cross-spectral density matrix $\boldsymbol{\Phi}_{vv}$.
(3) The source signal is uncorrelated with the noise signals and the Fourier coefficients of each process are independent in different frequencies.

With the establishment of the statistical model, we can proceed with the derivation of the aforementioned estimators. However, first we shall give a very brief description of the MVDR beamformer, since as already mentioned it possesses essential role in the derived solutions.

2.1. MVDR beamformer

An approach for estimating the source signal from its noisy instances is to process the vector $\boldsymbol{X}(k,\ell)$ which consists of the noisy observations, with a matrix operation $\boldsymbol{W}^{\mathrm{H}}(k,\ell)$, where $\boldsymbol{W}(k,\ell)$ is a column vector $N\times 1$ and $(\cdot)^{\mathrm{H}}$ denotes Hermitian transpose. This procedure is known as *filter and sum* beamforming (Johnson and Dudgeon, 1993). To obtain an optimal beamformer we have to minimize the power spectrum of the output[1] given by $\phi_{yy}=\boldsymbol{W}^{\mathrm{H}}\boldsymbol{\Phi}_{xx}\boldsymbol{W}$, where $\boldsymbol{\Phi}_{xx}$ is the auto-spectral density matrix of the noisy inputs. In order to avoid the trivial solution, $\boldsymbol{W}=\boldsymbol{0}$, we use the distortionless criterion, $\boldsymbol{W}^{\mathrm{H}}\boldsymbol{D}=1$, which demands that in the absence of noise, the output of the MVDR beamformer must equal with the desired signal.

The weight vector $\boldsymbol{W}^{\mathrm{H}}$ emerging from the solution of this constrained minimization problem, corresponds to the MVDR or superdirective beamformer and is given by (Bitzer and Simmer, 2001; Cox et al., 1987)

$$\boldsymbol{W}^{\mathrm{H}}=\frac{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}}{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}\boldsymbol{D}}. \tag{3}$$

An important property of the MVDR beamformer is that it maximizes the *array gain* $\frac{|\boldsymbol{W}^{\mathrm{H}}\boldsymbol{D}|^2}{\boldsymbol{W}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}\boldsymbol{W}}$ (Cox et al., 1987; Cox et al., 1986), which is a measure of the increase in signal-to-noise ratio (SNR) that is obtained by using an array rather than a single microphone.

2.2. Multichannel MMSE estimator

Since we have assumed that the source and noise signals are vector gaussian random processes, the MMSE estimator reduces to a linear estimator. Next, we derive this estimator under a vector space viewpoint (Kay, 1993).

The optimum weight vector $\boldsymbol{W}_{\mathrm{opt}}$ transforms the input signal vector $\boldsymbol{X}$, which is corrupted by additive noise $\boldsymbol{V}$, into the best MMSE approximation of the source signal S. To find this optimum weight vector, which constitutes the Multichannel Wiener filter, we have to minimize the MSE at the beamformer's output. In the joint time–frequency domain the error at the beamformer's output is defined as $\mathscr{E}=S-\boldsymbol{W}^{\mathrm{H}}\boldsymbol{X}$ and the optimum solution, assuming that matrix $\boldsymbol{\Phi}_{xx}$ is invertible, is given by

$$\boldsymbol{W}_{\mathrm{opt}}=\boldsymbol{\Phi}_{xx}^{-1}\boldsymbol{\Phi}_{xs}, \tag{4}$$

where $\boldsymbol{\Phi}_{xs}$ is the cross-spectral density vector between the source signal and the noisy inputs. Under the assumption that the source signal and the noise are uncorrelated, it has been shown in Simmer et al. (2001) and Trees (2002) that (4) can be further decomposed into a MVDR beamformer followed by a single channel Wiener filter, which operates at the output of the beamformer:

$$\boldsymbol{W}_{\mathrm{opt}}^{\mathrm{H}}=\underbrace{\frac{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}}{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}\boldsymbol{D}}}_{\boldsymbol{W}_{mvdr}^{\mathrm{H}}}\cdot\underbrace{\left(\frac{\phi_{ss}}{\phi_{ss}+\phi_{nn}}\right)}_{\text{Wiener post-filter}}, \tag{5}$$

where ϕ_{nn} is the power spectrum of the noise at the output of the beamformer. We determine ϕ_{nn} as

$$\phi_{nn}=\boldsymbol{W}_{mvdr}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}\boldsymbol{W}_{mvdr}=\left(\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}\boldsymbol{D}\right)^{-1}. \tag{6}$$

From (5) we can easily obtain the MMSE estimator as $\widehat{S}=\boldsymbol{W}_{\mathrm{opt}}^{\mathrm{H}}\boldsymbol{X}$.

2.3. Optimal nonlinear estimators

From a perceptual point of view, the information we get from the phase is insignificant compared to the information obtained from the speech spectral amplitude (Vary, 1985). Thus, it seems more suitable to estimate the speech spectral amplitude instead of the complex spectrum. If we write $S(k,\ell)=A(k,\ell)e^{j\psi(k,\ell)}$ where $A(k,\ell)$ is the short-time spectral amplitude and $\psi(k,\ell)$ is the phase, then the

[1] Without loss of generality we omit the dependency of k and ℓ, for simplicity.

