10 Adaptive Microphone Array Employing Spatial Quadratic Soft Constraints and Spectral Shaping

Sven Nordholm[1], Hai Quang Dam[1], Nedelko Grbić[2], and Siow Yong Low[1]

[1] Western Australian Telecommunications Research Institute (WATRI)
Crawley, WA 6009, Australia
E-mail: {sven, damhai, siowyong}@watri.org.au

[2] Blekinge Institute of Technology
Department of Telecommunications and Signal Processing
Ronneby, SE 37225, Sweden
E-mail: nedelko.grbic@bth.se

Abstract. The convenience and the ease of use provided by hands-free operation of speech communication devices mean that speech enhancement schemes are becoming indispensable. In this chapter, two subband adaptive microphone array schemes are presented, which aim to provide good speech enhancement capability in poor signal to noise ratio situations. The basic commonality of the adaptive microphone array schemes is that they approximate the Wiener solution in an adaptive manner as new data comes in. Furthermore, both schemes include a quadratic constraint to prevent the trivial zero solution of the weights and to avoid suppression of the source of interest. The constraint is included to provide robustness against model mismatch and good spatial capture of the target signal. Furthermore, by using a subband structure the processing allows a time-frequency operation for each channel. As such, both schemes utilize the spatial, spectral, and temporal domains in an efficient and concise manner allowing a computational effective processing while maintaining high performance speech enhancement. Evaluations on the same data set, gathered from a car, show that the proposed schemes achieve good noise suppression up to 20 dB while experiencing very low levels of speech distortion.

10.1 Introduction

The comfort and flexibility provided by hands-free communication systems have spurred the integration of hands-free voice interface into everyday essentials such as personal digital assistants (PDAs), mobile phones, speech recognition devices, etc. With such a great demand, speech enhancement with regard to hands-free communications particularly in adverse environments has been an area of intensive research [1], [2], [3], [4], [5], [6]. Numerous speech enhancement schemes have been presented over the years with microphone array based techniques dominating the field. This is because microphone arrays offer the invaluable spatial diversity to spatially extract (or form a beam towards) the source of interest (SOI) [7]. In particular, adaptive microphone arrays are reported to have good interference suppression capability [8], [9], [10]. However, adaptive microphone array such as the generalized sidelobe canceller (GSC) succumbs to target signal cancellation in the presence of steering vector errors (e.g. microphone positions, reverberation, etc) [11], [12]. One solution to overcome the problem is to employ a voice activity detector (VAD) or an energy detector in which the GSC is only adapted when there is no target signal (or the signal-to-interference ratio (SIR) is low). Another more straightforward approach to address the problem is to calibrate the microphone array in the actual environment [13], [14]. By doing so, all the information on the array geometry and imperfections will be reflected in the final solution. This approach seems efficient and robust at first glance, but the need for calibration makes its use rather limited in consumer applications. For instance, when the SOI spatially moves or the environment changes, it requires a re-calibration. Such inflexibility may not be practically viable.

In this chapter, we will present two subband based schemes, namely robust soft constrained adaptive microphone array (RSCAMA) and noise statistics updated adaptive microphone array (NSUAMA). Both schemes have their roots in the calibrated microphone array [13], [14] but circumvent the calibration phase which makes them considerably more versatile. Instead, a source model is carefully embedded in the solution whereas the noise statistics is estimated on-line. To complement the source model, the SOI power spectral density (PSD) is also estimated from the data to preserve the spectral shape of the SOI. The real objective is to achieve a solution that is close to the optimal Wiener solution [15], [16] whilst incorporating a tracking capability to handle non-stationary noise. Both structures differ in the way the SOI power spectral density and noise statistics are incorporated in the solution but share the commonality of having a 2-D space constrained source model. Unlike a point source model, the 2-D space model (the physical area of the SOI e.g. a person's mouth) effectively compensates for the large radial vector errors in the source location caused by the presence of erroneous steering vector in real life situations, making both proposed structures robust against errors.

The RSCAMA scheme is constrained to extract the SOI in a pre-defined area (as modelled by the 2-D space constraints). Basically, the idea is originally derived from [4] with the assumption that the power spectral density (PSD) of the source is constant over time and frequency range. However, speech signal is short-term stationary and this implies that the spectrum varies over time. Therefore, to better utilize the time-frequency information of the SOI, its PSD is recursively updated in the constraints using the most current time-frequency content of the output signal from the beamformer. The motivation behind the use of the output signal in the update comes from the fact that the optimum beamformer output in each subband, is an enhanced version of the spectral information of the SOI. In other words, the feedback from the beamformer output continuously shapes the SOI spectrum, thus providing a spectrally improved constraint at each time instant. The noise statistics on the other hand, are estimated recursively from the received data without the need of a VAD as the solution has been constrained to preserve any signal from the desired region. Needless to say, the performance will be very much improved if a VAD is used, however at the expense of a higher computational complexity.

As the name suggests, the NSUAMA scheme includes a noise statistics update to track variations in the background noise. Simply, the adaptive microphone array estimates the covariance information and decides if the estimated information can be used to update the noise statistics in the solution i.e. the noise covariance information. A modified VAD or a noise covariance detector which includes spatial information is introduced to ensure that "only noise covariance" information is used in the update. With the incorporation of the criterion, the microphone array behaves like a "noise only" detector which uses only the noise information to update its solution. This results in an efficient and fast converging adaptive microphone array even in highly non-stationary environment. Similar to RSCAMA, the source PSD is embedded in the optimum Wiener solution in each subband to fully utilize the time-frequency information of the target signal. However unlike RSCAMA, the source PSD is updated using a least-squares criterion [17]. As such, it tracks the variation in the spectral content of the target signal continuously, yielding a statistically optimized constraint for each time instant.

Clearly, the major difference between the RSCAMA and NSUAMA schemes is their computational complexities. The RSCAMA structure offers simplicity and is straightforward to implement in real-time. Naturally, the downside of it is less suppression capability when compared to the NSUAMA scheme. Evaluations in a real car hands-free scenario reveal that the NSUAMA scheme manages to achieve an impressive noise suppression level of 20 dB whilst the simpler RSCAMA performs around 16-17 dB. Most importantly, both schemes maintain negligible distortion on the target signal.

