AN ADAPTIVE BLIND BEAMFORMER WITH AN INTEGRATED SINGLE-CHANNEL NOISE REDUCTION METHOD FOR ROBUST REALTIME BLIND SPEECH EXTRACTION

Benny Sällberg, Nedelko Grbić, and Ingvar Claesson

Department of Signal Processing, Blekinge Institute of Technology, SE-372 25 Ronneby, Sweden
E-mails: {bsa, ngr, icl}@bth.se

ABSTRACT

The performance of single-channel temporal noise reduction methods generally deteriorate in high noise environments, whereas spatial beamformers can maintain some level of speech enhancement. This paper presents a solution where a low complexity single-channel noise reduction method is integrated into the feedback control loop of an adaptive blind beamformer with the purpose of robust blind speech extraction in high noise environments. The proposed combined system outperforms each of the individual methods with respect to signal-to-interference ratio improvement for a wide range of operating conditions, and where the loss in estimated perceptual speech quality due to the combined system is tolerably low. Furthermore, the excess processing load in a hardware solution is comparatively insignificant for the proposed extended approach.

Index Terms— Speech enhancement, Array signal processing.

1. INTRODUCTION

Classical approaches for speech enhancement in human communication are typically based on single-channel noise reduction methods, where the data from a single microphone is used to perform the noise reduction [1, 2, 3]. Inherent in single-channel methods is the necessity to trade-off the opposing design aspects of speech distortion and noise reduction. Due to the fact that single-channel techniques are limited to the temporal domain, they generally provide a high degree of speech distortion when the noise reduction level is increased, which is often needed in a high noise environment.

Blind adaptive beamforming has features that are attractive for speech enhancement in human communication. The motivation for employing a beamformer is that it uses several microphones, thus operating in the spatiotemporal domain [4], and has a higher degree of freedom as opposed to single-channel methods that utilize only the temporal domain. The inherent virtue of a blind control method in beamforming is that no knowledge about the spatiotemporal environment is needed, such as the position of the sources relative to the microphone array, or knowledge regarding the physical dimension of the array itself [5, 6, 7]. The merging of an adaptive beamformer with a blind control method results in a structure that continuously tracks sources in a changing environment [8].

This paper investigates an approach where a single-channel noise reduction method [2, 3] is integrated into the feedback control loop of a recently proposed adaptive blind beamforming technique [9, 10]. The intended application is Blind Speech Extraction (BSE), where a dominant speech source (dominant in the kurtosis measure) is extracted from an observed convolutive mixture of sources [5, 6, 7, 11]. The idea of integrating a noise reduction method into the feedback control loop of a blind beamformer is, to the best knowledge of the authors, novel. The approach provides a successful symbiosis where the spatial processing of the blind beamformer aids the temporal processing of the noise reduction method, and vice versa. This is emphasized in the evaluation where the performance of the proposed approach is increasingly better than any of the individual systems, and even better than the linear addition of the individual systems' performances. The speech quality deterioration (according to the ITU-T standard P.862, Perceptual Evaluation of Speech Quality (PESQ) [12]) of the proposed system is tolerably low, and the increased processing load due to the combined system is comparatively insignificant.

The outline of this paper is as follows: The assumed signal model, and the beamforming notation are given in Section 2. The proposed structure with a single-channel noise reduction method and an adaptive blind beamformer is presented in Section 3. Evaluation results are given in Section 4, and a summary with conclusions is provided in Section 5.

2. SIGNAL MODEL

In this paper we assume one dominant desired source (with highest Kurtosis) and one or many undesired sources. It is further assumed that the speech has a stationarity time that is much shorter than the interfering noise. The sources' relative positions to the beamformer are unknown, and the beamformer's spatial configuration is also unknown. The beamformer employs M microphones that senses the acoustical wavefield, and the recorded time signal for each microphone is denoted $x_m(t)$ for $m = \{1, 2, \ldots, M\}$ with time index t. The sampled received time signals are efficiently decomposed into a time-frequency representation, denoted $\mathbf{X}_k(n)$ where $k = \{1, 2, \ldots, K\}$ is the subband index, and n is the subband time index, using a poly-phase realization of a Discrete Fourier Transform (DFT) modulated analysis filterbank [13]. The observed convolutive mixture in the time domain corresponds to instantaneous mixtures in the frequency domain [7], and the observed subband signals are assumed to be

