Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 2006-73

A unification of adaptive multi-microphone noise reduction systems [1]

Ann Spriet[2], Simon Doclo [3], Marc Moonen[4] ,Jan Wouters[5]

To appear inProc. of the 2006 International Workshop on Acoustic Echo and Noise Control (IWAENC 2006), Paris, France, 2006

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.be* in the directory *pub/sista/spriet/reports/06-73.pdf*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium, Tel. 32/16/32 17 95, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.be/scd*. E-mail: *ann.spriet@esat.kuleuven.be*. K.U.Leuven, ExpORL/ O&N2, Herestraat 49 bus 721, 3000 Leuven, Belgium, WWW: *http://www.kuleuven.be/exporl/Lab/Default.htm*. Ann Spriet is a postdoctoral researcher funded by F.W.O.-Vlaanderen. This research was carried out at the ESAT laboratory and the ExpORL laboratory of K.U. Leuven, in the frame of IUAP P5/22 (2002-2007), the Concerted Research Action GOA-AMBioRICS, the K.U. Leuven Research Council CoE EF/05/006, FWO Projects nr. G.0504.04 and G.0334.06, IWT project 020540.

[3]K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium, WWW: *http://www.esat.kuleuven.be/scd*. E-mail: *simon.doclo@esat.kuleuven.be*. Simon Doclo is a postdoctoral researcher funded by F.W.O.-Vlaanderen.

[4]K.U.Leuven, Dept. of Electrical Engineering (ESAT), SISTA, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium, Tel. 32/16/32 17 09, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.be/scd*. E-mail: *marc.moonen@esat.kuleuven.be*. Marc Moonen is a professor at the Katholieke Universiteit Leuven.

[5]K.U.Leuven, ExpORL, O&N2, Herestraat 49 bus 721, 3000 Leuven, Belgium, Tel. 32/16/33 23 42, Fax 32/16/33 23 35, WWW: *http://www.kuleuven.be/exporl/Lab/Default.htm* E-mail: *jan.wouters@med.kuleuven.be*. Jan Wouters is a professor at the Katholieke Universiteit Leuven.

A UNIFICATION OF ADAPTIVE MULTI-MICROPHONE NOISE REDUCTION SYSTEMS

Ann Spriet[1,2]*, Simon Doclo*[1]*,Marc Moonen*[1]*, Jan Wouters*[2]

[1]K.U. Leuven, ESAT/SCD-SISTA
Kasteelpark Arenberg 10, 3001 Leuven, Belgium
{spriet,doclo,moonen}@esat.kuleuven.be

[2]K.U. Leuven - ExpORL, O&N2
Herestraat 49 bus 721, 3000 Leuven, Belgium
jan.wouters@med.kuleuven.be

ABSTRACT

In this paper a general cost function for adaptive multi-microphone noise reduction is proposed. From this cost function, many existing adaptive multi-microphone noise reduction techniques can be derived, such as linearly constrained minimum variance (LCMV) beamforming, transfer-function LCMV, soft-constrained beamforming and speech-distortion weighted multi-channel Wiener filtering as well as combined approaches.

1. INTRODUCTION

In speech communication applications, such as teleconferencing, hearing aids, handsfree telephony, the presence of background noise may seriously degrade the quality and intelligibility of the speech signal. To enhance the speech recordings, several adaptive multi-microphone noise reduction techniques have been proposed in the literature. Two categories of adaptive techniques can be distinguished: adaptive beamforming and multi-channel Wiener filtering based techniques.

Adaptive beamforming techniques typically solve a linearly constrained minimum variance (LCMV) optimization criterion, minimizing the output power subject to the (hard) constraint that signals coming from a certain region or direction (i.e., ideally the direction of the desired speech source) are preserved [1, 2]. The classical LCMV beamformer assumes free-field propagation. To improve performance in the presence of reverberation, an extension to the classical LCMV beamformer that incorporates arbitrary transfer functions, referred to as transfer function LCMV (TF-LCMV), has been suggested [3]. An efficient realization of the LCMV is the Generalized Sidelobe Canceller (GSC) [1, 2].