10.2 Signal Modelling and Problem Formulation

Consider a linear microphone array with I microphones. The target signal in this case is a person speaking, which can be modelled as an infinite number of point sources clustered closely in space. This space is modelled as a circular area $\mathbf{A}$ with radius r and a distance h from the array, see Fig. 10.1. Alternatively, the source constrained region can be modelled as a pie sliced area defined by radii $[R_a, R_b]$ and angles $[\theta_a, \theta_b]$ [8]. As mentioned previously, the advantage of the source constrained region in Fig. 10.1 as opposed to a point source is consistent with the fact that errors in the response vector cause large radial errors in the corresponding source location [11]. These errors are typically due to sensor misplacement and gain variations in the microphones. With the inclusion of the constrained area, the structure is made more robust

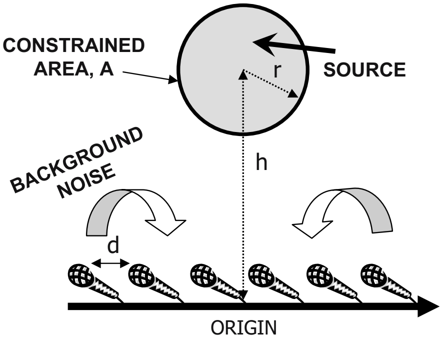

Fig. 10.1. Configuration of the linear microphone array with the inter-element distance d and the source constrained area defined by radius r and distance h.

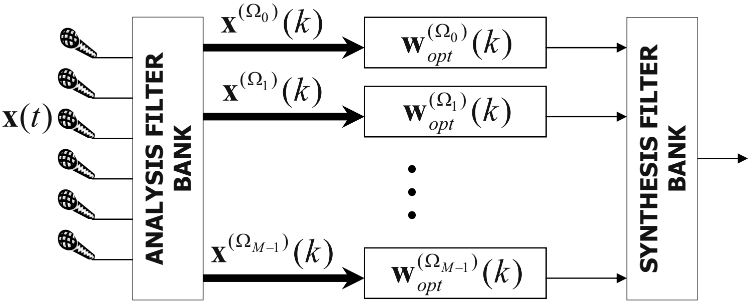
$\mathbf{x}(t)$
ANALYSIS FILTER BANK
$\mathbf{x}^{(\Omega_0)}(k)$
$\mathbf{x}^{(\Omega_1)}(k)$
$\mathbf{x}^{(\Omega_{M-1})}(k)$
$\mathbf{w}_{opt}^{(\Omega_0)}(k)$
$\mathbf{w}_{opt}^{(\Omega_1)}(k)$
$\mathbf{w}_{opt}^{(\Omega_{M-1})}(k)$
SYNTHESIS FILTER BANK

Fig. 10.2. Structure of the RSCAMA subband beamformer.

and more closely related to a real situation. Throughout the chapter, the SOI is assumed to be in the constrained region as shown in Fig. 10.1.

Figures 10.2 and 10.3 show the block diagrams of the proposed RSCAMA and NSUAMA respectively. Irrespective of the different structures, both the subband based RSCAMA and NSUAMA aim to extract the SOI in the constrained region. From the figures, the received signal is initially decomposed into M subbands by using an analysis filterbank. After the relevant processing independently (in each structure), the processed subband signals are then reconstructed by the synthesis filterbank into fullband representation.

10.2.1 Analysis and Synthesis Filterbanks

The main consideration in the design is to minimize aliasing in the subband signals as well as minimizing magnitude, phase and aliasing distortion in the reconstructed output. Literature associated with filterbanks can be found in the following references [18], [19]. In this work, an oversampled uniform analysis DFT filterbank is employed to decompose each of the I microphone input signals into M subbands with an oversampling decimation factor of

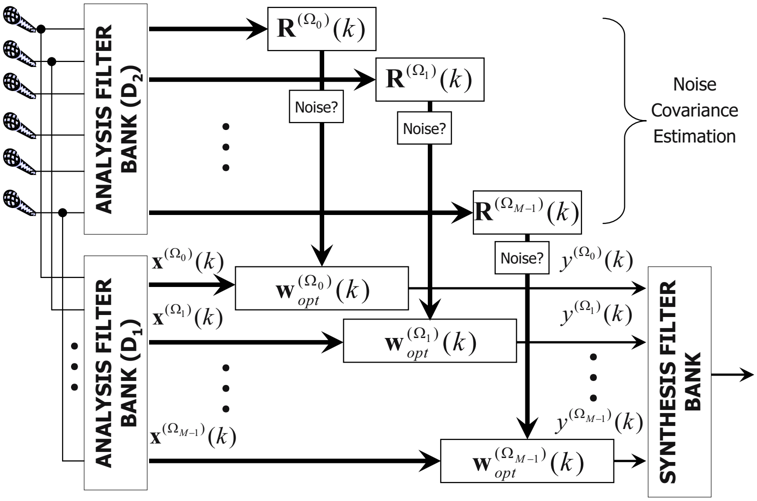
ANALYSIS FILTER BANK (D2)
$\mathbf{R}^{(\Omega_0)}(k)$
$\mathbf{R}^{(\Omega_1)}(k)$
$\mathbf{R}^{(\Omega_{M-1})}(k)$
Noise?
Noise?
Noise?
Noise Covariance Estimation
ANALYSIS FILTER BANK (D1)
$\mathbf{x}^{(\Omega_0)}(k)$
$\mathbf{x}^{(\Omega_1)}(k)$
$\mathbf{x}^{(\Omega_{M-1})}(k)$
$\mathbf{w}_{opt}^{(\Omega_0)}(k)$
$\mathbf{w}_{opt}^{(\Omega_1)}(k)$
$\mathbf{w}_{opt}^{(\Omega_{M-1})}(k)$
$y^{(\Omega_0)}(k)$
$y^{(\Omega_1)}(k)$
$y^{(\Omega_{M-1})}(k)$
SYNTHESIS FILTER BANK

Fig. 10.3. Structure of the NSUAMA subband beamformer. D_1 and D_2 are the different decimation factors with $D_1 \geq D_2$.

$D_1 = M/2$ unless otherwise stated. By oversampling, the inband aliasing effects is greatly reduced. The analysis and synthesis prototype filters are designed using a Hamming window with a cut off frequency π/M. The Hamming window has side-lobes that are 50 dB below its mainlobe and by using a factor two over-sampling, the overall distortion and aliasing will be kept small. Note that the noise covariance estimation in Fig. 10.3 has its own requirement, as it is decimated at a lower rate of D_2, where $D_1 > D_2$. This is to ensure the sufficiency of data in estimating the noise covariance matrix (i.e. to achieve low variance estimate for a better tracking in the noise statistics).