$$\mathbf{X}_k(n) \quad = \quad \mathbf{H}_k(n)S_k(n) + \mathbf{V}_k(n), \tag{1}$$

where $\mathbf{H}_k(n)$ represents a spatiotemporal transfer function related to the desired speech source with source signal $S_k(n)$, and $\mathbf{V}_k(n)$ represents the subband noise component, for subband index k. A linear weighting of this subband input signal using a time-varying beamformer filter vector $\mathbf{W}_k(n) = (W_{k,1}(n), W_{k,2}(n), \ldots, W_{k,M}(n))^T$, where $(\,\cdot\,)^T$ denotes the transpose, yields a subband output signal

$$Y_k(n) \quad = \quad \mathbf{W}_k^H(n)\mathbf{X}_k(n), \tag{2}$$

where $(\,\cdot\,)^H$ denotes the Hermitian transpose. The time output signal $y(t)$ is efficiently reconstructed from the subband output sig-

nals $Y_k(n)$ using a polyphase DFT modulated synthesis filterbank, matched to the analysis filterbank [13].

3. THE PROPOSED STRUCTURE

The original formulation of the adaptive blind beamformer in [9, 10] used two signals in its feedback control loop, the input signal vector $\mathbf{X}_k(n)$ and an *a-priori* beamformer output signal $\widetilde{Y}_k(n) = \mathbf{W}_k^H(n-1)\mathbf{X}_k(n)$. The idea of this paper is to integrate a low-complexity single-channel noise reduction method in the feedback control loop of this adaptive blind beamformer. The *a-priori* output signal of the adaptive blind beamformer, $\widetilde{Y}_k(n)$, is used as input to the noise reduction method. The output signal of the noise reduction method is denoted $\overline{Y}_k(n) = G_k(n)\widetilde{Y}_k(n)$, where a real valued gain function $G_k(n) \in [0,1]$ is applied in each subband to facilitate the noise reduction effect. Any phase mismatch between the input signal vector $\mathbf{X}_k(n)$ and the noise reduced signal $\overline{Y}_k(n)$ will deteriorate the beamformer's performance. The same gain function is therefore applied to the input signal vector prior to the beamformer's control loop in order to nullify this performance limitation. The input signal vector is $\overline{\mathbf{X}}_k(n) = \text{diag}\left\{G_k(n), G_k(n), \ldots, G_k(n)\right\}\mathbf{X}_k(n)$, where diag$\{\ \cdot\ \}$ produces a square diagonal matrix of size $M \times M$. This section will first present the single-channel noise reduction method, and thereafter the proposed adaptive blind beamforming method.

3.1. Single-channel Noise Reduction Method

The single-channel noise reduction method used in this paper is the Adaptive Gain Equalizer (AGE) [2, 3]. The AGE is selected due to its inherent simplicity and because it does not require a supplementary structure, like a Voice Activity Detector (VAD), which is required by many other noise reduction techniques such as the spectral subtraction type of methods [1]. The AGE operates in a subband domain and utilizes a real valued gain function $G_k(n)$ per each subband k in order to impose the noise reduction to its input signal. The input signal to the AGE method is in our case an *a-priori* output signal of the adaptive blind beamformer, $\widetilde{Y}_k(n)$, and the output signal of the AGE method is denoted $\overline{Y}_k(n) = G_k(n)\widetilde{Y}_k(n)$. Two averages, $A_{S,k}(n)$ and $A_{L,k}(n)$, are the key elements of the AGE. These averages are intended to track the speech bursts and the background noise floor level, respectively. The averages are realized using first order auto-regressive filters

$$A_{S,k}(n) = \alpha_{S,k}A_{S,k}(n-1) + (1-\alpha_{S,k})\left|\widetilde{Y}_k(n)\right|, \quad (3)$$

$$T = \alpha_{L,k}A_{L,k}(n-1) + (1-\alpha_{L,k})\left|\widetilde{Y}_k(n)\right|, \quad (4)$$

