

# Online Blind Speech Extraction based on a Locally Quadratic Kurtosis Criteria and a Preprocessing Automatic Gain Controller

Benny Sällberg, Nedelko Grbić and Ingvar Claesson

Department of Signal Processing  
Blekinge Institute of Technology  
SE-372 25 Ronneby, Sweden  
E-mail: *bsa@bth.se*

**Abstract** - This paper focuses on realtime speech extraction using blind adaptive beamforming. The speech extraction is carried out using an approximation of the kurtosis measure in a subband domain. The introduced kurtosis approximation is an improvement of a recently proposed approximation technique where a locally quadratic criterion was solved at each iteration. The improvement introduced in this paper regards an approach to normalize this same criterion using a pre-processing Automatic Gain Control unit, and thereby making the algorithm invariant to input signal scales. The proposed method outperforms the recent technique in terms of Signal to Interference Ratio improvement. In addition, the increased memory consumption and processing load due to the proposed improvement is comparably low and this is often desirable in a realtime Digital Signal Processor (DSP) implementation. Further, a real-time implementation of the method is conducted and results with real data is presented.

## 1 INTRODUCTION

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, and has a higher degree of freedom as opposed to corresponding single channel techniques that only utilize the temporal domain [1]. The inherent virtue of a blind control algorithm 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. The merging of an adaptive beamformer with a blind control algorithm results in a structure that continuously tracks sources in a changing environment [2, 3].

This paper presents an improvement to a recently proposed approach of approximating subband kurtosis using a locally quadratic criteria [3]. The improvement proposed in this paper regards a pre-processing Automatic Gain Control (AGC) stage that normalizes the input signals prior to the adaptive beamformer's filter update equation, thus yielding a scale-invariant solution. 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 [3, 4, 5, 6]. The proposed approach stands in contrast to Blind Signal Separation (BSS) in which all dominant sources, or groups of sources, are separated, see for example [7].

The outline of this paper is as follows: The assumed signal model, and the beamforming notation are given in Section 2. The improved kurtosis approximation is provided in Section 3, and a Newton-based

maximization thereof is given in Section 4. A realtime Digital Signal Processor (DSP) implementation is presented in Section 5 to illustrate the method's robustness and performance. Evaluation results are given in Section 6. Conclusions and topics for future work are provided in Section 7.

## 2 SIGNAL MODEL

In this paper we assume one dominant desired speech source and  $L - 1$  undesired sources (with lower kurtosis). 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 received time signals for each microphone  $m = \{1, 2, \dots, M\}$  at time index  $t$  is denoted  $x_m(t)$ , and the set of all microphone signals is represented by the vector  $\mathbf{x}(t) = (x_1(t), \dots, x_M(t))^T$ , where the superscript  $(\ )^T$  denotes the transpose. The received time signals are efficiently decomposed into a time-frequency representation, denoted  $\mathbf{X}_k(n) = (X_{1,k}(n), \dots, X_{M,k}(n))^T$  where  $k = \{1, 2, \dots, K\}$  is the subband index and  $n$  is the block time index, using a polyphase realization of a Discrete Fourier Transform (DFT) modulated analysis filterbank [8]. The convolutive mixture in the time domain corresponds to instantaneous mixtures in the frequency domain [6], and the observed subband signal, for subband index  $k$ , is assumed to be

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

where  $\mathbf{H}_k(n)$  represents a matrix of channels, the vector  $\mathbf{S}_k(n)$  contains the subband source signals, and  $\mathbf{V}_k(n)$  represents the subband noise component. The desired speech source's signal is the first signal  $S_k^1(n)$  in  $\mathbf{S}_k(n)$ , and the other signals  $S_k^i(n)$  ( $i \in \{2, \dots, L\}$ )

in  $\mathbf{S}_k(n)$  are considered as interferences. A linear weighting of this subband input signal using a time-varying beamformer filter vector  $\mathbf{W}_k(n) = (W_{k,1}(n), \dots, W_{k,M}(n))^T$ , denoted a filter-and-sum beamformer [1], yields a subband output signal

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

where the superscript  $(\ )^H$  denotes the Hermitian transpose. The time domain output signal  $y(t)$  is reconstructed from the subband output signals  $Y_k(n)$  using a polyphase DFT modulated synthesis filterbank, matched to the analysis filterbank [8].