MMSE–STSA estimator for the kth spectral component, is given by the conditional mean (Ephraim and Mallah, 1984):

$$\widehat{A} = E\{A|x_0(\cdot),\ldots,x_{N-1}(\cdot)\}, \tag{7}$$

where $E\{\cdot\}$ denotes statistical expectation. Since $\{x_0(\cdot),\ldots,x_{N-1}(\cdot)\}$ and $\{X_0(\cdot),\ldots,X_{N-1}(\cdot)\}$ are equivalent representations, and furthermore the Fourier coefficients of each process are uncorrelated at different frequencies, i.e. $X_i(k_1)$ is independent of $X_j(k_2)$ for $k_1 \neq k_2$, (7) can be rewritten as

$$\begin{aligned}\widehat{A} &= E\{A|\{X_1,\ldots,X_{N-1}\} = \boldsymbol{X}\} \\ &= \int_0^\infty A\left(\int_0^{2\pi} p(A,\psi|\boldsymbol{X})\mathrm{d}\psi\right)\mathrm{d}A,\end{aligned} \tag{8}$$

where $p(A,\psi)$ is the joint probability of the amplitude and phase signals.

In a similar way to the MMSE–STSA, the MMSE log-STSA minimizes the mean square error of the log-spectral amplitude. In fact this distortion measure according to (Ephraim and Mallah, 1985) seems more meaningful. For this case the estimator is given by the following conditional mean

$$\widehat{A}_{\log} = \exp(E\{\ln(A)|\boldsymbol{X}\}). \tag{9}$$

The assumed gaussian statistical model leads to Rayleigh distributed joint probability

$$p(A,\psi) = \frac{A}{\pi\phi_{ss}}\exp\left(-\frac{A^2}{\phi_{ss}}\right). \tag{10}$$

Moreover the conditional pdf $p(\boldsymbol{X}|A,\psi)$ is given by

$$p(\boldsymbol{X}|A,\psi) = \frac{1}{\pi^N \det(\boldsymbol{\Phi}_{vv})}\exp(-(\boldsymbol{X}^{\mathrm{H}} - S^*\boldsymbol{D}^{\mathrm{H}})\boldsymbol{\Phi}_{vv}^{-1}(\boldsymbol{X} - \boldsymbol{D}S)). \tag{11}$$

This conditional pdf can be factorized into the product of two functions as

$$p(\boldsymbol{X}|A,\psi) = g(A,T(\boldsymbol{X}))h(\boldsymbol{X}), \tag{12}$$

where g depends only on A and $T(\boldsymbol{X})$, h depends only on the matrix $\boldsymbol{X}$ of the noisy observations and $T(\boldsymbol{X})$ is the output of the MVDR beamformer

$$T(\boldsymbol{X}) = \frac{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}\boldsymbol{X}}{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}\boldsymbol{D}} = \boldsymbol{W}_{mvdr}^{\mathrm{H}}\boldsymbol{X}. \tag{13}$$

According to the Factorization Theorem (Poor, 1998) $T(\boldsymbol{X})$ turns out to be sufficient statistics for A. Moreover, the authors in Balan and Rosca (2002) state that $T(\boldsymbol{X})$ is sufficient statistics for S and any function of S, $\rho(S)$. The above lead to the conclusion that for any prior pdf of S, the conditional pdf of S or of a function $\rho(S)$ with respect to the noise observations $\boldsymbol{X}$, is equivalent with the conditional pdf with respect to $T(\boldsymbol{X})$:

$$p(\rho(S)|\boldsymbol{X}) = p(\rho(S)|T(\boldsymbol{X})). \tag{14}$$

Having this equivalence in mind, it is straightforward to prove that the conditional mean of $\rho(S)$ with respect to $\boldsymbol{X}$ reduces to (Balan and Rosca, 2002):

$$E\{\rho(S)|\boldsymbol{X}\} = E\{\rho(S)|T(\boldsymbol{X})\}. \tag{15}$$

The above result is of great importance and will be used for the derivation of the MMSE–STSA and MMSE log-STSA estimators.

2.3.1. Multichannel MMSE–STSA estimator

To derive the MMSE–STSA estimator we use (15) for the case of $\rho(S) = A$ obtaining

$$\widehat{A} = E\{A|Y = T(\boldsymbol{X})\}, \tag{16}$$

that is we have to estimate the conditional mean of the spectral amplitude with respect to the output of the MVDR beamformer. Recalling that the MVDR beamformer satisfies the distortionless criterion, we will have at its single channel output

$$Y = S + \frac{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}\boldsymbol{V}}{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}\boldsymbol{D}}. \tag{17}$$

The closed form expression of (16) can be obtained (Ephraim and Mallah, 1984) as

$$\widehat{A} = G(u)R, \tag{18}$$

$$G(u) = \Gamma(1.5)\frac{\sqrt{u}}{\gamma}\exp\left(-\frac{u}{2}\right)\left[(1+u)I_0\left(\frac{u}{2}\right) + uI_1\left(\frac{u}{2}\right)\right], \tag{19}$$

where R is the spectral amplitude of Y, $Y(k,\ell) = R(k,\ell)e^{j\vartheta(k,\ell)}$, Γ is the gamma function and I_0, I_1 are the modified Bessel functions of zero and first order respectively. The variable u is defined as

$$u = \frac{\xi}{1+\xi}\cdot\gamma, \tag{20}$$

where ξ and γ are known as *a priori* and *a posteriori* SNR, respectively and are defined as

$$\xi = \frac{\phi_{ss}}{\phi_{nn}}, \quad \gamma = \frac{R^2}{\phi_{nn}}. \tag{21}$$

Since we have estimated the spectral amplitude $\widehat{A}$, we can now use the phase of the noisy MVDR output to obtain the enhanced speech signal as $\widehat{S} = \widehat{A}e^{j\vartheta}$. The whole procedure is equivalent to first processing the noisy observations with the MVDR beamformer and then applying to the single channel output Y, a post-filter with transfer function $G(u)$ given by (19).