A second category are multi-channel Wiener filtering (MWF) based techniques such as the speech-distortion weighted MWF (SDW-MWF) [4] and the soft-constrained beamforming techniques [5]. In contrast to adaptive beamforming techniques, these techniques exploit both spectral and spatial differences between the speech and the noise sources, so that inevitably some speech distortion will be introduced.

In this paper, we show that the above mentioned adaptive noise reduction techniques as well as some combinations can be derived from one general cost function, trading off between output noise power and a speech distortion. Basically, the noise reduction techniques differ from each other in the use of an a-priori and/or online estimated speech model and the use of a soft or hard constraint on the amount of speech distortion.

Ann Spriet and Simon Doclo are postdoctoral researchers funded by F.W.O.-Vlaanderen. This research was carried out at the ESAT laboratory and the ExpORL laboratory of K.U. Leuven, in the frame of IUAP P5/22 (2002-2007), the Concerted Research Action GOA-AMBioRICS, the K.U. Leuven Research Council CoE EF/05/006, FWO Projects nr. G.0504.04 and G.0334.06, IWT project 020540.

2. GENERAL COST FUNCTION

2.1. Signal model

Let $X_i(f)$, $i = 1, \ldots, M$ denote the frequency-domain microphone signals[1]

$$X_i(f) = X_i^s(f) + X_i^n(f) \tag{1}$$

and let $\mathbf{X}(f) \in \mathbb{C}^{M\times 1}$ be defined as the stacked vector

$$\begin{aligned} \mathbf{X}(f) &= \begin{bmatrix} X_1(f) & X_2(f) & \cdots & X_M(f) \end{bmatrix}^T & (2)\\ &= \mathbf{X}^s(f) + \mathbf{X}^n(f) & (3)\end{aligned}$$

Defining $H_i^s(f)$ as the acoustic transfer function from the speech source $S(f)$ to the i-th microphone, $\mathbf{X}^s(f)$ can be written as

$$\mathbf{X}^s(f) = \mathbf{H}^s(f)S(f) = \tilde{\mathbf{H}}^s(f)X_1^s(f), \tag{4}$$

with $\tilde{\mathbf{H}}^s(f)$ the vector with transfer function ratios relative to the first microphone

$$\tilde{\mathbf{H}}^s(f) = \mathbf{H}^s(f)/H_1^s(f) = \begin{bmatrix} 1 & \frac{H_2^s(f)}{H_1^s(f)} & \cdots & \frac{H_M^s(f)}{H_1^s(f)} \end{bmatrix}^T. \tag{5}$$

To simplify notation, we define the power spectral density (PSD) of the speech and the noise in the i-th microphone signal as

$$\begin{aligned} P_{X_i}^s(f) &= \varepsilon\{X_i^s(f)X_i^{*,s}(f)\}, & (6)\\ P_{X_i}^n(f) &= \varepsilon\{X_i^n(f)X_i^{*,n}(f)\}. & (7)\end{aligned}$$

In addition, we define the noise and speech correlation matrix as:

$$\begin{aligned} \mathbf{R}^n(f) &= \varepsilon\{\mathbf{X}^n(f)\mathbf{X}^{n,H}(f)\}, & (8)\\ \mathbf{R}^s(f) &= \varepsilon\{\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\} = P_{X_1}^s(f)\tilde{\mathbf{H}}^s(f)\tilde{\mathbf{H}}^{s,H}(f). & (9)\end{aligned}$$

2.2. Free-field propagation model

Single point source

Assuming free-field propagation, the contribution $X_i(f,\mathbf{p})$ of a point source $S(f,\mathbf{p})$ at location $\mathbf{p}$ in the i-th microphone signal (with coordinates $\mathbf{p}_i$) equals

$$X_i(f,\mathbf{p}) = A_i(f,\mathbf{p})a_i(\mathbf{p})e^{-j2\pi f\tau_i(\mathbf{p})}S(f,\mathbf{p}), \tag{10}$$