### 3 APPROXIMATION OF SUBBAND KURTOSIS

A listing of 16 different definitions of the kurtosis for complex valued data is given in [9]. One of these kurtosis definitions was applied in [3] for the beamformer's subband output signal

$$K_{Y;k} = E \{ |Y_k(n)|^4 \} - 2E^2 \{ |Y_k(n)|^2 \} - |E \{ Y_k(n)^2 \}|^2, \quad (3)$$

where  $E\{ \}$  represents the expectation operator, and  $K_{Y;k}$  designates the kurtosis value of the signal  $Y_k(n)$ .

#### 3.1 Improvement - Scale Invariance

While a linear scale  $c$  applied to the beamformer's input signal  $c\mathbf{X}_k(n)$  yields a non-linear change in the beamformer output signal's kurtosis value  $c^4 K_{Y;k}$ , it may be concluded that the kurtosis criterion in (3) is not invariant to input signal scale. This scale variant behavior is in many cases undesirable, while it influences the predictability of the algorithm's behavior in a realtime implementation. An Automatic Gain Control (AGC) unit is proposed in this paper as a remedy to this shortfall. The AGC is applied to the input signal before the adaptive beamformer's filter update equation. It is stressed that the AGC is not used in the filtering part of the algorithm, and does thereby not influence the output signal directly, as indicated in Fig. 1. The AGC normalizes the input data, denoted

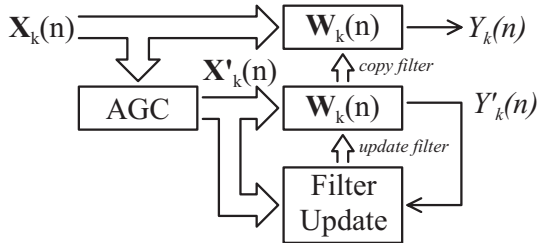


Figure 1: An Automatic Gain Control (AGC) in the control loop of an adaptive beamformer.

$\mathbf{X}'_k(n)$ , by the square root of its mean power

$$\mathbf{X}'_k(n) = \frac{\mathbf{X}_k(n)}{\sqrt{E \{ \mathbf{X}_k^H(n) \mathbf{X}_k(n) \}}}. \quad (4)$$

According to [3]; a normalized output signal  $Y'_k(n) = \mathbf{W}_k^H(n) \mathbf{X}'_k(n)$  and a normalized *a-priori* output signal  $\tilde{Y}'_k(n) = \mathbf{W}_k^H(n-1) \mathbf{X}'_k(n)$  are used in a scale invariant kurtosis criterion,  $\hat{K}_{Y';k}(n)$ , according to

$$\begin{aligned} \hat{K}_{Y';k}(n) = & E \left\{ |Y'_k(n)|^2 |\tilde{Y}'_k(n)|^2 \right\} \\ & - 2E \left\{ |\tilde{Y}'_k(n)|^2 \right\} Re \left\{ E \left\{ Y'_k(n) \tilde{Y}'_k(n) \right\} \right\} \\ & - Re \left\{ E^* \left\{ \tilde{Y}'_k(n) \right\} E \left\{ Y'_k(n) \tilde{Y}'_k(n) \right\} \right\}, \quad (5) \end{aligned}$$

where the operator  $Re\{ \}$  takes the real part of its argument, and  $(\ )^*$  designates the complex conjugate. The real-operator is introduced to ensure that this kurtosis approximation is real valued and that it forms an analytic function of  $\mathbf{W}_k(n)$ . The objective is now to maximize this kurtosis approximation  $\hat{K}_{Y';k}(n)$  in (5) by continuously updating the filter  $\mathbf{W}_k(n)$  using information in the previous filter vector  $\mathbf{W}_k(n-1)$ .