2.3.2. Multichannel MMSE log-STSA estimator

For the derivation of the MMSE log-STSA estimator we use once again (15) for the case of $\rho(S) = \ln(A)$ obtaining

$$\widehat{A}_{\log} = E\{\ln(A)|Y = T(\boldsymbol{X})\}, \tag{22}$$

i.e. we have to estimate the conditional mean of the log-spectral amplitude with respect to the output of the MVDR

beamformer. In this case the closed form expression of (22) can be obtained (Ephraim and Mallah, 1985) as

$$\widehat{A}_{\log} = G_{\log}(u)R, \tag{23}$$

$$G_{\log}(u) = \frac{\xi}{1+\xi}\exp\left(\frac{1}{2}\int_u^{\infty}\frac{\mathrm{e}^{-t}}{t}\mathrm{d}t\right), \tag{24}$$

where R is the spectral amplitude of Y (17) and ξ and γ are defined in (21). Once again, we can consider that the enhanced speech signal $\widehat{S}$ is obtained by processing the noisy observations $\boldsymbol{X}$ with the MVDR beamformer and then applying to the single channel output a post-filter with the transfer function provided in (24).

3. Post-filter estimation

In the case of the MVDR beamformer the weight vector $\boldsymbol{W}_{mvdr}^{\mathrm{H}}$ in (3) can be evaluated since it is data independent. In fact, even if there is no prior knowledge of the noise cross-spectral density matrix $\boldsymbol{\Phi}_{vv}$, we can prove that there exists a solution depending only on the auto-spectral density matrix of the noisy observations $\boldsymbol{\Phi}_{xx}$. Noting that $\boldsymbol{\Phi}_{xx}$ can be written as $\boldsymbol{\Phi}_{xx} = \phi_{ss}\mathbf{D}\boldsymbol{D}^{\mathrm{H}} + \boldsymbol{\Phi}_{vv}$, under the assumption that speech and noise are independent, and using the Matrix Inversion lemma (Kay, 1993) we can express $\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{xx}^{-1}$ as

$$\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{xx}^{-1} = \frac{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{vv}^{-1}}{1+(\phi_{ss}/\phi_{nn})}. \tag{25}$$

Then it is trivial to show that the following equality holds:

$$\boldsymbol{W}_{mvdr}^{\mathrm{H}} = \frac{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{xx}^{-1}}{\boldsymbol{D}^{\mathrm{H}}\boldsymbol{\Phi}_{xx}^{-1}\boldsymbol{D}}. \tag{26}$$

On the contrary, from an inspection on (5), (19) and (24) we can see that it is required first to estimate the quantities ϕ_{ss} and ϕ_{nn} in order to derive the studied post-filters. For the estimation of the above quantities we propose later a novel estimation method using the complex coherence function (Elko, 2001).

3.1. Noise field analysis

In microphone array applications, noise fields can be classified according to the degree of correlation between noise signals at different spatial locations. A common measure that is used to characterize a noise field is the *complex coherence function*. The coherence function between two signals x_i and x_j, located at discrete locations, is equal to the cross-power spectrum $\phi_{x_ix_j}$ of these two processes normalized by the square root of the product of the auto-power spectrums $\phi_{x_ix_i}$ and $\phi_{x_jx_j}$ (Elko, 2001):

$$C_{x_ix_j}(\omega) = \frac{\phi_{x_ix_j}(\omega)}{\sqrt{\phi_{x_ix_i}(\omega)\phi_{x_jx_j}(\omega)}}. \tag{27}$$

The coherence is a normalized cross-spectral density function; in particular, the normalization constrains (27) so that the magnitude-squared coherence lies in the range $0 \leqslant |C_{x_ix_j}|^2 \leqslant 1$.

In a diffuse or spherically isotropic noise field, noise of equal energy propagates in all directions simultaneously. The sensors of a microphone array will receive noise signals that are mainly correlated at low frequencies but have approximately the same energy. Diffuse noise field can serve as a model for many applications concerning noisy environments, e.g. cars and offices (Meyer and Simmer, 1997; McCowan and Bourlard, 2003). The complex coherence function for such a noise field can be approximated by (Elko, 2001)

$$C_{v_iv_j}(\omega) = \frac{\sin(\omega f_s r/c)}{\omega f_s r/c} \quad \forall\omega, \tag{28}$$

where $v_{i,j}$ stand for the noise in sensors i and j, r is the distance among the sensors, c is the velocity of sound and ω is the discrete-time angular frequency. For the experiments in this paper the assumption of a diffuse noise field will be considered.