$$A_{L,k}(n) = \min\left(T, A_{S,k}(n)\right), \quad (5)$$

where T is a temporary variable, and $\alpha_{S,k}$ and $\alpha_{L,k}$ are constants associated to the integration time of the two averages $A_{S,k}(n)$ and $A_{L,k}(n)$, respectively. The function $\min(a,b)$ selects the minimal value of its two parameters a and b, and it is used to ensure that $A_{S,k}(n) \geq A_{L,k}(n)$, i.e. that $\frac{A_{S,k}(n)}{A_{L,k}(n)} \geq 1$, for all k and n. If the parameters $\alpha_{S,k}$ and $\alpha_{L,k}$ are chosen so that the integration time of $A_{S,k}(n)$ is close to speech pseudo-stationarity time (20-50 ms), and the integration time of $A_{L,k}(n)$ has a time frame matched to the slowly varying background noise (in the order of seconds), then the quotient $\frac{A_{S,k}(n)}{A_{L,k}(n)}$ will be close to unity when speech is not present, and $\frac{A_{S,k}(n)}{A_{L,k}(n)} \gg 1$ when a speech burst is present. The different temporal properties of the speech and the background noise are used to form the real valued noise reducing gain function $G_k(n)$ that continuously tracks the speech level, i.e. speech bursts, without the need of a supplementary VAD structure. The AGE utilizes the quotient of the two averages in order to construct the gain function

$$G_k(n) = f_k\left(\frac{A_{S,k}(n)}{A_{L,k}(n)}\right), \quad (6)$$

where the function $f_k(\ \cdot\ )$ inhibits the quotient to never exceed unity. The inhibiting function $f_k(\ \cdot\ )$ can typically be selected as a hard clipping function [2, 3]

$$f_k(x) = \begin{cases} \frac{x}{G_{A,k}}, & \text{if } \frac{x}{G_{A,k}} < 1 \\ 1, & \text{if } \frac{x}{G_{A,k}} \geq 1 \end{cases}, \quad (7)$$

where $G_{A,k} > 1$ is a real valued subband specific maximal allowed noise reduction level. The resulting effect is that the gain function is bounded to $\frac{1}{G_{A,k}} \leq G_k(n) \leq 1$ for all k and n. This means that if no speech is present and $A_{S,k}(n) \approx A_{L,k}(n)$ then $G_k(n) \approx \frac{1}{G_{A,k}}$ and the noise reduction is maximal, whereas if a speech burst is present and $A_{S,k}(n) \gg A_{L,k}(n)$ then $G_k(n) \approx 1$ and it becomes an all-pass filter.

3.2. Proposed Adaptive Blind Beamforming Method

A listing of 16 different definitions of the Kurtosis for complex valued data is given in [14]. One of these Kurtosis definitions was applied in [9, 10] for the beamformer's subband output signal

$$K_{Y,k} = E\left\{|Y_k(n)|^4\right\} - 2E^2\left\{|Y_k(n)|^2\right\} - \left|E\left\{Y_k(n)^2\right\}\right|^2, \quad (8)$$

where $E\{\ \cdot\ \}$ represents the expectation operator, and $(\ \cdot\ )^*$ designates the complex conjugate. $K_{Y,k}$ designates the Kurtosis value of the signal $Y_k(n)$, and it was approximated in the previous works by the time-varying function $\widehat{K}_{Y,k}(n)$, i.e. $\widehat{K}_{Y,k}(n) \approx K_{Y,k}$. The approximation in [9, 10] utilized the *a-priori* output signal $\widetilde{Y}_k(n)$ in its control loop, whereas, in this paper, a set of noise-reduced signals $\overline{Y}_k(n)$ and $\overline{\mathbf{X}}_k(n)$ are used instead, according to

$$\begin{aligned}\widehat{K}_{Y,k}(n) = \mathbf{W}_k^H(n)E\left\{\overline{\mathbf{X}}_k(n)\overline{\mathbf{X}}_k^H(n)|\overline{Y}_k(n)|^2\right\}\mathbf{W}_k(n) \\ -2E\left\{|\overline{Y}_k(n)|^2\right\}Re\left\{\mathbf{W}_k^H(n)E\left\{\overline{\mathbf{X}}_k(n)\overline{Y}_k^*(n)\right\}\right\} \\ -Re\left\{E^*\left\{\overline{Y}_k^2(n)\right\}\mathbf{W}_k^H(n)E\left\{\overline{\mathbf{X}}_k(n)\overline{Y}_k(n)\right\}\right\}. \quad (9)\end{aligned}$$

where the operator $Re\{\ \cdot\ \}$ takes the real part of its argument. The real-operator is introduced to ensure that this Kurtosis approximation is a real valued function of $\mathbf{W}_k(n)$. The objective is now to maximize this Kurtosis approximation $\widehat{K}_{Y,k}(n)$ in (9) by continuously updating the filter $\mathbf{W}_k(n)$ using information in the previous filter vector $\mathbf{W}_k(n-1)$. The introduced approximation, using the *a-priori* output signal, was inspired by the derivation of the Projection Approximation Subspace Tracking (PAST) technique in [15].