[1]In the sequel, the superscripts s and n are used to refer to the speech and noise contribution of a signal.

where $A_i(f,\mathbf{p})$ represents the characteristic of the i-th microphone, $a_i(\mathbf{p})$ is the attenuation of the point source $S(f,\mathbf{p})$ at the position of the i-th microphone (near-field effect) and

$$\tau_i(\mathbf{p}) = \frac{\|\mathbf{p}-\mathbf{p}_i\|}{c} \tag{11}$$

with c the speed of sound (340 m/s), is the propagation delay from the point source $S(f,\mathbf{p})$ to the i-th microphone. Defining the first microphone signal $X_1(f,\mathbf{p})$ as reference signal,

$$\mathbf{X}(f,\mathbf{p}) = \tilde{\mathbf{d}}(f,\mathbf{p})X_1(f,\mathbf{p}) \tag{12}$$

where $\tilde{\mathbf{d}}(f,\mathbf{p})$ is the steering vector

$$\tilde{\mathbf{d}}(f,\mathbf{p}) = \begin{bmatrix} 1 \\ \frac{A_2(f,\mathbf{p})}{A_1(f,\mathbf{p})}\frac{a_2(\mathbf{p})}{a_1(\mathbf{p})}e^{-j2\pi f(\tau_2(\mathbf{p})-\tau_1(\mathbf{p}))} \\ \vdots \\ \frac{A_M(f,\mathbf{p})}{A_1(f,\mathbf{p})}\frac{a_M(\mathbf{p})}{a_1(\mathbf{p})}e^{-j2\pi f(\tau_M(\mathbf{p})-\tau_1(\mathbf{p}))} \end{bmatrix}. \tag{13}$$

Multiple point sources

If several point sources $S(f,\mathbf{p})$ at positions $\mathbf{p}\in\mathbf{P}$ are active, the microphone signals $\mathbf{X}(f)$ can be modeled as:

$$\mathbf{X}(f) = \int_{\mathbf{p}\in\mathbf{P}} \tilde{\mathbf{d}}(f,\mathbf{p})X_1(f,\mathbf{p}), \tag{14}$$

with $X_1(f,\mathbf{p})$ defined by (10). For uncorrelated point sources

$$\varepsilon\{X_1(f,\mathbf{p}_k)X_1(f,\mathbf{p}_l)\} = P_{X_1}(f,\mathbf{p}_k)\delta_{kl}. \tag{15}$$

2.3. Multi-microphone noise reduction

In a multi-microphone noise reduction system, the microphone signals $X_i(f)$ are filtered by (adaptive or fixed) filters $W_i(f)$ and combined in order to obtain an enhanced speech signal $Z(f)$. Define

$$\mathbf{W}(f) = \begin{bmatrix} W_1(f) & W_2(f) & \cdots & W_M(f) \end{bmatrix}^H, \tag{16}$$

then the output $Z(f)$ of the multi-channel noise reduction algorithm is

$$Z(f) = \underbrace{\mathbf{W}^H(f)\,\mathbf{X}^s(f)}_{Z^s(f)} + \underbrace{\mathbf{W}^H(f)\mathbf{X}^n(f)}_{Z^n(f)}. \tag{17}$$

The goal of the filter $\mathbf{W}(f)$ is to minimize the output noise power as much as possible without severely distorting the speech signal. The amount of speech distortion is measured with respect to a reference speech signal $D^s(f)$. This reference signal can be the speech component $X_1^s(f)$ in the first microphone, the speech source signal $S(f)$ or the speech component in the output of a fixed beamformer (e.g., the speech reference in the spatially pre-processed SDW-MWF [4]).

2.4. General cost function

A general cost function $J(\mathbf{W}(f))$ for the filter $\mathbf{W}(f)$ is (18) on the following page. The first two terms in $J(\mathbf{W}(f))$ correspond to the output noise energy. This output noise energy can be:

- estimated online (i.e., the term $\mathbf{W}^H\mathbf{R}^n(f)\mathbf{W}(f)$)
- and/or based on a prior knowledge $\mathbf{R}_m^n(f)$ of the noise correlation matrix, which is constructed through calibration measurements or mathematical models.