10.2.2 The Wiener Solution

In this section, the multichannel Wiener filter in each subband is formulated. To begin, let $\mathbf{w}_{opt}^{(\Omega)}(k)$ be the optimum weight vector at time index k for each frequency $\Omega \in [\Omega_0, \cdots, \Omega_{M-1}]$ as

$$\mathbf{w}_{opt}^{(\Omega)}(k) = [w_1^{(\Omega)}(k), w_2^{(\Omega)}(k), \cdots, w_I^{(\Omega)}(k)]^T. \tag{10.1}$$

The optimal weight vector at time point k above can be readily found from the Wiener solution as follows,

$$\mathbf{w}_{opt}^{(\Omega)}(k) = [\mathbf{R}_s^{(\Omega)}(k) + \mathbf{R}_n^{(\Omega)}(k)]^{-1}\mathbf{r}_s^{(\Omega)}(k), \tag{10.2}$$

where $\mathbf{R}_s^{(\Omega)}(k)$ and $\mathbf{r}_s^{(\Omega)}(k)$ are the covariance matrix and the cross-covariance vector for the SOI for frequency band Ω, respectively. The covariance matrix

$\mathbf{R}_s^{(\Omega)}(k)$ can be resolved into a normalized spatial covariance matrix $\bar{\mathbf{R}}_s^{(\Omega)}(k)$ and a non-negative spectral weighting as

$$\mathbf{R}_s^{(\Omega)}(k) = S^{(\Omega)}(k)\bar{\mathbf{R}}_s^{(\Omega)}. \tag{10.3}$$

Likewise, the cross-covariance vector $\mathbf{r}_s^{(\Omega)}$ can be decomposed as

$$\mathbf{r}_s^{(\Omega)}(k) = S^{(\Omega)}(k)\bar{\mathbf{r}}_s^{(\Omega)}, \tag{10.4}$$

where $\bar{\mathbf{r}}_s^{(\Omega)}$ is the normalized spatial cross-covariance vector. Substituting (10.3) and (10.4) into (10.2) yields,

$$\mathbf{w}_{opt}^{(\Omega)}(k) = [S^{(\Omega)}(k)\bar{\mathbf{R}}_s^{(\Omega)} + \mathbf{R}_n^{(\Omega)}(k)]^{-1}S^{(\Omega)}(k)\bar{\mathbf{r}}_s^{(\Omega)}. \tag{10.5}$$

Equation (10.5) forms the basis for the development of both microphone array schemes RSCAMA and NSUAMA. From (10.5), it is clear that there are two varying parameters that need to be estimated continuously i.e. $S^{\Omega}(k)$ and $\mathbf{R}_n^{(\Omega)}(k)$. The former functions as the source spectral moulder (to reduce spectral distortion) and the latter is to track the noise statistics for optimal noise suppression. The SOI spatial covariance matrix $\bar{\mathbf{R}}_s^{(\Omega)}$ and the spatial cross-covariance vector $\bar{\mathbf{r}}_s^{(\Omega)}$ on the other hand, are determined by the spatial location of the SOI. For many applications such as internet telephony, hands-free mobile telephony, etc, the SOI is typically located more or less in a fixed position (in front of the array). In keeping with this, the SOI is assumed to be spatially stationary in a pre-defined region and a constraint called the space constraint is used to model it. Both the RSCMA and NSUAMA schemes employ the space constraint to model the SOI spatial information. In the following section, the space constraint is explained.

10.2.3 The Space Constrained Source Covariance Information

Let us denote $S^{(\Omega)}$ as the PSD of the source at frequency Ω. Note that the PSD will be time varying and can be thought of as short-term stationary. However, for the following model it is kept constant. As mentioned previously, the source is assumed to be in the pre-defined area $\mathbf{A}$ afore-mentioned (see Fig. 10.1). Thus, the spatio-temporal covariance matrix of source in the spectral band $[\Omega_a, \Omega_b]$ can be computed as

$$\mathbf{R}_s = \int_{\Omega_a}^{\Omega_b} \iint\limits_{\mathbf{A}} S^{(\Omega)}\mathbf{d}^{(\Omega)}(\overrightarrow{\mathbf{a}})(\mathbf{d}^{(\Omega)}(\overrightarrow{\mathbf{a}}))^H d\overrightarrow{\mathbf{a}}d\Omega, \tag{10.6}$$

where $\overrightarrow{\mathbf{a}}$ is the point source localization vector and $(\cdot)^H$ denotes the Hermitian transposition. The response vector $\mathbf{d}^{(\Omega)}(\overrightarrow{\mathbf{a}})$ is defined as

$$\mathbf{d}^{(\Omega)}(\overrightarrow{\mathbf{a}}) = \left[\frac{1}{R_1}e^{-j\Omega\tau_1(\overrightarrow{\mathbf{a}})}, \frac{1}{R_2}e^{-j\Omega\tau_2(\overrightarrow{\mathbf{a}})}, \cdots, \frac{1}{R_I}e^{-j\Omega\tau_I(\overrightarrow{\mathbf{a}})}\right]^T, \tag{10.7}$$

where $\tau_i(\overrightarrow{\mathbf{a}})$, $1 \leq i \leq I$ is the time delay from a point source in the predefined area to sensor i, R_i is the distance between the source and sensor i and $[\cdot]^T$ denotes the transposition operator. The reference point for the microphone array response is defined at the origin of the coordinates.

Therefore for a frequency Ω, the spatial covariance matrix in (10.3) is,

$$\mathbf{R}_s^{(\Omega)} = S^{(\Omega)} \bar{\mathbf{R}}_s^{(\Omega)}, \tag{10.8}$$

where the normalized spatial covariance matrix is defined from (10.6) as,

$$\bar{\mathbf{R}}_s^{(\Omega)} = \iint\limits_{\mathbf{A}} \mathbf{d}^{(\Omega)}(\overrightarrow{\mathbf{a}}) \mathbf{d}^{(\Omega)}(\overrightarrow{\mathbf{a}})^H d\overrightarrow{\mathbf{a}}. \tag{10.9}$$

The spatial cross covariance vector is given by

$$\mathbf{r}_s^{(\Omega)} = S^{(\Omega)} \bar{\mathbf{r}}_s^{(\Omega)}, \tag{10.10}$$

where

$$\bar{\mathbf{r}}_s^{(\Omega)} = \iint\limits_{\mathbf{A}} \mathbf{d}^{(\Omega)}(\overrightarrow{\mathbf{a}}) d\overrightarrow{\mathbf{a}}. \tag{10.11}$$

With the space constrained model readily available, the task at hand is to efficiently estimate the varying parameters $S^{(\Omega)}(k)$ and $\mathbf{R}_n^{(\Omega)}(k)$ in (10.5). This is where the distinction between the RSCAMA and NSUAMA structures comes in. The simpler RSCAMA estimates the information directly irrespective of whether the SOI is active or inactive whereas the NSUAMA performs otherwise. Sections 10.3 and 10.4 explain both the RSCAMA and NSUAMA structures in detail.