3.2.1. Newton-based Kurtosis Maximization

The approximation of the beamformer's output signal Kurtosis value in (9) is (locally) quadratic in the filter vector $\mathbf{W}_k(n)$, and the optimization of this approximative Kurtosis value, according to a modified Recursive Least Squares (RLS) method [16], follows

$$\mathbf{W}_k(n) = \frac{\mathbf{W}_k(n-1) - \gamma_k\mathbf{P}_k(n)\boldsymbol{\Delta}_k(n)}{\|\mathbf{W}_k(n-1) - \gamma_k\mathbf{P}_k(n)\boldsymbol{\Delta}_k(n)\|_2}, \quad (10)$$

where

$$\mathbf{\Delta}_k(n) \quad = \quad 2a_k(n)\mathbf{A}_k(n) + b_k^*(n)\mathbf{B}_k(n). \qquad (11)$$

The parameter $\gamma_k \in [0,1]$ is introduced in order to control the fluctuations in the filter weights due to the random input data. The normalization in (10) has been incorporated in order to avoid the trivial solution, $\mathbf{W}_k(n) = \mathbf{0}$. The variables $a_k(n)$, $b_k(n)$, $\mathbf{A}_k(n)$, and $\mathbf{B}_k(n)$ are herein implemented using first order auto-regressive averages to approximate the various statistical measures in (9), as

$$\begin{aligned} a_k(n) &= \lambda_k a_k(n-1) + (1-\lambda_k)\left|\overline{Y}_k(n)\right|^2, & (12)\\ b_k(n) &= \lambda_k b_k(n-1) + (1-\lambda_k)\,\overline{Y}_k^2(n), & (13)\\ \mathbf{A}_k(n) &= \lambda_k \mathbf{A}_k(n-1) + (1-\lambda_k)\,\overline{\mathbf{X}}_k(n)\overline{Y}_k^*(n), & (14)\\ \mathbf{B}_k(n) &= \lambda_k \mathbf{B}_k(n-1) + (1-\lambda_k)\,\overline{\mathbf{X}}_k(n)\overline{Y}_k(n), & (15) \end{aligned}$$

where the parameter $\lambda_k \in [0,1]$ controls the convergence rate (and the source tracking performance) of the method. The matrix $\mathbf{P}_k(n)$ is computed according to the matrix inversion lemma [16] as

$$\begin{aligned} \mathbf{P}_k(n) &= \lambda_k^{-1}\mathbf{P}_k(n-1) \\ &\quad - \frac{\mathbf{P}_k(n)\overline{\mathbf{X}}_k(n)\overline{\mathbf{X}}_k^H(n)|\overline{Y}_k(n)|^2\mathbf{P}_k(n)}{\lambda_k^2 + \lambda_k\left|\overline{Y}_k(n)\right|^2\overline{\mathbf{X}}_k^H(n)\mathbf{P}_k(n)\overline{\mathbf{X}}_k(n)}. \qquad (16) \end{aligned}$$

A suitable initialization of this method is $\mathbf{P}_k(0) = \mathbf{I}_M$ where $\mathbf{I}_M$ is the $(M \times M)$ identity matrix, $a_k(0) = b_k(0) = 0$, $\mathbf{A}_k(0) = \mathbf{B}_k(0) = (0,0,\ldots,0)^T$, and $\mathbf{W}_k(0) = (1,1,\ldots,1)^T$.

4. EVALUATION

The performance of the proposed method is analyzed using an off-line setting with real measured data and two microphones. This setting allows comparison between the single-channel method, the blind adaptive beamformer, and the combined proposed structure.

4.1. Evaluation Measures

A measure of the Signal to Interference Ratio (SIR) improvement, and an objective measure that reflects the perceptual speech quality, through the Perceptual Evaluation of Speech Quality (PESQ) [12] measure, is used to evaluate the proposed approach. The filter weights at each iteration are stored, and used for filtering the original convolved, but unmixed, source signals. This enables direct access to the evaluation measures. The SIR improvement performance measure, denoted P_{SIR}, is defined as

$$P_{\text{SIR}} \quad = \quad \frac{\widehat{\text{Var}}\{y_s(t)\}\widehat{\text{Var}}\{x_{1;v}(t)\}}{\widehat{\text{Var}}\{x_{1;s}(t)\}\widehat{\text{Var}}\{y_v(t)\}}, \qquad (17)$$

where $\widehat{\text{Var}}\{\ \cdot\ \}$ denotes an estimator of variance, $y_s(t)$ and $y_v(t)$ represent the speech and noise components of the enhanced output signal, and similarly, the signals $x_{1;s}(t)$ and $x_{1;v}(t)$ represent the speech and noise components of the first microphone signal. The first microphone is acting as a reference in the analysis.