In this paper, we focus on an online estimated noise model. For extensions with a pre-defined noise model (including fixed beamformers), we refer to [6].
The last two terms in $J(\mathbf{W})$ denote the distortion energy between the output speech component $\mathbf{W}^H(f)\mathbf{X}^s(f)$ (or $\mathbf{W}^H(f)\mathbf{X}_m^s(f)$) and a reference speech signal $D^s(f)$ (or $D_m^s(f)$). Again, the output speech distortion energy may be

- estimated online (i.e., as $\varepsilon\{(D^s(f) - \mathbf{W}^H(f)\mathbf{X}^s(f))(D^s(f) - \mathbf{W}^H(f)\mathbf{X}^s(f))^H\}$)
- and/or based on prior knowledge $\mathbf{X}_m^s(f)$ for the microphone signals (i.e., as $\varepsilon\{(D_m^s(f) - \mathbf{W}^H(f)\mathbf{X}_m^s(f))(D_m^s(f) - \mathbf{W}^H(f)\mathbf{X}_m^s(f))^H\}$). Again, this model can be constructed based on calibration data or based on mathematical models.

Parameters μ_1, μ_2 trade off between speech distortion and noise reduction: the larger μ_1 or μ_2, the more emphasis is put on speech distortion. Depending on the use of prior knowledge of the speech correlation matrix and the use of a hard constraint on the speech distortion term (i.e. $\mu_{1,2}=\infty$ or $\mu_{1,2}\neq\infty$), different adaptive multi-microphone noise reduction techniques can be obtained, as indicated in Table 1. When using a hard constraint (i.e.,$\mu_1=\infty$ or $\mu_2=\infty$), noise suppression is only achieved in the subspace orthogonal to the defined or actual speech subspace. Signals in the (defined or actual) speech subspace are passed through undistorted by the noise reduction algorithm. The use of a soft-constraint ($\mu_1\neq\infty$ or $\mu_2\neq\infty$) typically results in a spectral filtering of the desired speech component $D^s(f)$ since the speech and noise subspace are generally not orthogonal (often, the noise subspace spans the complete space).
In the next sections, the different techniques are explained in more detail.

3. A-PRIORI SPEECH MODEL ($\mu_1 = 0$)

The classical LCMV beamformer [1, 2] and the soft-constrained beamformer [5] exploit a-priori knowledge about the speech statistics. Assumptions are made about the microphones (microphone characteristics, positions), the location of the desired speaker and the room acoustics (e.g., no reverberation). These assumptions are often violated in practice so that the performance may be suboptimal.

3.1. Hard constraint ($\mu_2 = \infty$): LCMV

The LCMV beamformer [1, 2] minimizes the output noise power subject to the constraint that signals coming from a certain location or region of interest are preserved. This corresponds to the cost function (18) with $\mu_2=\infty$ and $\mu_1=0$. Typically, the free-field propagation model (12)-(13) is assumed for the speech signal:

$$\mathbf{X}_m^s(f) = \tilde{\mathbf{d}}^s(f,\mathbf{p}_m^s)X_{m,1}^s(f), \tag{19}$$

where $\mathbf{p}_m^s$ refers to the position of the speech source. The reference signal $D_m^s(f)$ equals $X_{m,1}^s(f)$.
The filter $\mathbf{W}(f)$ equals

$$\left(\mathbf{R}^n(f) + \mu_2 P_{X_1}^s(f)\tilde{\mathbf{d}}^s\tilde{\mathbf{d}}^{s,H}\right)^{-1}\mu_2 P_{X_1}^s(f)\tilde{\mathbf{d}}^s(f,\mathbf{p}^s). \tag{20}$$

