ONLINE MAXIMIZATION OF SUBBAND KURTOSIS FOR BLIND ADAPTIVE BEAMFORMING IN REALTIME 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

This paper presents a method for blind beamforming with application in realtime speech extraction in a non-stationary environment. The blind beamforming is carried out using an online kurtosis maximization approach where the optimization is based on Newton's method. The main novelty of the paper lies in the formulation of the subband kurtosis approximation, where a locally quadratic criterion is solved at each iteration. Further, a real-time digital signal processor (DSP) implementation of the method is conducted and results with real data is presented.

Index Terms— Blind Speech Extraction, Adaptive Beamforming, Kurtosis Maximization.

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 one channel techniques that utilize only 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].

This paper presents a novel approach of approximating subband kurtosis using a locally quadratic criteria. 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]. The proposed method in this paper is an extension of [6] where the new formulation of the kurtosis measure provides significantly faster convergence. This 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, 8].

The outline of this paper is as follows: The background to the proposed method is presented in Section 2. The assumed signal model, and the beamforming notation are given in Section 3. The proposed kurtosis approximation is provided in Section 4, and a gradient based maximization thereof is given in Section 5. A real-time Digital Signal Processor (DSP) implementation is presented in Section 6 to illustrate the method's robustness and performance. Evaluation results are given in Section 7, and a summary with conclusions and topics for future research are provided in Section 8.

2. BACKGROUND

One fundamental approach for ICA is to exploit the non-Gaussianity of the observed signals [3, 4]. This may be achieved by utilizing the kurtosis as an estimate of Gaussianity. The estimated kurtosis of a beamformer's output signal can then control the update of a beamformer so as to maximize the output kurtosis [2]. By doing so, it is intended that the beamformer blocks the noise signal(-s) (low kurtosis) and passes the speech component (high kurtosis).

In order to successfully apply blind algorithms in the application of speech extraction it is necessary to construct a time-varying and robust structure to allow for source movement tracking and to compensate for a changing background noise environment. Many of the fundamental ICA methods require stationary instantaneous mixtures (at least for the duration of each data batch) to yield robust performance [3, 9]. Hence, some of these methods are not suitable in a real environment where the spatiotemporal settings may vary rapidly. The main focus in this paper is therefore on Newton's method utilizing an approximative kurtosis estimate in each subband, allowing for fast source tracking in such a time-varying environment.

In many BSE/BSS algorithms there is a need for a whitening (decorrelation) of the input signals in order to improve the convergence speed [3, 4]. This is often done in a pre-processing step for instance by using the Principle Component Analysis (PCA), [10, 11]. However, in a non-stationary environment, such as the environment intended here, the tracking performance of the PCA method may potentially deteriorate the performance of the following online ICA algorithm. Therefore, the proposed method does not utilize pre-whitening and operates directly upon the observed signals using an online update of the beamformer filter.

The permutation ambiguity of subband ICA may pose a potential problem in many applications and may therefore require special attention [8]. However, the permutation problem is not considered an issue in this paper while we focus on extraction of one dominant speech source (dominant in the kurtosis measure) that is convolved and mixed with one or many noise sources with low kurtosis.

3. SIGNAL MODEL

In this paper we assume one dominant desired source and one or many 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 recorded 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 $\mathbf{x}(t) = (x_1(t), x_2(t), \dots, x_M(t))^T$, where the superscript $(\)^T$ denotes the transpose. The received time signals

are efficiently decomposed into a time-frequency representation, denoted $\mathbf{X}_k(n)$ where $k = \{1, 2, \dots, K\}$ is the subband index and n is the block time index, using a polyphase realization of Discrete Fourier Transform (DFT) modulated analysis filterbank [12]. The observed convolutive mixture in the time domain corresponds to instantaneous mixtures in the frequency domain [5], and the observed subband signals are assumed to be

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

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 $\mathbf{W}_k(n) = (W_{k,1}(n), W_{k,2}(n), \dots, W_{k,M}(n))^T$, 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 $(\cdot)^H$ denotes the Hermitian transpose. The time domain output signal $y(t)$ is efficiently reconstructed from the subband output signals $Y_k(n)$ using a polyphase DFT modulated synthesis filterbank, matched to the analysis filterbank [12].