The PESQ standard is an automated method for objective assessment of perceptual sound quality, and it uses a perceptual model of how sound quality is perceived by humans. The PESQ computes a perceptual model for a clean received reference speech signal $x_{1;s}(t)$, and a perceptual model for the processed output speech component $y_s(t)$. The perceptual difference between the clean received speech signal and the processed speech signal is mapped on the Mean Opinion Score (MOS), which yields a value between one and five, where the score one indicates a bad speech quality and the score five is used to indicate an excellent speech quality.

Program block	Est. proc. load
Dual-channel analysis filter bank	25.4 %
Dual-channel adaptive blind beamformer	58.0 %
Single-channel noise reduction method	2.9 %
Single-channel synthesis filter bank	13.7 %

Table 1. *Estimated processing load of a dual-channel implementation of the proposed method on an ADSP-21262 DSP. The total program package requires* 13.5 *% of the DSP's processing resources.*

4.2. System Configuration

The filterbank configuration used $K = 64$ subbands, with two times oversampling. The prototype filter was designed using the window method with a Hamming window. The method parameters λ_k, γ_k, $\alpha_{S,k}$, and $\alpha_{L,k}$ were set such that their integration times were 60 ms, 30 ms, 60 ms, and 2 s, respectively. The maximal allowed attenuation in the AGE method was $G_{A,k} = 15$ dB (i.e., $10^{15/20}$). It should be noted that these parameter values were selected empirically, and the same values were used for all subbands. A future analysis should encompass the influence of various parameter values on the method's performance in order to find their optimal values.

4.3. Signal Configuration

Human speech (male and female) was sent through a loudspeaker and recorded using two microphones separated by 5 cm in an office room (reverberation time $RT_{60} = 130$ ms) with sampling frequency 8 kHz. Previously recorded ferry engine noise and factory noise were subsequently emitted and recorded using the same setup. The speech signal is then mixed at various levels of SIR with each of the two interfering noise signals.

4.4. Estimated Processing Load

The estimated processing load[1] for a realtime Digital Signal Processor (DSP) implementation of the proposed method on an ADSP-21262 type DSP is provided in Table. 1. As can be seen from this analysis is that the filterbanks (analysis and synthesis) together with the adaptive blind beamformer comprises the lions share of the required processing load, and the additional noise reduction method require merely 2.9 % of the overall processing load. The total program package requires 13.5 % of the DSP's available processing resources.

4.5. Evaluation Results

The evaluated performance of the single-channel AGE technique, the original adaptive blind beamformer, and the proposed combined structure are presented in Fig. 1 and Fig. 2 for the cases when speech is mixed with ferry engine noise and speech is mixed with factory noise. The results indicate that the combined system outperforms each of the individual systems with respect to SIR improvement.

[1]The estimation of the processing load for the proposed method is performed in a simulation environment provided by the DSP manufacturer. The test-software is written in C language, where the compiler is set to operate at the highest optimization level.

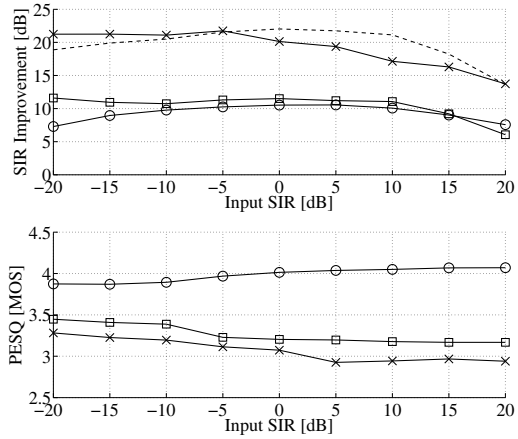

Fig. 1. Evaluated performance when speech is mixed with ferry engine noise for the single-channel AGE method (circles), the adaptive blind beamformer (squares), and the proposed combined structure (crosses). The linear addition of SIR improvement of the AGE method and the adaptive blind beamformer (dashed).