$$J(\mathbf{W}(f)) = (1-\lambda)\mathbf{W}^H(f)\mathbf{R}^n(f)\mathbf{W}(f) + \lambda\mathbf{W}^H(f)\mathbf{R}_m^n(f)\mathbf{W}(f) + \mu_1\varepsilon\{(D^s(f)-\mathbf{W}^H(f)\mathbf{X}^s(f))(D^s(f)-\mathbf{W}^H(f)\mathbf{X}^s(f))^H\} + \mu_2\varepsilon\{(D_m^s(f)-\mathbf{W}^H(f)\mathbf{X}_m^s(f))(D_m^s(f)-\mathbf{W}^H(f)\mathbf{X}_m^s(f))^H\}. \quad (18)$$

Speech model	Hard/Soft constraint on speech distortion		Technique
A-priori (Section 3)	$\mu_1 = 0$	$\mu_2 = \infty$	LCMV
	$\mu_1 = 0$	$\mu_2 \neq \infty$	Soft-constrained beamforming
Online (Section 4)	$\mu_1 = \infty$	$\mu_2 = 0$	TF-LCMV
	$\mu_1 \neq \infty$	$\mu_2 = 0$	SDW-MWF
Combination (Section 5)	$\mu_1 \neq \infty$	$\mu_2 = \infty$	SDR-GSC
	$\mu_1 \neq \infty$	$\mu_2 \neq \infty$	Combination SDW-MWF/soft-constrained

Table 1: Classification of adaptive multi-microphone noise reduction techniques.

Applying the matrix inversion lemma

$$\left(\mathbf{R}^n(f) + \mu_2 P_{X_1}^s(f)\tilde{\mathbf{d}}^s(f,\mathbf{p}^s)\tilde{\mathbf{d}}^{s,H}(f,\mathbf{p}^s)\right)^{-1} = \mathbf{R}^{n^{-1}}(f) - \frac{\mathbf{R}^{n^{-1}}(f)\mu_2 P_{X,1}^s(f)\tilde{\mathbf{d}}^s(f,\mathbf{p}^s)\tilde{\mathbf{d}}^{s,H}(f,\mathbf{p}^s)\mathbf{R}^{n^{-1}}(f)}{1+\mu_2 P_{X_1}^s(f)\tilde{\mathbf{d}}^{s,H}(f,\mathbf{p}^s)\mathbf{R}^{n-1}(f)\tilde{\mathbf{d}}^s(f,\mathbf{p}^s)}, \quad (21)$$

and setting $\mu_2 = \infty$, results in

$$\mathbf{W}(f) = \frac{\mathbf{R}^{n^{-1}}(f)\tilde{\mathbf{d}}^s(f,\mathbf{p}^s)}{\tilde{\mathbf{d}}^{s,H}(f,\mathbf{p}^s)\mathbf{R}^{n-1}(f)\tilde{\mathbf{d}}^s(f,\mathbf{p}^s)}. \quad (22)$$

3.2. Soft constraint ($\mu_2 \neq \infty$): soft-constrained beamformer

In [5], MWF techniques are proposed that use a (partially) precomputed speech correlation matrix. These techniques, called soft-constrained beamforming, minimize the output noise power with a soft constraint on a (partially) modelled speech distortion term. This corresponds to (18) with $\mu_2 \neq \infty$ and $\mu_1 = 0$. A fixed model is used for the spatial characteristics $\tilde{\mathbf{H}}^s(f)$ of the speech while the speech PSD $P_{X_1}^s(f)$ is estimated online. The speech source is modeled as an infinite number of (uncorrelated) point sources with true PSD $P_{X_1}^s(f)$ clustered closely in space within a pre-defined area $\mathbf{P}$:

$$\mathbf{X}_m^s(f) = \int_{\mathbf{p}\in\mathbf{P}} X_{m,1}^s(f,\mathbf{p})\tilde{\mathbf{d}}^s(f,\mathbf{p})d\mathbf{p} \quad (23)$$

$$D_m^s(f) = \int_{\mathbf{p}\in\mathbf{P}} X_{m,1}^s(f,\mathbf{p})d\mathbf{p} \quad (24)$$

with

$$\varepsilon\{X_{m,1}^s(f,\mathbf{p}_k)X_{m,1}^{s,*}(f,\mathbf{p}_l)\} = P_{X_1}^s(f)\delta_{kl} \ \forall \mathbf{p}_k,\mathbf{p}_l \in \mathbf{P}. \quad (25)$$