4. APPROXIMATION OF SUBBAND KURTOSIS

A listing of 16 different definitions of the kurtosis for complex valued data is given in [13]. One of these kurtosis definitions is applied herein 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 $E\{\cdot\}$ 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 is approximated in this paper by the time-varying function $\hat{K}_{Y;k}(n)$, i.e. $K_{Y;k} \approx \hat{K}_{Y;k}(n)$. This approximation utilizes an *a-priori* output signal denoted $\tilde{Y}_k(n) = \mathbf{W}_k^H(n-1)\mathbf{X}_k(n)$, and is formulated as

$$\begin{aligned} \hat{K}_{Y;k}(n) = & E \{ |Y_k(n)|^2 |\tilde{Y}_k(n)|^2 \} \\ & - 2E \{ |\tilde{Y}_k(n)|^2 \} Re \{ E \{ Y_k(n) \tilde{Y}_k^*(n) \} \} \\ & - Re \{ E^* \{ \tilde{Y}_k^2(n) \} E \{ Y_k(n) \tilde{Y}_k(n) \} \}, \quad (4) \end{aligned}$$

where the operator $Re\{\cdot\}$ takes the real part of its argument. The real-operator is introduced to ensure that this kurtosis approximation is 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 (4) 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, is inspired by the derivation of the Projection Approximation Subspace Tracking (PAST) technique in [10].

5. GRADIENT-BASED KURTOSIS MAXIMIZATION

We may rewrite the kurtosis approximation in (4) using (2) as

$$\begin{aligned} \hat{K}_{Y;k}(n) = & \mathbf{W}_k^H(n) E \{ \mathbf{X}_k(n) \mathbf{X}_k^H(n) |\tilde{Y}_k(n)|^2 \} \mathbf{W}_k(n) \\ & - 2E \{ |\tilde{Y}_k(n)|^2 \} Re \{ \mathbf{W}_k^H(n) E \{ \mathbf{X}_k(n) \tilde{Y}_k^*(n) \} \} \\ & - Re \{ E^* \{ \tilde{Y}_k^2(n) \} \mathbf{W}_k^H(n) E \{ \mathbf{X}_k(n) \tilde{Y}_k(n) \} \}. \quad (5) \end{aligned}$$

The approximation of the beamformer's output signal kurtosis value in (5) is (locally) quadratic in the filter vector $\mathbf{W}_k(n)$, and the gradient at block time index n with respect to the filter vector $\mathbf{W}_k(n)$ is

$$\begin{aligned} \frac{\partial \hat{K}_{Y;k}(n)}{\partial \mathbf{W}_k^*(n)} = & E \{ \mathbf{X}_k(n) \mathbf{X}_k^H(n) |\tilde{Y}_k(n)|^2 \} \mathbf{W}_k(n) \\ & - 2E \{ |\tilde{Y}_k(n)|^2 \} E \{ \mathbf{X}_k(n) \tilde{Y}_k^*(n) \} \\ & - E^* \{ \tilde{Y}_k^2(n) \} E \{ \mathbf{X}_k(n) \tilde{Y}_k(n) \}. \quad (6) \end{aligned}$$

Optimization of the approximative kurtosis value, according to Newton's method [14], follows

$$\begin{aligned} \mathbf{W}_k(n) = & \mathbf{W}_k(n-1) \\ & - E^{-1} \left\{ \mathbf{X}_k(n) \mathbf{X}_k^H(n) |\tilde{Y}_k(n)|^2 \right\} \\ & \left(2E \{ |\tilde{Y}_k(n)|^2 \} E \{ \mathbf{X}_k(n) \tilde{Y}_k^*(n) \} \right. \\ & \left. + E^* \{ \tilde{Y}_k^2(n) \} E \{ \mathbf{X}_k(n) \tilde{Y}_k(n) \} \right). \quad (7) \end{aligned}$$