### 4 NEWTON-BASED KURTOSIS MAXIMIZATION

The approximation of the beamformer output signal's kurtosis value in (5) is (locally) quadratic in the filter vector  $\mathbf{W}_k(n)$ . Optimization of the approximative kurtosis value follows Newton's method [10]. A feasible implementation is achieved through the introduction of a set of suitable approximations, where auto-regressive averaging is common in this kind of approximations [4, 5]

$$\mathbf{P}_k(n) \approx E^{-1} \left\{ \mathbf{X}'_k(n) \mathbf{X}'_k{}^H(n) \left| \tilde{Y}'_k(n) \right|^2 \right\}, \quad (6)$$

$$a_k(n) \approx E \left\{ \left| \tilde{Y}'_k(n) \right|^2 \right\}, \quad (7)$$

$$b_k(n) \approx E \left\{ \tilde{Y}'_k{}^2(n) \right\}, \quad (8)$$

$$\mathbf{A}_k(n) \approx E \left\{ \mathbf{X}'_k(n) \tilde{Y}'_k{}^*(n) \right\}, \quad (9)$$

$$\mathbf{B}_k(n) \approx E \left\{ \mathbf{X}'_k(n) \tilde{Y}'_k(n) \right\}, \quad (10)$$

$$\eta_k(n) \approx E \left\{ \mathbf{X}'_k{}^H(n) \mathbf{X}_k(n) \right\}. \quad (11)$$

The matrix  $\mathbf{P}_k(n)$  in (6) is recursively updated according to the matrix inversion lemma as [10]

$$\begin{aligned} \mathbf{P}_k(n) = & \lambda_k^{-1} \mathbf{P}_k(n-1) \\ & - \frac{\mathbf{P}_k(n-1) \mathbf{X}'_k(n) \mathbf{X}'_k{}^H(n) \left| \tilde{Y}'_k(n) \right|^2 \mathbf{P}_k(n-1)}{\lambda_k^2 + \lambda_k \left| \tilde{Y}'_k(n) \right|^2 \mathbf{X}'_k{}^H(n) \mathbf{P}_k(n-1) \mathbf{X}'_k(n)}, \quad (12) \end{aligned}$$

where the parameter  $\lambda_k \in [0, 1]$  controls the convergence rate (i.e. the tracking performance) of the method. First order auto-regressive averages are used in (7) to (11)

$$a_k(n) = \lambda_k a_k(n-1) + (1 - \lambda_k) \left| \tilde{Y}'_k(n) \right|^2, \quad (13)$$

$$b_k(n) = \lambda_k b_k(n-1) + (1 - \lambda_k) \tilde{Y}'_k{}^2(n), \quad (14)$$

$$\mathbf{A}_k(n) = \lambda_k \mathbf{A}_k(n-1) + (1 - \lambda_k) \mathbf{X}'_k(n) \tilde{Y}'_k{}^*(n), \quad (15)$$

$$\mathbf{B}_k(n) = \lambda_k \mathbf{B}_k(n-1) + (1 - \lambda_k) \mathbf{X}'_k(n) \tilde{Y}'_k(n), \quad (16)$$

$$\eta_k(n) = \lambda_k \eta_k(n-1) + (1 - \lambda_k) \mathbf{X}'_k{}^H(n) \mathbf{X}_k(n). \quad (17)$$

The input power estimate in (17) is used to normalize the input signal  $\mathbf{X}'_k(n) = \mathbf{X}_k(n) / \sqrt{\eta_k(n)}$ . The filter update equation, according to Newton's method (with an additional normalization step), becomes

$$\mathbf{W}_k(n) = \frac{\mathbf{W}_k(n-1) - \gamma_k \mathbf{P}_k(n) \Delta_k(n)}{\|\mathbf{W}_k(n-1) - \gamma_k \mathbf{P}_k(n) \Delta_k(n)\|_2}, \quad (18)$$