To separate the estimation of the spectral and spatial characteristics, the technique is implemented in the frequency-domain. The filter $\mathbf{W}(f)$ equals

$$W(f) = (\mu_2\mathbf{R}_m^s(f) + \mathbf{R}^n(f))^{-1}\mu_2\varepsilon\{\mathbf{X}_m^s(f)D_m^{s,*}(f)\}. \quad (26)$$

Assuming uncorrelated point sources, $\mathbf{R}_m^s(f)$ and $\varepsilon\{\mathbf{X}_m^s(f)D_m^s(f)\}$ in (26) can be computed as:

$$\mathbf{R}_m^s(f) = \int_{\mathbf{p}\in\mathbf{P}} \tilde{\mathbf{d}}^s(f,\mathbf{p})\tilde{\mathbf{d}}^{s,H}(f,\mathbf{p})\varepsilon\{X_{m,1}^s(f,\mathbf{p})X_{m,1}^{s,*}(f,\mathbf{p})\}d\mathbf{p}, = P_{X_1}^s(f)\int_{\mathbf{p}\in\mathbf{P}} \tilde{\mathbf{d}}^s(f,\mathbf{p})\tilde{\mathbf{d}}^{s,H}(f,\mathbf{p})d\mathbf{p}, \quad (27)$$

$$\varepsilon\{\mathbf{X}_m^s(f)D_m^s(f)\} = P_{X_1}^s(f)\int_{\mathbf{p}\in\mathbf{P}} \tilde{\mathbf{d}}^s(f,\mathbf{p})d\mathbf{p}, \quad (28)$$

where $P_{X_1}^s(f)$ is estimated online.
Instead of using a mathematical speech model, the speech correlation matrix $\mathbf{R}_m^s(f)$ and the cross-correlation $\varepsilon\{\mathbf{X}_m^s(f)D_m^{s,*}(f)\}$ can also be computed based on calibration data [7].

4. ONLINE SPEECH MODEL ($\mu_2 = 0$)

In this section, techniques that use an online estimate of the speech statistics are discussed, i.e., the TF-LCMV [3] and the SDW-MWF [4]. Since the source signal $S(f)$ is unknown, these techniques estimate the speech component in one of the microphones (e.g., the first microphone), i.e., $D^s(f) = X_1^s(f)$ (or in the output of a fixed beamformer). These techniques typically exploit a voice activity detection (VAD) mechanism and assume the noise statistics to be more stationary than the speech statistics. Hence, VAD errors or highly non-stationary noise may affect the performance.

4.1. Hard constraint ($\mu_1 = \infty$): TF-LCMV

The TF-LCMV beamformer [3] minimizes the output noise power subject to the constraint that the speech component in the first microphone signal is preserved, i.e.,

$$\mathbf{W}^H\mathbf{X}^s(f) = X_1^s(f) \text{ or } \mathbf{W}^H\tilde{\mathbf{H}}^s(f) = 1, \quad (29)$$

with $\tilde{\mathbf{H}}^s(f)$ is the relative transfer function ratio vector defined in (5). This corresponds to (18) with $\mu_1 = \infty$, $\mu_2 = 0$ and $D^s(f) = X_1^s(f)$, resulting in (cf. the derivation in Section 3.1)

$$\mathbf{W}(f) = \frac{\mathbf{R}^{n^{-1}}(f)\tilde{\mathbf{H}}^s(f)}{\tilde{\mathbf{H}}^{s,H}(f)\mathbf{R}^{n-1}(f)\tilde{\mathbf{H}}^s(f)}. \quad (30)$$