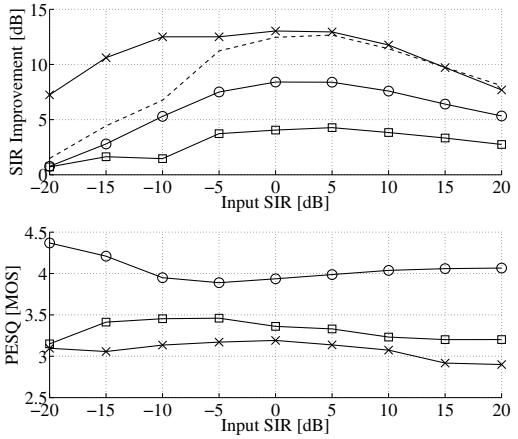

Fig. 2. Evaluated performance when speech is mixed with factory noise for the single-channel AGE method (circles), the adaptive blind beamformer (squares), and the proposed combined structure (crosses). The linear addition of SIR improvement of the AGE method and the adaptive blind beamformer (dashed).

In some cases, the performance of the combined system also outperforms the linear addition of performance of each of the two subsystems. This indicates a successful symbiosis, where the spatial processor aids the temporal processor, and vice versa. In addition, the perceptual speech degradation of the combined system never falls below 0.3 MOS-units in relation to the blind beamformer's MOS, and this further motivates the proposed solution.

5. SUMMARY AND CONCLUSIONS

This paper presents the integration of a single-channel noise reduction technique in the feedback control loop of a recently proposed adaptive blind beamformer. The proposed combined system provides a SIR improvement that outperforms the individual systems. In some cases, the performance of the combined system also outperforms the linear addition of performance of each of the two subsystems. The introduced degradation in perceptual speech quality is tolerably low, and the extra processing load due to the extended structure is small, and this further motivates the proposed combined structure. The current method parameters were selected empirically, and an important part for future research is the design of optimal parameter values that will further improve the method's performance. The proposed approach has been successfully validated in realtime using a DSP implementation with the purpose of blind speech extraction in high-noise human communication applications.

6. REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech and Sig. Proc.*, vol. ASSP-27, pp. 113–120, April 1979.

[2] N. Westerlund, M. Dahl, and I. Claesson, "Speech enhancement for personal communication using an adaptive gain equalizer," *Elsevier Signal Processing*, vol. 85, no. 6, pp. 1089–1101, 2005.

[3] B. Sällberg, H. Åkesson, M. Dahl, and I. Claesson, "A mixed analog-digital hybrid for speech enhancement purposes," *IEEE ISCAS*, vol. 1, pp. 852–855, 2005.

[4] D. Johnson and D. Dudgeon, *Array Signal Processing – Concepts and Techniques*, Prentice Hall, 1993.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, 2001.

[6] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*, John Wiley and Sons, 2003.

[7] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Elsevier Neurocomputing*, vol. 22, no. 1–3, pp. 21–34, 1998.

[8] Z. Ding, "A new algorithm for automatic beamforming," *Asilomar Conference on Signals, Systems and Computers*, vol. 2, no. 25, pp. 689–693, 1991.

[9] B. Sällberg, N. Grbić, and I. Claesson, "Online maximization of subband kurtosis for blind adaptive beamforming in realtime speech extraction," *IEEE DSP*, 2007.

[10] B. Sällberg, N. Grbić, and I. Claesson, "Online blind speech extraction based on a locally quadratic kurtosis criteria and a preprocessing automatic gain controller," *IEEE ELMAR*, 2007.

[11] N. Grbić, X.-J. Tao, S. Nordholm, and I. Claesson, "Blind signal separation using overcomplete subband representation," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 5, pp. 524–533, 2001.

[12] ITU-T p.862, *Perceptual evaluation of speech quality (PESQ)*.

[13] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, 1993.

[14] C. Nikias and A. Petropulu, *Higher-Order Spectral Analysis - A Nonlinear Signal Processing Framework*, Prentice Hall, 1993.

[15] B. Yang, "Projection approximation subspace tracking," *IEEE Trans. on Signal Proc.*, vol. 43, no. 1, pp. 95–107, 1995.

[16] S. Haykin, *Adaptive Filter Theory*, John Wiley and Sons, 2002.