3.2. Generalized estimation approach

In the current section we propose a novel estimation method for the derivation of the studied post-filters, which is appropriate for a variety of different noise fields and optimal for all the discussed minimization criteria (i.e. MSE, MSE-STSA, MSE log-STSA). An overview of the overall multichannel-based noise reduction system is shown in Fig. 1. Note that the various cases (different minimization criteria) differ with respect to the kind of the post-filter used at the output of the MVDR beamformer. In particular, the overall estimator includes the following stages:

(1) The multichannel input signals are fed into a time alignment module. The outputs of this module are the scaled and aligned inputs to account for the effects of propagation. The output signals can be

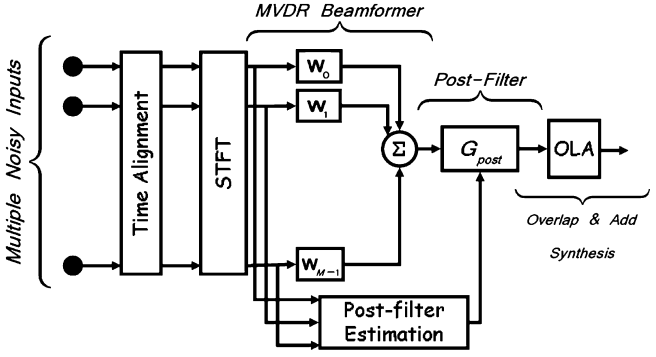

Fig. 1. Multichannel speech enhancement system with post-filter.

denoted in matrix form as $\boldsymbol{X}' = \boldsymbol{I} \cdot S + \boldsymbol{V}'$, with $\boldsymbol{I} = [1, \ldots, 1]^T N \times 1$ column vector.[2]

(2) The multichannel noisy observations are projected to a single channel output Y (17) with minimum noise variance, through the MVDR beamformer.
(3) One of the examined post-filters, according to the utilized criterion, is applied to the output Y.

3.2.1. Source signal spectral estimation

Under the adopted assumptions and the additional hypothesis of a homogeneous noise field, i.e. the noise power spectrum is the same on all sensors ($\phi_{v_i v_i} = \phi_{vv} \forall i$), the computation of the auto- and cross-power spectrums of the time aligned input signals on sensors i and j, results to

$$\phi_{x_i x_j} = \phi_{ss} + \phi_{v_i v_j}, \tag{29}$$

$$\phi_{x_i x_i} = \phi_{ss} + \phi_{vv}. \tag{30}$$

If we have available an estimation of the coherence function then immediately emerges, by replacing in (27) x_i and x_j with v_i and v_j, respectively, that the noise cross-spectral density $\phi_{v_i v_j}$ is given by

$$\phi_{v_i v_j} = \phi_{vv} C_{v_i v_j}. \tag{31}$$

Eqs. (29)–(31) form a 3×3 linear system. By noting that $\phi_{x_i x_i} = \phi_{x_j x_j}$ and solving for ϕ_{ss} we obtain:

$$\hat{\phi}_{ss}^{ij} = \frac{\mathrm{Re}\{\hat{\phi}_{x_i x_j}\} - \frac{1}{2}(\hat{\phi}_{x_i x_i} + \hat{\phi}_{x_j x_j})\mathrm{Re}\{\hat{C}_{v_i v_j}\}}{1 - \mathrm{Re}\{\widehat{C}_{v_i v_j}\}}, \tag{32}$$

which is the derived estimation of ϕ_{ss} using the auto- and cross-spectral densities between sensors i and j. The notation $\hat{(\cdot)}$ stands for the estimated quantity. The average between the auto-power spectrums of channels i and j improves robustness. The use of the real operator $\mathrm{Re}\{\cdot\}$ is justified by the fact that the power spectrum is by definition real. Robustness of the estimation is further improved by taking the average over all $\binom{N}{2}$ possible combinations of channels i and j, resulting in

$$\hat{\phi}_{ss} = \frac{2}{N(N-1)} \sum_{i=0}^{N-2} \sum_{j=i+1}^{N-1} \hat{\phi}_{ss}^{ij}. \tag{33}$$

This result was first derived in McCowan and Bourlard (2003) for the estimation of the Wiener post-filter numerator (5) but is also a part of our extended method which generalizes to all the minimization criteria. The authors in McCowan and Bourlard (2003), in order to obtain the overall transfer function, estimated the denominator $\phi_{ss} + \phi_{nn}$ (5) as the average of the sum of the N auto-power spectrums $\phi_{x_i x_i}$:

$$\phi_{ss} + \phi_{nn} = \sum_{i=0}^{N-1} \phi_{x_i x_i}. \tag{34}$$

[2] In the following we will use $\boldsymbol{X}$ and refer to these aligned signal versions.

This estimation approach leads to a sub-optimal solution (McCowan and Bourlard, 2003; Fischer and Kammeyer, 1997), since it over-estimates the noise power spectrum at the output of the MVDR beamformer. This is attributed to the fact that the noise attenuation already provided by the beamformer is not taken into account.

3.2.2. Noise spectral estimation

We propose a more accurate method for the estimation of ϕ_{nn} which leads to the optimal solution. Furthermore, with the proposed method, in contrast to (McCowan and Bourlard, 2003), we obtain a separate estimation of the noise power spectral density at the output of the beamformer, ϕ_{nn}, which can also be used for the derivation of the nonlinear post-filter transfer functions provided in (19) and (24).