10.3 Robust Soft Constrained Adaptive Microphone Array (RSCAMA)

10.3.1 Problem Formulation

Let $\mathbf{w}_{opt}^{(\Omega)}$ be the optimum weight vector for frequency Ω,

$$\mathbf{w}_{opt}^{(\Omega)} = [w_1^{(\Omega)}, w_2^{(\Omega)}, \ldots, w_I^{(\Omega)}]^T, \tag{10.12}$$

where $w_i^{(\Omega)}$ is the optimum coefficient for the i^{th} sensor. The optimum weight vector is then calculated as

$$\mathbf{w}_{opt}^{(\Omega)} = \left[\mathbf{R}_s^{(\Omega)} + \mathbf{R}_n^{(\Omega)}\right]^{-1} \mathbf{r}_s^{(\Omega)}, \tag{10.13}$$

where $\mathbf{R}_n^{(\Omega)}$ is the noise covariance matrix. Suppose that we have knowledge of the PSD of the SOI $S^{(\Omega)}$, then (10.13) can be rewritten as

$$\begin{aligned}\mathbf{w}_{opt}^{(\Omega)} &= \left[\mathbf{R}_s^{(\Omega)}/S(\Omega) + \mathbf{R}_n^{(\Omega)}/S(\Omega)\right]^{-1}\left(\mathbf{r}_s^{(\Omega)}/S(\Omega)\right)\\ &= \left[\bar{\mathbf{R}}_s^{(\Omega)} + \bar{\mathbf{R}}_n^{(\Omega)}\right]^{-1}\bar{\mathbf{r}}_s^{(\Omega)},\end{aligned} \tag{10.14}$$

where $\bar{\mathbf{R}}_s^{(\Omega)}$ is the normalized spatial covariance matrix given in (10.9) and $\bar{\mathbf{r}}_s^{(\Omega)}$ is the normalized spatial cross covariance vector as defined in (10.11). The implication of (10.14) is that the SOI PSD $S^{(\Omega)}$ is incorporated in the solution and both $\bar{\mathbf{R}}_s^{(\Omega)}$ and $\bar{\mathbf{r}}_s^{(\Omega)}$ can be calculated for a given constraint region without the knowledge of the PSD of the source, given that they are spatially invariant.

The remaining issue is to recursively estimate the noise parameters $\bar{\mathbf{R}}_n^{(\Omega)}$. Since data containing only the active noise is not available, the noise covariance matrix $\bar{\mathbf{R}}_n^{(\Omega)}$ is estimated by using K samples of the received data $\mathbf{x}^{(\Omega)}(k)$, where K is a fixed positive number and the index k is the subband time index. Moreover, the exact PSD of the source $S^{(\Omega)}(k)$ is not available, particularly in a car environment where strong speech masking components of noise exists. Thus, we propose to use the previous microphone array outputs for the estimation of $S^{(\Omega)}(k)$ as

$$\mathbf{z}^{(\Omega)}(k) = \frac{\mathbf{x}^{(\Omega)}(k)}{|\mathbf{w}_{opt}^{(\Omega)}(k-1)^H\mathbf{x}^{(\Omega)}(k-1)| + \delta}, \tag{10.15}$$

where $|.|$ is the absolute value operator and δ is a positive number to avoid zero division. At iteration k, $\bar{\mathbf{R}}_n^{(\Omega)}(k)$ can be estimated based on $\mathbf{z}^{(\Omega)}(m)$ where $\max(0, k-K) \leq m \leq k$ as follows,

- if $k \leq K$ then

$$\bar{\mathbf{R}}_n^{(\Omega)}(k) = \frac{1}{k}\sum_{m=1}^{k}\mathbf{z}^{(\Omega)}(m)\mathbf{z}^{(\Omega)}(m)^H, \tag{10.16}$$

- if $k > K$ then

$$\bar{\mathbf{R}}_n^{(\Omega)}(k) = \frac{1}{K}\sum_{m=k-K+1}^{k}\mathbf{z}^{(\Omega)}(m)\mathbf{z}^{(\Omega)}(m)^H. \tag{10.17}$$

In the next section, a recursive algorithm is developed to efficiently update the beamforming weights according to (10.16) and (10.17) based on the received data.

10.3.2 A Recursive Algorithm for the RSCAMA

The algorithm runs in parallel/sequentially for each subband with mid-frequency $\Omega = 2\pi f_s m/M$, $0 \le m \le M-1$, where f_s is the sampling frequency. Let

$$\bar{\mathbf{R}}^{(\Omega)}(k) = \bar{\mathbf{R}}_s^{(\Omega)} + \bar{\mathbf{R}}_n^{(\Omega)}(k) \tag{10.18}$$

and

$$\mathbf{P}^{(\Omega)}(k) = [\bar{\mathbf{R}}^{(\Omega)}(k)]^{-1}. \tag{10.19}$$

The optimal weight vector for the iteration k is then reduced to

$$\mathbf{w}_{opt}^{(\Omega)}(k) = \mathbf{P}^{(\Omega)}(k)\bar{\mathbf{r}}_s^{(\Omega)}. \tag{10.20}$$

It follows from (10.17) that for $k > K$, $\bar{\mathbf{R}}^{(\Omega)}(k)$ can be obtained from the previous estimate as

$$\bar{\mathbf{R}}^{(\Omega)}(k) = \bar{\mathbf{R}}^{(\Omega)}(k-1) + \frac{1}{K}\mathbf{z}^{(\Omega)}(k)\mathbf{z}^{(\Omega)}(k)^H - \frac{1}{K}\mathbf{z}^{(\Omega)}(k-K)\mathbf{z}^{(\Omega)}(k-K)^H. \tag{10.21}$$

Thus, the inverse matrix $\mathbf{P}^{(\Omega)}(k)$ for $k > K$ can be updated efficiently by using the matrix inversion lemma

$$\mathbf{D} = \mathbf{P}^{(\Omega)}(k-1) - \frac{\mathbf{P}^{(\Omega)}(k-1)\mathbf{z}^{(\Omega)}(k)\mathbf{z}^{(\Omega)}(k)^H\mathbf{P}^{(\Omega)}(k-1)}{\left(K + \mathbf{z}^{(\Omega)}(k)^H\mathbf{P}^{(\Omega)}(k-1)\mathbf{z}^{(\Omega)}(k)\right)} \tag{10.22}$$

and

$$\mathbf{P}^{(\Omega)}(k) = \mathbf{D} + \frac{\mathbf{D}\mathbf{z}^{(\Omega)}(k-K)\mathbf{z}^{(\Omega)}(k-K)^H\mathbf{D}}{\left(K - \mathbf{z}^{(\Omega)}(k-K)^H\mathbf{D}\mathbf{z}^{(\Omega)}(k-K)\right)}, \tag{10.23}$$

where $\mathbf{D}$ in this case is an intermediate matrix of the same size as $\mathbf{P}^{(\Omega)}(k)$. The recursive algorithm is now summarized in the following steps,

- *Step 1: Choose a number of subbands M, a block size K and a weight smoothing factor λ*[1]*.*
- *Step 2: Initialize $k = 1$ and the weight vector $\mathbf{w}_{opt}^{(\Omega)}(0)$ as an $I \times 1$ zero vector.*
- *Step 3: Calculate the matrix $\bar{\mathbf{R}}_s^{(\Omega)}$ and the vector $\bar{\mathbf{r}}_s^{(\Omega)}$ according to (10.9) and (10.11), respectively.*

[1] The factor λ is employed because the target speech signal adds spatial coherent power to the pre-calculated covariance matrix, and this in turn leads to small weight power fluctuations.