where

$$\Delta_k(n) = 2a_k(n) \mathbf{A}_k(n) + b_k^*(n) \mathbf{B}_k(n). \quad (19)$$

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 (18) has been incorporated in order to avoid the trivial solution,  $\mathbf{W}_k(n) = \mathbf{0}$ . However, this normalization is not optimal with respect to the spectral whitening of the enhanced speech since only  $\|\mathbf{W}_k^H(n)\|_2 = 1$  is ensured, and the term  $|\mathbf{W}_k^H(n) \mathbf{H}_k^1(n)|$  ( $\mathbf{H}_k^1(n)$  is the transfer function vector related to the desired speech source) is thereby not guaranteed to equal unity. It may be noted that, except for the normalization, the filter update in (18) is similar to the well known Recursive Least Squares (RLS) algorithm [10]. A suitable initialization of this algorithm is  $\mathbf{P}_k(0) = \mathbf{I}_M$  where  $\mathbf{I}_M$  is the  $(M \times M)$  identity matrix,  $a_k(0) = b_k(0) = 0$ ,  $\mathbf{A}_k(0) = \mathbf{B}_k(0) = (0, 0, \dots, 0)^T$ , and  $\mathbf{W}_k(0) = (1, 1, \dots, 1)^T$ .

## 5 REALTIME IMPLEMENTATION

The implementation of the proposed method is made on a floating point DSP named ADSP-21262 from Analog Devices. This is a high-performance DSP that

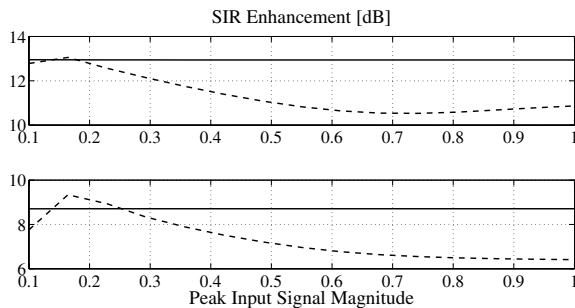


Figure 2: Mean SIR enhancement at different peak input signal magnitudes for speech and ferry engine noise signal (upper), and speech and factory noise signal (lower). The improved scale-invariant algorithm (solid), and the original algorithm [3] (dashed).

supports effective parallel computations through the Single Instruction Multiple Data (SIMD) mode, suitable for vector-based operations. The algorithm is efficiently implemented using a transformed approach to reduce the overhead. A rule-of-thumb for the transformed approach is to perform all computations so that the largest data dimension is used in the element-wise vector operations. The transformation approach reduces the number of computations in this DSP platform by 7.5 times for a typical vector operation [11]. The same hardware platform was used in [3] to realize the predecessor to the proposed blind beamformer.

## 5.1 System Configuration

The filterbank configuration used  $K = 64$  subbands, with four times oversampling. The prototype filter was designed using the window method with a Hamming window. This configuration was selected as a tradeoff between the audio quality performance and the introduced signal delay, measured to 8.5 ms. The algorithmic parameters were selected such that the integration time of the  $\lambda_k$  parameter was 0.46 s, and the integration time of the  $\gamma_k$  parameter was 5 ms.

## 6 EVALUATION RESULTS

The performance of the improved proposed method is first analyzed using an offline setting with real measured data. This setting allows for comparison with the recently proposed method [3]. The realtime implementation is analyzed using short-term power estimates where the algorithm may be switched on and off.

### 6.1 Evaluation Measures

The mean Signal to Interference Ratio (SIR) enhancement  $Q$  is analyzed in the time domain using

$$Q = \frac{\widehat{\text{Var}}\{y_S(t)\} \widehat{\text{Var}}\{x_V(t)\}}{\widehat{\text{Var}}\{y_V(t)\} \widehat{\text{Var}}\{x_S(t)\}}, \quad (20)$$

where  $\widehat{\text{Var}}\{\cdot\}$  designates an estimator of variance, and the signals  $x_S(t)$ ,  $x_V(t)$ ,  $y_S(t)$ , and  $y_V(t)$  represents the components of the input and output signals that are related to the speech  $S$  and interference  $V$ , respectively.

### 6.2 Offline Results using Real Data

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. A previously recorded ferry engine noise signal and a factory noise signal were subsequently emitted and recorded using the same setup, although from a different direction. The speech signal was then mixed at 0 dB SIR level with each of the two interfering noise signals. Each mixed signal was thereafter scaled such that the peak input signal magnitude corresponded to a certain level - this is to test the algorithm's ability to extract speech in different operating conditions.