To impose the hard constraint (29), the relative transfer function ratios $\tilde{\mathbf{H}}^s(f)$ need to be identified. In [3], an unbiased estimate of $\tilde{\mathbf{H}}^s(f)$ is computed during speech periods by exploiting the nonstationarity of the desired signal and the stationarity of the noise.
Remark: The GSC with switching adaptive filters [8] and the GSC with adaptive blocking matrix [9, 10] also belong to this class. Here, $\tilde{\mathbf{H}}^s(f)$ is estimated throug a least-squares match

between the microphone signals and the first microphone signal [8] *or the output of a fixed beamformer* [9, 10]. *Due to the presence of noise, this estimate is biased.*

4.2. Soft constraint ($\mu_1 \neq \infty$): SDW-MWF

The SDW-MWF [4] minimizes the output noise power subject to a soft constraint on the speech distortion, corresponding to (18) with $\mu_1 \neq \infty$ and $D^s(f) = X_1^s(f)$, resulting in

$$\mathbf{W}(f) = (\mathbf{R}^n(f) + \mu_1 \mathbf{R}^s(f))^{-1} \mu_1 \varepsilon\{\mathbf{X}^s(f) X_1^{s,H}(f)\}. \quad (31)$$

The speech correlation matrix $\mathbf{R}^s(f)$ is estimated by exploiting stationarity of the noise and a VAD mechanism.
Assuming that $\mathbf{R}^s(f)$ is rank-one,$\mathbf{W}(f)$ can be decomposed into a TF-LCMV with a single-channel SDW postfilter [4]

$$\underbrace{\frac{\mathbf{R}^{n^{-1}}(f)\tilde{\mathbf{H}}^s(f)}{\tilde{\mathbf{H}}^s(f)\mathbf{R}^{n^{-1}}(f)\tilde{\mathbf{H}}^s(f)}}_{\text{TF-LCMV}} \underbrace{\left(\frac{\mu_1 P_{X_1}^s(f)}{\mu_1 P_{X_1}^s(f) + \frac{1}{\tilde{\mathbf{H}}^{s,H}(f)\mathbf{R}^{n^{-1}}\tilde{\mathbf{H}}^s(f)}}\right)}_{\text{postfilter}}.$$

Hence, the soft constraint on the speech distortion term introduces spectral filtering of the speech component $X_1^s(f)$ (unless the speech and the noise subspace are orthogonal such that $\frac{1}{\tilde{\mathbf{H}}^{s,H}(f)\mathbf{R}^{n^{-1}}\tilde{\mathbf{H}}^s(f)} = 0$).

5. COMBINATION OF AN ONLINE AND A-PRIORI SPEECH MODEL

So far, either an a-priori speech model or an online estimated speech model was used in (18). However, also a combination of a-priori knowledge and online estimation (based on incoming data) can be used. This approach allows for a (partial) update of the speech model while it is expected to increase robustness to an erroneous estimation of the speech model (e.g., due to VAD failures).

5.1. Hard constraint on a-priori model ($\mu_2 = \infty$, $\mu_1 \neq \infty$): speech distortion regularized GSC (SDR-GSC)

In the SDR-GSC [4], the LCMV beamformer is combined with the SDW-MWF. A hard constraint is imposed on an a-priori speech model (i.e., $\mu_2 = \infty$), e.g.,

$$\mathbf{X}_m^s(f) = \tilde{\mathbf{d}}^s(f, \mathbf{p}^s) X_{m,1}^s(f), \quad (32)$$
$$D_m^s(f) = X_{m,1}^s(f). \quad (33)$$