- *Step 4: If $k \leq K$, the matrix $\mathbf{P}^{(\Omega)}(k)$ is calculated according to (10.15), (10.16) and (10.19) by using pseudo-inverse operation instead of the conventional matrix inverse operation due to rank deficiency. Otherwise, the matrix $\mathbf{P}^{(\Omega)}(k)$ is updated recursively by using (10.22) and (10.23). The weight vector is then updated as*

  $$\mathbf{w}_{opt}^{(\Omega)}(k) = \lambda \mathbf{w}_{opt}^{(\Omega)}(k-1) + (1-\lambda)\mathbf{P}^{(\Omega)}(k)\bar{\mathbf{r}}_s^{(\Omega)},$$

  and the output is given by

  $$y^{(\Omega)}(k) = \mathbf{w}_{opt}^{(\Omega)}(k)^H \mathbf{x}^{(\Omega)}(k).$$

- *Step 5: Set $k = k+1$ and return to Step 4 until the end of the data.*

10.4 Noise Statistics Updated Adaptive Microphone Array (NSUAMA)

10.4.1 Problem Formulation

In the formulation of the RSCAMA scheme, the update of the noise covariance matrix estimate $\bar{\mathbf{R}}_n^{(\Omega)}(k)$ is performed continuously sample by sample. This means that the covariance information also contains the SOI. Naturally, if the noise covariance estimation is free from the SOI, the advantages will be twofold i.e. better noise suppression and consequently better source PSD estimate. Here, the NSUAMA scheme employs a "noise covariance detector" to avoid the inclusion of the SOI in the noise covariance matrix.

In order to explain this formulation, we consider the Wiener solution [eq. (10.5)] again

$$\mathbf{w}_{opt}^{(\Omega)}(k) = [S^{(\Omega)}(k)\bar{\mathbf{R}}_s^{(\Omega)} + \mathbf{R}_n^{(\Omega)}(k)]^{-1} S^{(\Omega)}(k)\bar{\mathbf{r}}_s^{(\Omega)}. \tag{10.24}$$

As before, both the $\bar{\mathbf{R}}_s^{(\Omega)}$ and $\bar{\mathbf{r}}_s^{(\Omega)}$ can be precalculated according to (10.9) and (10.11) respectively as long as the SOI is spatially invariant. Similar to the RSCAMA scheme, the objective is to calculate the Wiener solution above by efficiently estimating the power spectrum of the SOI $S^{(\Omega)}(k)$ and the noise covariance matrix $\mathbf{R}_n^{(\Omega)}(k)$.

10.4.2 The Noise Covariance Detector

From the pre-defined source area model $\mathbf{A}$, the matrix $\mathbf{R}_s^{(\Omega)}$ in (10.9) for frequency Ω has non-zero determinant and is therefore a full rank matrix[2]. Thus, this matrix can be decomposed as follows,

$$\bar{\mathbf{R}}_s^{(\Omega)} = \mathbf{V}^{(\Omega)}\mathbf{\Lambda}^{(\Omega)}\mathbf{V}^{(\Omega)H}, \tag{10.25}$$

[2] Depending on how much of the space it spans, it will have a few dominating eigenvalues.

where

$$\mathbf{V}^{(\Omega)} = [\mathbf{v}_1^{(\Omega)}, \cdots, \mathbf{v}_I^{(\Omega)}] \tag{10.26}$$

is a matrix that contains the eigenvectors and

$$\boldsymbol{\Lambda}^{(\Omega)} = \mathrm{diag}\{\lambda_1^{(\Omega)}, \cdots, \lambda_I^{(\Omega)}\} \tag{10.27}$$

is a diagonal matrix that consists of the eigenvalues. Since the SOI and noise are assumed to be uncorrelated and by using the proposed source covariance model, the total covariance matrix can be written as

$$\mathbf{R}^{\Omega}(k) = S^{\Omega}(k)\bar{\mathbf{R}}_s^{\Omega} + \mathbf{R}_n^{\Omega}(k), \tag{10.28}$$

where the total covariance matrix $\mathbf{R}^{\Omega}(k)$ can be calculated from the received signal $\mathbf{x}^{(\Omega)}(k)$ by K of its samples as follows

$$\mathbf{R}^{(\Omega)}(k) = \frac{1}{K} \sum_{m=k-K+1}^{k} \mathbf{x}^{(\Omega)}(m)\mathbf{x}^{(\Omega)}(m)^H. \tag{10.29}$$

By multiplying the left and right side of (10.28) with the eigenvector $\mathbf{v}_{max}^{(\Omega)}$ that corresponds to the largest eigenvalue of $\bar{\mathbf{R}}_s^{(\Omega)}$, $\lambda_{max}^{(\Omega)}$, we have the following equation

$$\mathbf{v}_{max}^{(\Omega)}{}^H \mathbf{R}^{(\Omega)}(k)\mathbf{v}_{max}^{(\Omega)} =$$

$$S^{(\Omega)}(k)\ \mathbf{v}_{max}^{(\Omega)}{}^H \bar{\mathbf{R}}_s^{(\Omega)}\mathbf{v}_{max}^{(\Omega)} + \mathbf{v}_{max}^{(\Omega)}{}^H \mathbf{R}_n^{(\Omega)}(k)\mathbf{v}_{max}^{(\Omega)}. \tag{10.30}$$

The right hand side of (10.30) can be simplified to

$$\mathbf{v}_{max}^{(\Omega)}{}^H \mathbf{R}^{(\Omega)}(k)\mathbf{v}_{max}^{(\Omega)} =$$

$$S^{(\Omega)}(k)\ \lambda_{max}^{(\Omega)} + \mathbf{v}_{max}^{(\Omega)}{}^H \mathbf{R}_n^{(\Omega)}(k)\mathbf{v}_{max}^{(\Omega)}. \tag{10.31}$$

The purpose of using $\mathbf{v}_{max}^{(\Omega)}$ is consistent with the fact that it represents the strongest component in the target signal subspace. By denoting

$$F^{(\Omega)}(k) = \mathbf{v}_{max}^{(\Omega)}{}^H \mathbf{R}^{(\Omega)}(k)\mathbf{v}_{max}^{(\Omega)}, \tag{10.32}$$

(10.31) can be rewritten as,

$$F^{(\Omega)}(k) = S^{(\Omega)}(k)\lambda_{max}^{(\Omega)} + \mathbf{v}_{max}^{(\Omega)}{}^H \mathbf{R}_n^{(\Omega)}(k)\mathbf{v}_{max}^{(\Omega)}. \tag{10.33}$$

In the following, we will propose a criterion for the case when the noise is assumed to be long-term stationary (such as in a car or helicopter environment) whereas the speech signal is short-term stationary. This means that the statistics of the noise remain unchanged for at least 1 second. Using this assumption, there exists a number of sample points $L >> K$ which corresponds to 1 second in time, where the noise is stationary during the interval $[k-L, k]$. As such, if the SOI is silent, then the number of K sample points in (10.29) will be sufficient to capture the noise statistics for that particular period.