Under the assumption of a homogeneous noise field and employing (6), ϕ_{nn} can be written as

$$\phi_{nn} = \phi_{vv} \boldsymbol{W}_{mvdr}^{\mathrm{H}} \boldsymbol{C}_{vv} \boldsymbol{W}_{mvdr} = \frac{\phi_{vv}}{\boldsymbol{D}^{\mathrm{H}} \boldsymbol{C}_{vv}^{-1} \boldsymbol{D}}, \tag{35}$$

where $\boldsymbol{C}_{vv}$ is the coherence matrix of the noise field defined as

$$\boldsymbol{C}_{vv} = \begin{pmatrix} 1 & C_{v_0 v_1} & \ldots & C_{v_0 v_{N-1}} \\ C_{v_1 v_0} & 1 & & \\ \vdots & & \ddots & \\ C_{v_{N-1} v_0} & \ldots & & 1 \end{pmatrix}. \tag{36}$$

Thus, in order to estimate ϕ_{nn} we need only to estimate ϕ_{vv}. Solving the system of Eqs. (29)–(31) for ϕ_{vv}, results in

$$\hat{\phi}_{vv}^{ij} = \frac{\frac{1}{2}(\hat{\phi}_{x_i x_i} + \hat{\phi}_{x_j x_j}) - \mathrm{Re}\{\hat{\phi}_{x_i x_j}\}}{1 - \mathrm{Re}\{\widehat{C}_{v_i v_j}\}}, \tag{37}$$

which is the estimation of ϕ_{vv} using the auto- and cross-spectral densities between sensors i and j. Using a similar rational with ϕ_{ss}, improved robustness is achieved by taking the average of the auto-power spectrums between channels i and j and by averaging over all combinations of channels:

$$\hat{\phi}_{vv} = \frac{2}{N(N-1)} \sum_{i=0}^{N-2} \sum_{j=i+1}^{N-1} \hat{\phi}_{vv}^{ij}. \tag{38}$$

It should be noted that the estimation of $\hat{\phi}_{ss}^{ij}$ (32) and $\hat{\phi}_{vv}^{ij}$ (37) leads to an indeterminate solution in the case that $\widehat{C}_{v_i v_j} = 1$, for all $i \neq j$. A simple approach to avoid this problem is to bound the model of the coherence function so as $\widehat{C}_{v_i v_j} < 1$, for all $i \neq j$.

An alternative approach only for the estimation of the Wiener post-filter denominator $\phi_{ss} + \phi_{nn}$ (5), is to estimate the power spectrum ϕ_{yy}, directly from the output of the MVDR beamformer. However, in such case the estimation lacks robustness since we have available only one output signal to make the estimation, instead of the N signals we use in our approach.

For practical purposes, one can cope with the deficiency of the MVDR to remove sufficiently the noise for low frequencies, by using instead of ϕ_{nn} a modified version expressed as

$$\phi_{nn} = \begin{cases} \phi_{vv} & \text{for } \omega \leqslant \omega_1, \\ \phi_{nn} & \text{for } \omega > \omega_1, \end{cases}$$

where ω_1 sets the bound for the low frequency region. Once we have estimated the quantities ϕ_{ss} and ϕ_{nn} the derivation of the discussed post-filters provided in (5), (19) and (24) can be accomplished in a straightforward manner.

4. Experiments and results

To validate the effectiveness of the proposed post-filter estimation method, we compare its performance to other multichannel noise reduction techniques, including the MVDR beamformer (Bitzer and Simmer, 2001), the generalized Zelinski post-filter (Zelinski, 1988) and the McCowan post-filter (McCowan and Bourlard, 2003), under the assumption of a diffuse noise field. In addition, we provide comparisons with the noise reduction results obtained by using at the output of the MVDR beamformer the "decision directed" estimation approach (Ephraim and Mallah, 1984). This is a single channel method used to estimate the transfer function of the post-filter.

4.1. Speech corpus and system realization

The microphone data set used for the experiments is the CMU microphone array database (Sullivan, 1996). The recordings were collected in a computer lab by a linear microphone array with eight sensors spaced 7 cm apart, at a sampling rate of 16 kHz. The array was placed on a desk and the speaker was seated directly in front of it at a distance of 1 m from its center. For each array recording there exists a corresponding clean control recording. The room had multiple noise sources, including several computer fans and overhead air blowers. These noise conditions can be effectively modeled by a diffuse noise field. The reverberation time of the room was measured to be 240 ms and the average SNR of the recordings is 6.5 dB. The corpus consists of 130 utterances, 10 speakers of 13 utterances each.

The time aligned noisy input microphone signals are divided in time into frames of 400 samples (25 ms) with overlap of 300 samples (≈19 ms) between adjacent frames. At each frame a Hamming window is applied and a STFT analysis takes place. Afterwards, the transformed inputs are fed into the MVDR beamformer. In order to overcome the gain and phase errors of the microphones and the problem of the self-noise, the weight vector of the MVDR beamformer is computed under a *white noise gain* constraint (Cox et al., 1986). The post-filter transfer function of each studied method is derived by applying as inputs in the noise reduction system (see Fig. 1), the noisy speech signals. The auto- and cross-spectral densities $\phi_{x_i x_i}$ and $\phi_{x_i x_j}$ are computed using the short-time spectral estimation method proposed in Allen et al. (1977):

$$\hat{\phi}_{x_i x_j}(k,\ell) = \alpha \hat{\phi}_{x_i x_j}(k,\ell-1) + (1-\alpha) x_i(k,\ell) x_j^*(k,\ell), \qquad (39)$$

which can be viewed as a recursive Welch periodogram; this method yields smoother spectra and improved estimates. The term α in (39) is a number close to unity and $*$ denotes conjugate. Finally, the enhanced output of the post-filter is transformed back to the time-domain using the overlap and add synthesis (OLA) method (Rabiner and Schafer, 1978).