A feasible implementation is achieved through the introduction of a set of suitable approximations, where auto-regressive averaging is common in these kinds of approximations [4]

$$\mathbf{P}_k(n) \approx E^{-1} \left\{ \mathbf{X}_k(n) \mathbf{X}_k^H(n) |\tilde{Y}_k(n)|^2 \right\}, \quad (8)$$

$$a_k(n) \approx E \{ |\tilde{Y}_k(n)|^2 \}, \quad (9)$$

$$b_k(n) \approx E \{ \tilde{Y}_k^2(n) \}, \quad (10)$$

$$\mathbf{A}_k(n) \approx E \{ \mathbf{X}_k(n) \tilde{Y}_k^*(n) \}, \quad (11)$$

$$\mathbf{B}_k(n) \approx E \{ \mathbf{X}_k(n) \tilde{Y}_k(n) \}. \quad (12)$$

The matrix $\mathbf{P}_k(n)$ in (8) is updated according to the matrix inversion lemma [14] as

$$\begin{aligned} \mathbf{P}_k(n) = & \lambda_k^{-1} \mathbf{P}_k(n-1) \\ & - \frac{\mathbf{P}_k(n) \mathbf{X}_k(n) \mathbf{X}_k^H(n) |\tilde{Y}_k(n)|^2 \mathbf{P}_k(n)}{\lambda_k^2 + \lambda_k |\tilde{Y}_k(n)|^2 \mathbf{X}_k^H(n) \mathbf{P}_k(n) \mathbf{X}_k(n)}, \quad (13) \end{aligned}$$

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 (9) to (12)

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

$$b_k(n) = \lambda_k b_k(n-1) + (1 - \lambda_k) \tilde{Y}_k^2(n), \quad (15)$$

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

$$\mathbf{B}_k(n) = \lambda_k \mathbf{B}_k(n-1) + (1 - \lambda_k) \mathbf{X}_k(n) \tilde{Y}_k(n). \quad (17)$$

The update equation in (7) is now approximated as

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

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 $|\mathbf{W}_k^H(n)\mathbf{H}_k(n)|$ 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 [14]. 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 $\mathbf{W}_k(0) = (1, 1, \dots, 1)^T$.

6. REAL-TIME IMPLEMENTATION

The implementation of the proposed method is made on a floating point DSP named SHARC-21262 from Analog Devices. This is a high-performance DSP that 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. This alternative way of managing the data was introduced in [6] for a MATLAB-based real-time implementation. The transformation approach reduces the number of computations in this DSP platform by 7.5 times for a typical vector operation [15].

6.1. 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 set such that the integration time of the λ_k parameter was 0.2 s, and the integration time of the γ_k parameter was 1.5 s.

7. EVALUATION RESULTS

The performance of the proposed method is analyzed first using an offline setting with real measured data. This setting allows comparison with the analytically tractable Wiener beamformer [1]. The real-time implementation is analyzed using short-term power estimates where the algorithm may be switched on and off.

7.1. Evaluation Measures

The mean Signal to Interference Ratio (SIR) enhancement Q_k is analyzed in the subband domain using

$$Q_k = \frac{\widehat{\text{Var}}\{Y_{S,k}(n)\}\widehat{\text{Var}}\{X_{V,k}(n)\}}{\widehat{\text{Var}}\{Y_{V,k}(n)\}\widehat{\text{Var}}\{X_{S,k}(n)\}}, \quad (20)$$

where $\widehat{\text{Var}}\{\cdot\}$ designates an estimator of variance, and the subband signals $X_{S,k}(n)$, $X_{V,k}(n)$, $Y_{S,k}(n)$, and $Y_{V,k}(n)$ represents the components of the input and output signals that are related to the speech S and interference V , respectively. The mean SIR enhancement in the time domain, Q , is evaluated as

$$Q = \frac{\widehat{\text{Var}}\{y_S(t)\}\widehat{\text{Var}}\{x_V(t)\}}{\widehat{\text{Var}}\{y_V(t)\}\widehat{\text{Var}}\{x_S(t)\}}. \quad (21)$$

Fig. 1. Estimated subband power (left) for the speech sequence (solid), ferry engine noise (dotted), and factory noise (dashed). Beamformer's spectral effect on the enhanced speech (right) with original speech (solid), enhanced speech in the ferry engine noise case (dotted), and enhanced speech in the factory noise case (dashed).