The hard constraint is imposed through a GSC-structure with a fixed beamformer $\mathbf{W}_q(f)$ (e.g., $\mathbf{W}_q(f) = \frac{\tilde{\mathbf{d}}^s(f,\mathbf{p}^s)}{M}$) and a blocking matrix $\mathbf{B}(f)$ with $\mathbf{B}^H(f)\mathbf{W}_q(f) = 0$, i.e.,

$$\mathbf{W}(f) = \mathbf{W}_q(f) + \mathbf{B}(f)\mathbf{W}_a(f), \quad (34)$$

with $\mathbf{W}_a(f)$ the adaptive noise canceller.
In addition to the hard constraint, a soft constraint ($\mu_1 \neq \infty$) is imposed on the online estimated speech distortion between the speech component in the speech reference $D^s(f) = \mathbf{W}_q^H \mathbf{X}^s(f)$ and the speech component in the output, i.e., $\mathbf{W}^H(f)\mathbf{X}^s(f)$. Using (34), the online estimated speech distortion term in (18) equals:

$$\varepsilon\{\mathbf{W}_a^H(f)\mathbf{B}^H(f)\mathbf{X}^s(f)\mathbf{X}^{s,H}(f)\mathbf{B}(f)\mathbf{W}_a(f)\}, \quad (35)$$

which corresponds to the regularization term in the SDR-GSC. Using (35) in (18), results in the SDR-GSC cost function in [4].

5.2. Soft constraint on a-priori model ($\mu_1 \neq \infty$, $\mu_2 \neq \infty$,): combination soft constrained/SDW-MWF

Setting $\mu_1 \neq \infty$ and $\mu_2 \neq \infty$ in (18), results in a combination of the SDW-MWF (cf. Section 4.2) and the soft constrained beamformer (cf. Section 3.2). The speech model is then partially updated based on incoming data and partially computed a-priori using (23)-(24) or calibration data [7]. The filter $\mathbf{W}(f)$ equals

$$\mathbf{W}(f) = (\mu_1 \mathbf{R}^s(f) + \mu_2 \mathbf{R}_m^s(f) + \mathbf{R}^n(f))^{-1} \cdot (\mu_1 \varepsilon\{\mathbf{X}^s(f) D^{s,*}(f)\} + \mu_2 \varepsilon\{\mathbf{X}_m^s(f) D_m^{s,*}(f)\}) \quad (36)$$

with $\mathbf{R}_m^s(f)$ and $\varepsilon\{\mathbf{X}_m^s(f) D_m^{s,*}(f)\}$ computed as (27)-(28) or computed based on calibration data.
In the future, this combined approach will be compared with the SDW-MWF and the soft constrained beamformer in terms of performance and robustness.

6. REFERENCES

[1] K. M. Buckley, "Broad-band beamforming and the Generalized Sidelobe Canceller," *IEEE Trans. ASSP*, vol. 34, no. 5, pp. 1322–1323, Oct. 1986.

[2] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. AP*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and non-stationarity with applications to speech," *IEEE Trans. SP*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[4] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, Dec. 2004.

[5] S. Nordholm, H.Q. Dam, N. Grbíc, and S. Y. Low, *Adaptive microphone array employing spatial quadratic soft constraints and spectral shaping*, chapter 10 in "Speech Enhancement" (Benesty, J., Makino, S. and Chen, J.,Eds.), pp. 229–246, Springer-Verlag, 2005.

[6] A. Spriet, S. Doclo, M. Moonen, and J. Wouters, "Unification of multi-microphone noise reduction systems," Tech. Rep. ESAT-SISTA/TR 2006-72, K.U. Leuven, Belgium, Apr. 2006.

[7] S. Nordholm, I. Claesson, and M. Dahl, "Adaptive microphone array employing calibration signals: an analytical evaluation," *IEEE Trans. SAP*, vol. 7, no. 3, pp. 241–252, May 1999.

[8] D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," in *Proc. of ICASSP*, Albuquerque, Apr. 1990, vol. 2, pp. 833–836.

[9] W. Herbordt and W. Kellermann, *Adaptive Beamforming for Audio Signal Acquisition*, chapter 6 in "Adaptive Signal Processing: Applications to Real-World Problems" (Benesty, J. and Huang, Y., Eds.), pp. 155–188, Springer-Verlag, 2003.

[10] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE TSP*, vol. 47, pp. 2677–2683, 1999.