It follows from (10.33) that when there is no SOI at sample instant k, the first term $S^{(\Omega)}(k)\lambda_{max}^{(\Omega)}$ will be approximately zero. The second term ${\mathbf{v}_{max}^{(\Omega)}}^H \mathbf{R}_n^{(\Omega)}(k)\mathbf{v}_{max}^{(\Omega)}$ will reduce to a minimum value for $F^{(\Omega)}(k)$ during $[k-L, k]$ period due to the stationarity of the noise in that time frame. This term essentially represents the lower bound for the function in (10.33). Naturally, if speech (i.e SOI is active) is present, then the value in $F^{(\Omega)}(k)$ will be higher than its lower bound. Strictly speaking, it is a function that contains information on the periods of "speech-silence" in the constrained area. Having said so, a criterion can be formulated as follows,

$$F^{(\Omega)}(k) - \min_{[k-L,k]} F^{(\Omega)} < \lambda_{max}^{(\Omega)}\varepsilon, \forall\, \Omega, \tag{10.34}$$

where $\min_{[k-L,k]}(F^{(\Omega)})$ denotes the minimum of $F^{(\Omega)}$ over the period[3] $[k-L, k]$. The parameter ε in this case is the threshold in the detector. If the criterion in (10.34) is met for all frequency bands, then the covariance matrix $\mathbf{R}^{(\Omega)}(k)$ is used to update the estimated noise covariance matrix $\mathbf{R}_n^{(\Omega)}(k)$, through a first order smoothing function given by,

$$\mathbf{R}_n^{(\Omega)}(k) = (1-\lambda)\mathbf{R}_n^{(\Omega)}(k-1) + \lambda\mathbf{R}^{(\Omega)}(k). \tag{10.35}$$

The constant λ in this case is the smoothing factor. If the condition in (10.34) is not met, then

$$\mathbf{R}_n^{(\Omega)}(k) = \mathbf{R}_n^{(\Omega)}(k-1). \tag{10.36}$$

Since the speech signal has most of its energy in the frequency range 500 Hz to 2000 Hz, the criterion in (10.34) can be performed only in the speech dominant subbands. In other words, the detector can be implemented in the frequency range where the speech energy mainly concentrates.

10.4.3 Estimation of Power Spectrum of SOI

The SOI PSD can be estimated by using the least-squares approach given as

$$S^{(\Omega)}(k) = \arg \min_{S^{(\Omega)}, S^{(\Omega)}>0} \parallel \mathbf{R}^{(\Omega)}(k) - \mathbf{R}_n^{(\Omega)}(k) - S^{(\Omega)}\bar{\mathbf{R}}_s^{(\Omega)} \parallel_{\mathcal{F}}^2, \tag{10.37}$$

[3] This period is the interval in which the noise statistics remains unchanged. Since the noise considered is long-term stationary, a suitable duration will be around one second.

where $\| \cdot \|_{\mathcal{F}}$ is the Frobenius norm operator. For ease of computation, (10.37) can be efficiently solved by stacking the columns of each matrix to form a I^2 long vector. This problem can then be reduced to a quadratic optimization problem with I^2 variables. By setting the first derivative of (10.37) to zero, the optimum $S^{(\Omega)}(k)$ can be obtained. This PSD is estimated at every iteration of the received signal covariance matrix to provide a spectrally optimized constraint on the source. In simple terms, it attempts to preserve the spectrum of the source like a spectrum moulder.

10.4.4 The NSUAMA Algorithm

For simplicity, the noise covariance matrix can be updated in the algorithm only every one second due to the assumption on the long-term stationarity for the noise (otherwise, the noise covariance matrix can be re-evaluated by (10.35) in every iteration). Equations (10.24) and (10.37) can be reformulated as follows

$$\mathbf{w}_{opt}^{(\Omega)}(k) = [S^{(\Omega)}(k)\bar{\mathbf{R}}_s^{(\Omega)} + \mathbf{R}_n^{(\Omega)}]^{-1} S^{(\Omega)}(k)\bar{\mathbf{r}}_s^{(\Omega)} \tag{10.38}$$

and

$$S^{(\Omega)}(k) = \\ \arg \min_{S^{(\Omega)}, S^{(\Omega)} > 0} \| \mathbf{R}^{(\Omega)}(k) - \mathbf{R}_n^{(\Omega)} - S^{(\Omega)}\bar{\mathbf{R}}_s^{(\Omega)} \|_{\mathcal{F}}^2, \tag{10.39}$$

where $\mathbf{R}_n^{(\Omega)}$ is the most current evaluated noise covariance matrix from the noise detector during this time. The NSUAMA algorithm can be summarized in the following steps.

- *Step 1: Choose the number of subbands M, decimation factors D_1 and D_2 (in our algorithm $D_1 = M/2$ and $D_2 = M/4$), a block size K, a length of noise evaluation period L and a weight smoothing factor λ.*
- *Step 2: Initialize $k = 1$, the weight vector $\mathbf{w}_{opt}^{(\Omega)}(0)$ as an $I \times 1$ zero vector, the noise covariance matrix $\mathbf{R}_n^{(\Omega)}$ to an $I \times I$ identity matrix.*
- *Step 3: Calculate the matrix $\bar{\mathbf{R}}_s^{(\Omega)}$ and the vector $\bar{\mathbf{r}}_s^{(\Omega)}$ according to (10.9) and (10.11), respectively and the eigenvector $\mathbf{v}_{max}^{(\Omega)}$ that corresponds to the largest eigenvalue $\lambda_{max}^{(\Omega)}$ of $\bar{\mathbf{R}}_s^{(\Omega)}$.*
- *Step 4: Calculate $\mathbf{x}^{(\Omega)}(k)$ with D_1 decimation factor and $\mathbf{R}^{(\Omega)}(k)$ using the samples with D_2 decimation factor. The SOI PSD $S^{(\Omega)}(k)$ and the weight vector $\mathbf{w}_{opt}^{(\Omega)}(k)$ are calculated by using (10.38) and (10.39). The output is given by*

$$y^{(\Omega)}(k) = \mathbf{w}_{opt}^{(\Omega)}(k)^H \mathbf{x}^{(\Omega)}(k).$$

- *Step 5: Update $\mathbf{R}_n^{(\Omega)}(k)$ by checking the criterion (10.34) using (10.35) or (10.36). If k is within L, set $\mathbf{R}_n^{(\Omega)} = \mathbf{R}_n^{(\Omega)}(k)$.*
- *Step 6: Set $k = k + 1$ and return to Step 4 until the end of the data.*

10.5 Evaluations

10.5.1 The Simulation Scenario

The performance evaluation of the proposed microphone arrays was made in a real car hands-free situation. A six-sensor array with an inter-element distance of 5 cm was mounted on the visor at the passenger side in a Volvo station wagon. Data were gathered on a multichannel DAT-recorder with a sampling rate of 12 kHz and bandlimited to 300-3400 Hz. The car was moving at the speed of 110 km/h on a paved road.