Fig. 2. Subband mean SIR enhancement for speech and ferry engine noise (left), and speech and factory noise (right). The proposed method (solid), and Wiener Beamformer (dashed).

A short-term power estimate $q(t)$ is used in the time domain to evaluate the method's real-time performance. The short-term power of a signal $x(t)$ is estimated according to

$$q(t) = \alpha q(t-1) + (1-\alpha)x^2(t), \quad (22)$$

where the parameter $\alpha \in [0, 1]$ is selected so as to yield a 10 ms integration time.

7.2. 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. Previously recorded ferry engine noise and factory noise were subsequently emitted and recorded using the same setup. The signals' subband power estimates are given in the left side of Fig. 1. The speech signal is then mixed at 0 dB SIR with each of the two interfering noise signals. The subband mean SIR enhancement is provided in Fig. 2. This figure shows that the proposed method increases the subband mean SIR at several of the higher frequencies. The beamformer's spectral effect on the enhanced speech is given in the right side of Fig. 1. The time domain mean SIR enhancement in the ferry engine noise case is 10.6 dB for the proposed method and 12.4 dB for the Wiener beamformer. In the factory noise case the mean SIR enhancement was 5.6 dB for the proposed method, and 8.8 dB for the Wiener beamformer.

7.3. Online Results using Real Data

The online evaluation includes prerecorded signals consisting of male speech spatially added with music and it is provided in [16]. 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 ordering. 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.

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

8. SUMMARY AND CONCLUSIONS

This paper presents a blind beamformer based on the well known kurtosis maximization strategy. The novelty of this contribution lies in the introduced approximations, where the subband kurtosis is approximated locally using a quadratic criteria. This criteria is continuously updated online according to a Newton based search method, a variant of the RLS method. The method successfully extracts speech in convolutive mixtures of speech and ferry engine noise, speech and factory noise, and speech and music, at the cost of introduced spectral whitening of the enhanced speech. It is noticeable that, except for the spectral whitening, the speech distortion is perceptually very low. This method shows promising performance for speech enhancement in real environments, however some questions 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 natural extension of our proposed method is to introduce a power normalization of the subband kurtosis approximation in order to yield a scale-invariant solution.

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.

9. 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*, vol. 2, no. 25, pp. 689–693, 1991.
- [3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, John Wiley and Sons, 2001.
- [4] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*, John Wiley and Sons, 2003.
- [5] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Elsevier Neurocomputing*, vol. 22, no. 1–3, pp. 21–34, 1998.
- [6] B. Sällberg, M. Swartling, N. Grbić, and I. Claesson, “Real-time implementation of a blind beamformer for subband speech enhancement using kurtosis maximization,” *International Workshop on Acoustics, Echo and Noise Control*, pp. 485–489, 2006.
- [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.*, vol. 9, no. 5, pp. 524–533, 2001.
- [8] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Blind source separation of many signals in the frequency domain,” *IEEE ICASSP*, vol. 5, 2006.
- [9] F. J. Theis and Y. Inouye, “On the use of joint diagonalization in blind signal processing,” *IEEE ISCAS*, pp. 3586–3589, 2006.
- [10] B. Yang, “Projection approximation subspace tracking,” *IEEE Trans. on Signal Proc.*, vol. 43, no. 1, pp. 95–107, 1995.
- [11] R. Badeau, K. Abed-Meraim, G. Richard, and B. David, “Sliding window orthonormal past algorithm,” *IEEE ICASSP*, vol. 5, pp. 261–264, 2003.
- [12] P. P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice Hall, 1993.
- [13] C. Nikias and A. Petropulu, *Higher-Order Spectral Analysis - A Nonlinear Signal Processing Framework*, Prentice Hall, 1993.
- [14] S. Haykin, *Adaptive Filter Theory*, John Wiley and Sons, 2002.
- [15] 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.
- [16] T. W. Lee, “Blind source separation: Audio examples,” [Online] http://www.cnl.salk.edu/fewon/Blind/blind_audio.html, Feb. 17 2007.