4.2. Speech enhancement experiments

In order to compare the proposed post-filtering approach with the other multichannel reduction methods and the single-channel "decision directed" estimation method, we use five different objective speech quality measures. To evaluate the noise reduction we use the segmental signal-to-noise ratio enhancement (SSNRE). This is the dB difference between the segmental SNRs of the enhanced output and the noisy inputs average. The segmental SNR

Table 1
Speech quality results from speech enhancement experiments on the CMU database

	SSNRE (dB)	IS	LAR	LLR (dB)	LSD (dB)
Noisy input	–	1.973	8.314	6.920	8.341
MVDR	0.024	1.260	10.969	6.409	6.794
Zelinski	0.097	5.950	13.774	9.912	6.166
McCowan	5.707	3.279	6.764	7.156	3.775
MMSEdd[a]	3.071	13.518	8.575	8.666	4.921
STSAdd	2.137	6.984	9.439	6.936	5.519
Log-STSAdd	2.621	11.559	9.392	7.417	5.135
MMSE	6.361	0.988	4.425	4.742	3.511
STSA	6.221	0.992	4.431	4.727	3.525
Log-STSA	6.320	0.989	4.425	4.733	3.512

[a] Suffix "dd" refers to the "decision directed" method.

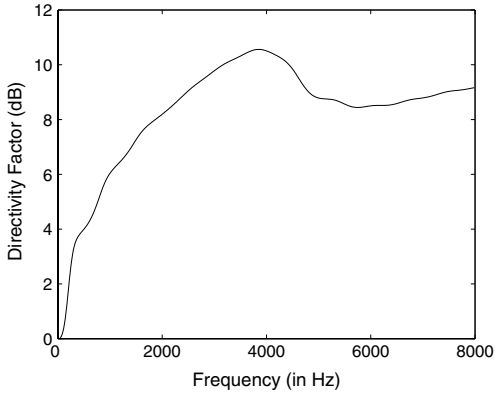

Fig. 2. MVDR beamformer directivity factor that describes the ability of the beamformer to suppress the noise field. For the low frequency region it shows a low gain.

is defined in Hansen and Pellom (1998) and is a more appropriate performance criterion for speech enhancement than the standard SNR. Since, frames with SNRs above 35 dB do not contribute significantly to the overall speech quality and frames consisting of silence can have SNRs with extreme negative values, that do not reflect the percep-