The mean SIR enhancement for various input signal magnitudes is presented in Fig. 2. It can be concluded from this figure that, the proposed algorithm is invariant to different input signal magnitudes as opposed to the original algorithm [3]. Furthermore, the performance of the proposed algorithm exceeds that of the original method in most cases. Some spectral whitening can be noticed in the extracted speech signal.

### 6.3 Online Results using Real Data

The online evaluation includes prerecorded signals consisting of male speech spatially added with music and it is provided in [12]. The speaker and the music was recorded in an office room using two microphones, where the distance between the sources and the microphones is 60 cm in a square. The two input channels are used as input to the proposed realtime DSP implementation, and the resulting output time signal is presented in Fig. 3 with a corresponding short-term power estimate. The proposed method is activated after 6.8 s. The two input channels are exchanged momentarily at 13.7 s in order to illustrate the methods capability to track signals in a non-stationary environment. The SIR enhancement is 15 dB to 20 dB in this case.

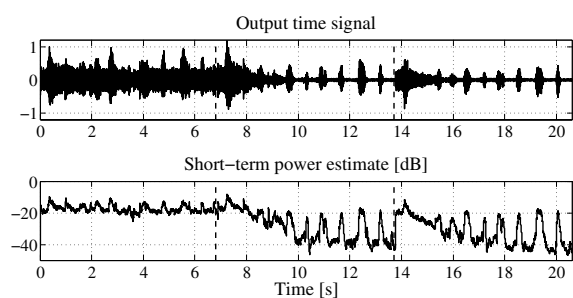


Figure 3: Output time signal for speech and music (upper), and short-term power estimate (lower). The proposed method is activated at 6.8 s, and the input channels are exchanged at 13.7 s.

## 7 CONCLUSIONS AND FUTURE WORK

This paper presents an improvement to a blind beamformer based on the well known kurtosis maximization strategy. The novelty of this contribution lies in a normalization approach of the recently proposed (locally) quadratic kurtosis criteria [3]. The improved criteria is continuously updated online according to a Newton based search method. The method successfully extracts speech in convolutive mixtures of speech and ferry engine noise, speech and factory noise, and speech and music. It is noticeable that, except for some spectral whitening, the speech distortion is perceptually very low. Furthermore, the performance of the proposed method is superior to the recently proposed method for a variety of input signal scales. This method shows promising performance for speech enhancement in real environments, however some ques-

tions are still open for future research:

The strategy for normalization of the adaptive filter used here, (18), is not optimal with respect to the spectral whitening of the enhanced speech. Other normalization strategies may reduce this undesired spectral whitening.

A rigorous analysis of the spatiotemporal behavior of this proposed method is required in order to relate and compare its performance to existing state-of-the-art solutions.

## REFERENCES

- [1] D. Johnson and D. Dudgeon. *Array Signal Processing – Concepts and Techniques*. Prentice Hall, 1993.
- [2] Z. Ding. A new algorithm for automatic beamforming. *Asilomar Conference on Signals, Systems and Computers*, 2(25):689–693, 1991.
- [3] B. Sällberg, N. Grbić, and I. Claesson. Online maximization of subband kurtosis for blind adaptive beamforming in realtime speech extraction. *IEEE DSP*, 2007.
- [4] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.
- [5] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*. John Wiley and Sons, 2003.
- [6] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Elsevier Neurocomputing*, 22(1–3):21–34, 1998.
- [7] N. Grbić, X.-J. Tao, S. E. Nordholm, and I. Claesson. Blind signal separation using overcomplete subband representation. *IEEE Trans. on Speech and Audio Proc.*, 9(5):524–533, 2001.
- [8] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.
- [9] C. Nikias and A. Petropulu. *Higher-Order Spectral Analysis - A Nonlinear Signal Processing Framework*. Prentice Hall, 1993.
- [10] S. Haykin. *Adaptive Filter Theory*. John Wiley and Sons, 2002.
- [11] Z. Yermeche, B. Sällberg, N. Grbić, and I. Claesson. Real-time dsp implementation of a subband beamforming algorithm for dual microphone speech enhancement. *IEEE ISCAS*, 2007.
- [12] T. W. Lee. Blind source separation: Audio examples. [Online] <http://www.cnl.salk.edu/~tewon/Blind/blind.audio.html>, Feb. 17 2007.