For all the evaluations, the length of the speech signal (female) was 4 seconds long and the matrix in (10.9) and the vector (10.11) were calculated by using numerical integration according to the constrained region given in Fig. 10.1. Here, the circular constrained area of the SOI was set to be 30 cm from the center of the array with a radius of 10 cm. The only parameter in the RSCAMA structure, the weight smoothing factor λ was chosen to be $\lambda = 0.99$. As for the NSUAMA scheme, the detector threshold was set to $\varepsilon = 0.01$ and both the K and L number of samples were chosen to be 30 ms and 1 s long respectively. The decimation factor for D_1 was made oversampled and set to $M/2$ in order to reduce the aliasing effects between the adjacent subbands. The decimation factor D_2 for the covariance estimation on the other hand, was chosen to be $D_2 = M/4$ to ensure the sufficiency of data.

10.5.2 Results for RSCAMA and NSUAMA Beamformers

Figure 10.4 shows the time-domain plots of the original speech, the noisy speech at the 4^{th} microphone and the microphone array outputs for RSCAMA and NSUAMA beamformers respectively. The SNR is -7 dB and the noise level of the signal at other microphones is approximately the same as the 4^{th} microphone. Clearly, Figs. 10.4(c) and 10.4(d) show that the background noise is suppressed significantly by both beamformers respectively. The plots also suggest good timbre of the output signal as the envelope of the SOI follows that of the original SOI [Fig. 10.4(a)].

To quantify the performance of the beamformers, the following noise suppression measure is defined as,

$$NS = 10\log_{10}\left(\frac{\int_{-\pi}^{\pi}\hat{P}_{in,n}(\omega)d\omega}{\int_{-\pi}^{\pi}\hat{P}_{out,n}(\omega)d\omega}\right) - 10\log_{10}(C_d), \tag{10.40}$$

where $\hat{P}_{in,n}(\omega)$ and $\hat{P}_{out,n}(\omega)$ are the spectral power estimates of the reference sensor observation and the output respectively, when the noise is active alone. The constant C_d normalizes the performance measure such that if the SOI is attenuated by the beamformer, the measure is reduced correspondingly (i.e. normalizes the noise suppression to unity SOI gain). Table 10.1 presents

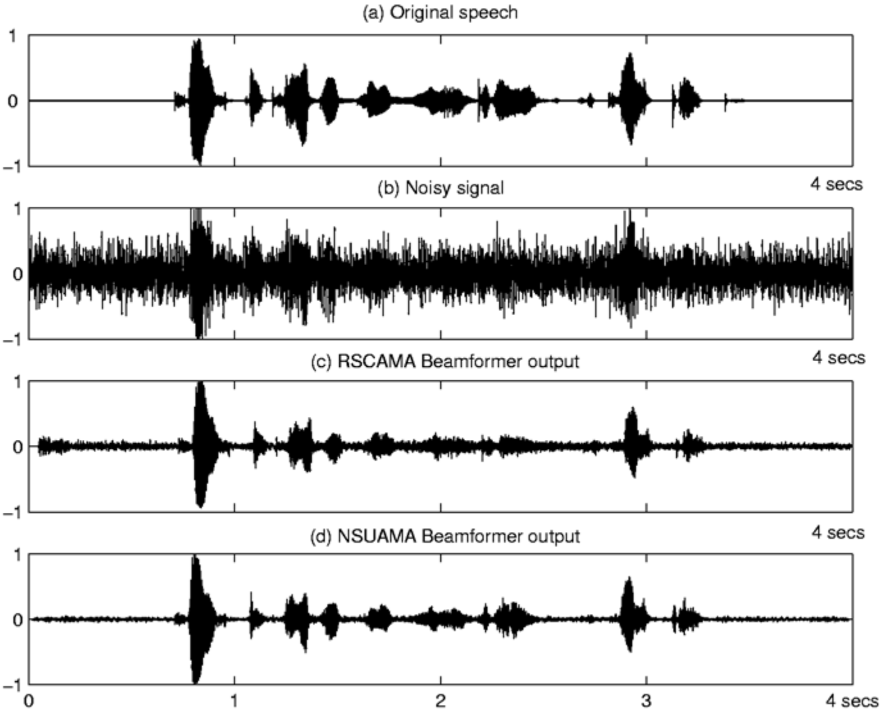

Fig. 10.4. Plots of the RSCAMA and NSUAMA data (a) clean target signal, (b) received signal, (c) RSCAMA beamformer output, and (d) NSUAMA beamformer output.

Table 10.1 Noise suppression (NS) for the RSCAMA and NSUAMA beamformers with different number of subbands.

Subbands M	NS for RSCAMA (dB)	NS for NSUAMA (dB)
16	12.8	15.7
32	14.1	18.3
64	15.2	20.1

the noise suppression levels with the number of subbands increases from 16 to 64 for both the RSCAMA and the NSUAMA schemes. The suppression levels for both the beamformers improve as the number of subbands increases. Evidently, the NSUAMA achieves $4-5$ dB noise suppression improvement over the RSCAMA structure irrespective of the number of subbands, yielding an impressive noise suppression level of 20.1 dB for the case of $M = 64$ subbands.

For completeness, Figs. 10.5(a) and 10.5(b) show the normalized output power plots of both the source and noise before and after the processing for both beamformers. From the power spectral plots, it is evident that the signal integrity of the source is maintained whilst the noise is suppressed uniformly across the frequency for both schemes. As mentioned previously, the NSUAMA algorithm achieves better noise suppression compared to the RSCAMA algorithm. More significantly, Fig. 10.5(a) reveals that the NSUAMA offers less spectral distortion to the SOI than the RSCAMA structure. This is attributed to the noise detector, which prevents the inclusion of the SOI in the update of the noise information. Nevertheless, the RSCAMA scheme has its merits as far as computational burden is concerned. For instance, in the RSCAMA algorithm, the update routine uses the matrix inversion lemma only twice (see Section 10.3.2). The NSUAMA algorithm on the other hand, requires the use of matrix inversion lemma to update all the eigenvectors of the SOI (see Section 10.4.4). However, depending on how much the space the SOI spans, it will have a few dominating eigenvectors. Therefore only half of the eigenvectors are updated in this evaluation and thus the NSUAMA scheme requires more computational requirements than the RSCAMA structure. Informal listening tests suggest good quality outputs from both the RSCAMA and the NSUAMA beamformers, with the NSUAMA offering more superior sound quality.