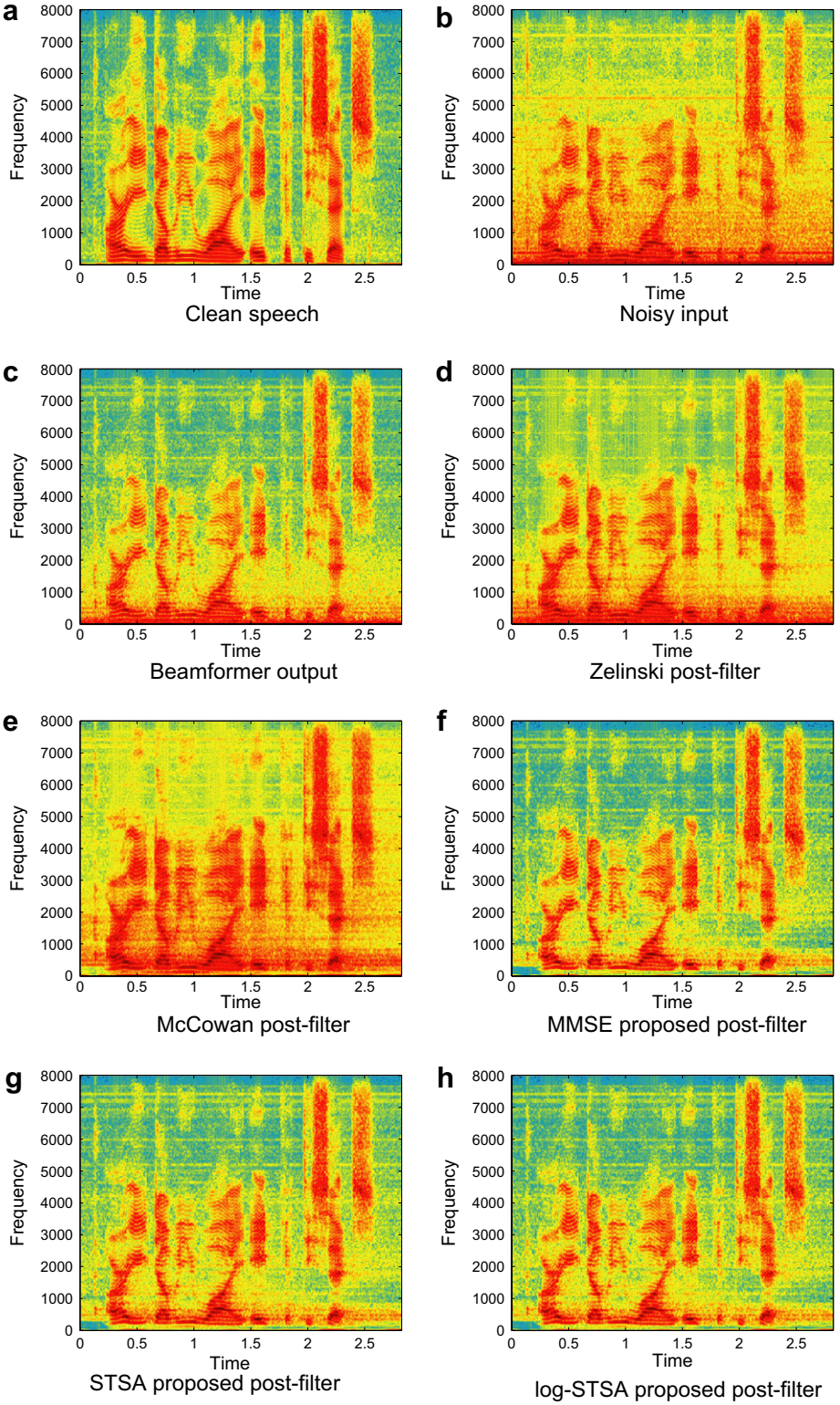

Fig. 3. Speech spectrograms for an utterance "r-e-w-y-8-56". (a) Original clean speech. (b) Noisy signal at central sensor (IS = 1.44). (c) Beamformer output (SSNRE = 0.02 dB, IS = 0.90). (d) Zelinski post-filter (SSNRE = 0.17 dB, IS = 2.89). (e) McCowan post-filter (SSNRE = 3.95 dB, IS = 2.08). (f) MMSE (SSNRE = 4.54 dB, IS = 0.81). (g) STSA (SSNRE = 4.46 dB, IS = 0.82). (h) log-STSA (SSNRE = 4.52 dB, IS = 0.81).

tual contribution of the signal, the SNR at each frame is limited to the range of $(-10, 35)$ dB. To assess the speech quality of the enhanced output signal we use the log-area-ratio distance (LAR), the log-likelihood ratio (LLR), the Itakura–Saito distortion (IS) (Hansen and Pellom, 1998) and the log-spectral distance (LSD) (Cohen, 2004). These measures are found to have a high correlation with the human perception. Low values of the above four quality measures denote high speech quality.

The SSNRE, LAR, LLR, IS and LSD results, averaged across the entire database, are shown in Table 1, for all the studied enhancement algorithms and the noisy input at the central sensor of the microphone array. With the suffix "dd" are the results obtained using the "decision directed" method. In the last three rows of Table 1 the objective speech quality results for the post-filters, estimated with the proposed method, are demonstrated. In addition, in Fig. 3 typical speech spectrograms are presented for comparison between the clean signal, the central noisy input and the output signals of the studied multichannel methods.

From both the table results and the speech spectrograms it can be clearly seen that neither the beamformer alone nor the Zelinski post-filter can provide sufficient noise reduction compared to the other four multichannel methods and the single channel "decision directed" approach. Specifically, from Fig. 3c and d we note that these two methods are incapable of removing the noise in the low frequency region. For the MVDR beamformer this inadequacy 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, as shown in Fig. 2. The poor performance of the Zelinski post-filter is expected since this method is based on the assumption of a spatially uncorrelated noise field, which leads to an inappropriate model for the noise conditions. By making the global assumption that for all frequencies the noise is uncorrelated among the channels, Zelinski post-filter improves the noise reduction for mid and high frequencies but has no effect at low frequencies where the correlation is significant. An additional explanation is provided in Fischer and Simmer (1996), where it is shown that Zelinski's method, can have an affordable performance only for reverberation times above 300 ms. For very low reverberation times, the output speech quality is found to be poorer than the input speech quality. On the other hand, McCowan post-filter performs better than the previous two methods, since the estimation of the source signal spectrum is performed using the correlation of the noise among the different channels. Still its performance is inferior to the post-filters derived by the proposed method, for the reasons we have already discussed. Finally, with the "decision directed" method the noise reduction is greater than the one provided by the first two methods, but at the cost of poor speech quality due to musical noise.

From the provided results, it is evident that the proposed enhancement algorithms outperform the other examined techniques, since they consistently produce better results for all the objective measures in the given database (Sullivan, 1996). Moreover, it can also be seen from Fig. 3a–h that the spectrograms closest to the clean speech are those derived by applying the post-filters estimated by the proposed approach. This is justified by the fact that the proposed post-filters, due to the accurate estimation of the noise spectral density, perform a sufficient noise reduction on every frequency region (low-mid-high) while still providing the highest speech quality signal with no further distortion. Furthermore, the similar, improved results obtained under the different criteria (MSE, MSE-STSA, MSE log-STSA), imply the simultaneous satisfaction of all three. This intuitively motivates the use of the proposed scheme as a general and possibly optimum estimation approach.

In a different direction, a by-product of some previous multichannel speech enhancement works was also to investigate possible improvements in automatic speech recognition (ASR) performance. Clearly, dealing with the ASR problem is by itself a very broad topic which goes far beyond the scope of this paper. Our main focus and effort in this paper was placed on how to give an analysis and provide an optimum estimation method that can be used for the realizations of the linear and nonlinear post-filters, derived under various speech enhancement criteria. However, in a previous work (Leukimmiatis et al., 2006), we had obtained some preliminary ASR results to test how our method behaves with respect to other multichannel approaches. These experiments considered only the case where we estimate the post-filter under the minimization of the MSE criterion. The derived results seemed quite promising and motivated us for further research in multichannel robust feature extraction.

5. Conclusions

In this paper we have presented a multichannel post-filtering estimation approach that is appropriate for a variety of different noise conditions and can be applied for the derivation of both linear and nonlinear post-filters. For the case of the MSE speech enhancement criterion, the proposed method is an improvement of the existing McCowan post-filter, since it produces a robust and more accurate estimation of the noise power spectrum at the beamformer output, which satisfies the MMSE optimality of the Wiener post-filter. In contrast to McCowan method the proposed technique is also applicable to post-filters satisfying other enhancement criteria than MSE.

In experiments with real noise multichannel recordings from the CMU database (Sullivan, 1996), the proposed technique obtained a significant gain over established reference methods as it consistently improved the enhancement performance in terms of five objective speech quality measures. Namely the relative % average improvements achieved compared to the best of the reference approaches were 11.5% in segmental SNR, 21.6% in Itakura–Saito

distortion, 34.5% in log area ratio, 26.2% in log-likelihood ratio and 7% in log spectral distance. Apart from the quantitative evaluation, both auditory and visual inspection of the speech waveforms and spectrograms verified the potential of the generalized estimation as a robust, multichannel enhancement approach.

Acknowledgement

The authors would like to thank G. Evangelopoulos and V. Pitsikalis for their helpful comments during the writing of this paper.

References

Allen, J.B., Berkley, D.A., Blauert, J., 1977. Multimicrophone signal-processing technique to remove room reverberation from speech signals. J. Acoust. Soc. Amer. 62 (4), 912–915.

Balan, R., Rosca, J., 2002. Microphone array speech enhancement by bayesian estimation of spectral amplitude and phase. In: Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop, pp. 209–213.

Bitzer, J., Simmer, K.U., 2001. Superdirective microphone arrays. In: Brandstein, M., Ward, D. (Eds.), Microphone Arrays: Signal Processing Techniques and Applications. Springer Verlag, pp. 19–38 (Chapter 2).

Cohen, I., 2004. Multichannel post-filtering in nonstationary noise environments. IEEE Trans. Signal Process. 52 (5), 1149–1160.

Cohen, I., Berdugo, B., 2002. Microphone array post-filtering for nonstationary noise suppression, In: International Conference on Acoustiscs, Speech, Signal Processing (ICASSP), Vol. 1. pp. 901–904.

Cox, H., Zeskind, R.M., Kooij, T., 1986. Practical supergain. IEEE Trans. Speech Audio Process. 34 (3), 393–398.

Cox, H., Zeskind, R.M., Owen, M.W., 1987. Robust adaptive beamforming. IEEE Trans. Speech Audio Process. 35 (10), 1365–1376.

Doclo, S., Moonen, M., 2003. Design of far-field and near-field broadband beamformers using eigenfilters. Speech Commun. 83, 2641–2672.

Elko, G.W., 2001. Spatial coherence function for differential microphones in isotropic noise fields. In: Brandstein, M., Ward, D. (Eds.), Microphone Arrays: Signal Processing Techniques and Applications. Springer Verlag, pp. 61–85 (Chapter 4).

Ephraim, Y., Mallah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32 (6), 1109–1121.

Ephraim, Y., Mallah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33 (2), 443–445.

Fischer, S., Kammeyer, D., 1997. Broadband beamforming with adaptive postfiltering for speech acquisition in noisy environments, In: International Conference on Acoustics, Speech, Signal Processing (ICASSP), Vol. 1, pp. 359–362.

Fischer, S., Simmer, K.U., 1996. Beamforming microphone arrays for speech acquisition in noisy environments. Speech Commun. 20, 215–227.

Hansen, J.H.L., Pellom, B.L. 1998. An effective quality evaluation protocol for speech enhancement algorithms. In: International Conference on Spoken Language Processing (ICSLP), pp. 2819–2822.

Johnson, D.H., Dudgeon, D.E., 1993. Array Signal Processing: Concepts and Techniques. Prentice Hall.

Kay, S.M., 1993. Fundamentals of Statistical Signal Processing: Estimation Theory. Prentice Hall.

Leukimmiatis, S., Dimitriadis, D., Maragos, P., 2006. An optimum microphone array post-filter for speech applications. In: Proceedings of the Interspeech–Eurospeech, pp. 2142–2145.

Marro, C., Mahieux, Y., Simmer, K.U., 1988. Analysis of noise reduction techniques based on microphone arrays with postfiltering. IEEE Trans. Speech Audio Process. 6 (3), 240–259.

McCowan, I.A., Bourlard, H., 2003. Microphone array post-filter based on noise field coherence. IEEE Trans. Speech Audio Process. 11 (6), 709–716.

Meyer, J., Simmer, K.U., 1997. Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction. In: International Conference on Acoustics, Speech, Signal Processing (ICASSP), Vol. 2. pp. 1167–1170.

Poor, H.V., 1998. An Introduction to Signal Detection and Estimation. Springer Verlag.

Rabiner, L.R., Schafer, R.W., 1978. Digital Signal Processing of Speech Signals. Prentice Hall.

Simmer, K.U., Bitzer, J., Marro, C., 2001. Post-filtering techniques. In: Brandstein, M., Ward, D. (Eds.), Microphone Arrays: Signal Processing Techniques and Applications. Springer Verlag, pp. 39–60 (Chapter 3).

Sullivan, T., 1996. CMU microphone array database. <http://www.speech.cs.cmu.edu/databases/micarray>.

Trees, H.L.V., 2002. Optimum Array Processing. Wiley.

Vary, P., 1985. Noise suppression by spectral magnitude estimation – mechanism and theoritical limits. Signal Process. 8 (4), 387–400.

Veen, B.D.V., Buckley, K.M., 1988. Beamforming: A versatile approach to spatial filtering. IEEE ASSP Mag. 5, 4–24.

Zelinski, R., 1988. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In: International Conference on Acoustics, Speech, Signal Processing (ICASSP), Vol. 5. pp. 2578–2581.