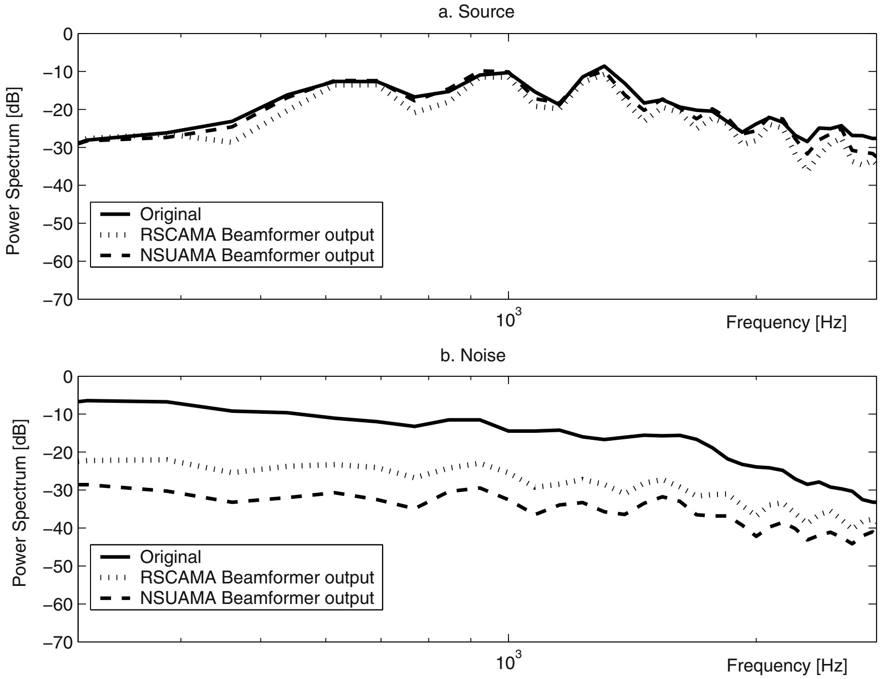

Fig. 10.5. Normalized output PSD plots for the RSCAMA and NSUAMA before and after processing of (a) source and (b) noise.

10.6 Conclusions

Two new space constrained adaptive microphone arrays with noise statistics updates have been presented. The novelty of both the structures lies in their space constraints, SOI spectral information and noise information updates. The space constraints provide robustness against steering vector errors and the update allows the noise statistics to be efficiently tracked in the Wiener solution. Also, the inclusion of the SOI PSD update in the solution offers a spectrally optimized constraint on the target signal integrity. The combination of both the PSD and space in the constraints makes full use of the available spatio-temporal domain. The major difference between the RSCAMA and NSUAMA algorithms is the manner that the SOI PSD and noise information updates are estimated. Whilst the RSCAMA is more computationally straightforward compared to the NSUAMA scheme, the NSUAMA achieves higher noise suppression capability. Results in a real hands-free car scenario show that the RSCAMA manages to achieve a good noise suppression level up to 15 dB and an impressive noise suppression of 20 dB for the NSUAMA.

References

1. Y. Grenier, "A microphone array for car environment," *Speech Communication*, vol. 12, pp. 25–39, Dec. 1993.
2. S. Nordholm, I. Claesson, and B. Bengtsson, "Adaptive array noise suppression of handsfree speaker input in cars," *IEEE Trans. on Vehicular Technology*, vol. 42, pp. 514–518, Nov. 1993.
3. N. Grbić, S. Nordholm, and A. Johansson, "Speech enhancement for hands-free terminals," in *Proc. IEEE Int. Sym. on Image and Signal Process. and Analysis*, 2001, pp. 435–440.
4. N. Grbić and S. Nordholm, "Soft constrained subband beamforming for hands-free speech enhancement," in *Proc. IEEE ICASSP*, 2002, vol. 1, pp. 885–888.
5. E. Jan and J. Flanagan, "Microphone arrays for speech processing," in *Proc. IEEE Int. Sym. on Signals, Systems, and Electronics*, 1995, pp. 373–376.
6. M. Brandstein and D. B. Ward, Editors, *Microphone Arrays: Signal Processing Techniques and Applications*, Ch. 3, pp. 39–60, Springer-Verlag, 2001.
7. B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Acoust., Speech and Signal Process. Magazine*, vol. 5, pp. 4–24, Apr. 1988.
8. H. Q. Dam, S. Nordholm, N. Grbic, and H. H. Dam, "Speech enhancement employing adaptive beamformer with recursively updated soft constraints," in *Proc. IWAENC*, 2003, pp. 307–310.
9. S. Y. Low, S. Nordholm, and N Grbić, "Subband generalized sidelobe canceller - a constrained region approach," in *Proc. IEEE Int. Workshop on Apps. of Signal Process. to Audio and Acoust.*, 2003, pp. 41–44.
10. H. Q. Dam, S. Y. Low, S. Nordholm, and H. H. Dam, "Adaptive microphone array with noise statistics updates," in *Proc. IEEE Int. Sym. on Circuits and Systems*, 2004, vol. 3, pp. 433–436.

11. O. Hoshuyama, A. Sugiyama, and A. Hirano, “A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters,” *IEEE Trans. on Signal Process.*, vol. 47, pp. 2677–2684, June 1999.
12. I. Claesson and S. Nordholm, “A spatial filtering approach to robust beamforming,” *IEEE Trans. on Antennas and Propagation*, vol. 40, pp. 1093–1096, Sept. 1992.
13. S. Nordholm, I. Claesson, and M. Dahl, “Adaptive microphone array employing calibration signals: analytical evaluation,” *IEEE Trans. on Speech and Audio Process.*, vol. 7, pp. 241–252, May 1999.
14. M. Dahl and I. Claesson, “Acoustic noise and echo cancelling with microphone array,” *IEEE Trans. on Speech and Audio Process.*, vol. 48, pp. 1518–1526, Sept. 1999.
15. N. Grbić, S. Nordholm, and A. Cantoni, “Optimal FIR subband beamforming for speech enhancement in multipath environments,” *IEEE Signal Process. Letters*, vol. 10, pp. 335–338, Nov. 2003.
16. N. Grbić, S. Nordholm, and A. Cantoni, “Limits in FIR subband beamforming for spatially spread near-field speech sources,” in *Proc. IEEE Int. Sym. on Circuits and Systems*, 2003, vol. 2, pp. 516–519.
17. H. Q. Dam, S. Y. Low, H. H. Dam, and S. Nordholm, “Space constrained beamforming with source PSD updates,” in *Proc. IEEE ICASSP*, 2004, vol. 4, pp. 93–96.
18. J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, “Filter bank design for subband adaptive microphone arrays,” *IEEE Trans. on Speech and Audio Process.*, vol. 11 , pp. 14–23, Jan. 2003.
19. K. F. C. Yiu, N. Grbić, S. Nordholm, and K. L. Teo, “Multicriteria design of oversampled uniform DFT filter banks,” *IEEE Signal Process. Letters*, vol. 11, pp. 541–544, June 2004.