A CALIBRATED SUBBAND BEAMFORMING ALGORITHM FOR SPEECH ENHANCEMENT

Zohra Yermeche, Pilar Márquez García, Nedelko Grbić and Ingvar Claesson

Blekinge Institute of Technology
Department of Telecommunications and Signal Processing
372 25 Ronneby, Sweden. E-mail: zye@bth.se

ABSTRACT

This paper proposes a new calibrated adaptive frequency domain beamformer for speech enhancement. The beamformer is based on the principle of a soft constraint formed from calibration data, rather than precalculated from free-field assumptions. The benefit is that the real room acoustical properties will be taken into account. The proposed algorithm continuously estimates the spatial information for each frequency band, based on weighting of the received data. The update of the beamforming weights is done recursively where the initial precalculated correlation estimates of the speech constitute a soft constraint. The soft constraint secures the spatial-temporal passage of the desired source signal, without the need of any speech detection. The performance is evaluated in real world scenarios, both in a car- and a restaurant- environment. Interference and noise suppression of more than 15 dB is accomplished, while very small distortion is measured for the source signal.

1. INTRODUCTION

The merge between computers and telephony techniques brings up the demand for convenient hands-free communication. Teleconference systems, hands-free telephone sets, and speech recognition systems are becoming increasingly more common. In such systems, the user wishes to lead a conversation in much the same way as in a normal person-to-person conversation. However, since the speaker is distant from the device, many problems are caused by room reverberation, ambient noise accompanying the desired speech signal, as well as acoustic coupling between loudspeaker and microphone. Microphone array techniques have the potential to reduce the additive noise and enhance the speech perception by directional processing. Several beamforming techniques have been suggested in order to enhance the desired speech source [1, 2, 3]. Many algorithms rely on voice activity detection (VAD) in order to turn off adaptation when the source signal is active. This is done in order to avoid source signal cancellation effects [4], which may result in high levels of speech distortion.

This paper proposes a new calibrated adaptive frequency domain beamformer for speech enhancement without the use of a VAD. The beamformer is based on the principle of a soft constraint, formed from calibration data, rather than precalculated from free-field assumptions as it is done in [5]. The benefit is that the real room acoustical properties will be taken into account.

Information about the speech location is first put into the algorithm in an initial acquisition. The objective is formulated in the frequency domain as a weighted recursive least squares solution. In order to track variations in the surrounding noise environment, the proposed algorithm continuously estimates the spatial information for each frequency band, based on weighting of the received data. The update of the beamforming weights is done recursively where the initial precalculated correlation estimates constitutes a soft constraint.

Performance of the algorithm is evaluated on real data recordings conducted in different hands-free environments. Results show a significant noise- and hands-free- interference reduction within the conventional telephone bandwidth.

2. PROBLEM FORMULATION

The source is assumed to be a wideband source, as in the case of a speech signal, located in the near field of a uniform linear array of I microphones. The beamformer uses finite length digital linear filters at each microphone. The output of the beamformer is given by,

$$y[n] = \sum_{i=1}^I \sum_{j=0}^{L-1} w_i[j] x_i[n-j] \quad (1)$$

where $L-1$ is the order of the FIR filters and $w_i[j]$, $j = 0, 1, \dots, L-1$, are the FIR filter taps for channel number i . The signals, $x_i[n]$, are digitally sampled microphone observations and the beamformer output signal is denoted $y[n]$.

The computational complexity of the finite impulse response filters is substantially reduced by introducing a sub-band beamforming scheme. The proposed structure con-

sists of a uniform over-sampled analysis filter-bank, used to decompose the received array signals into K subband signals, and a set of adaptive beamformers, each adapting on the multichannel subband signals. The outputs of the beamformers are reconstructed by a synthesis filter-bank in order to create a time domain output signal. The spatial characteristics of the input signal are maintained when using modulated filter-banks (analysis and synthesis), defined by single prototype filters, which leads to efficient polyphase realizations. Figure 1 illustrates the subband beamformer structure. The filter-banks are designed with the methodology described in [6], where in-band and reconstruction aliasing effects are minimized.

Fig. 1. Structure of the subband beamformer.

In the frequency domain formulation the filtering operations become multiplications with number I complex frequency domain representation weights, $w_i^{(f)}$. For a specific frequency, f , the output is given by

$$y^{(f)}[n] = \sum_{i=1}^I w_i^{(f)} x_i^{(f)}[n] \quad (2)$$

where the signals, $x_i^{(f)}[n]$ and $y^{(f)}[n]$, are time domain signals as specified before but they are narrow band, containing essentially components with frequency f .

2.1. Beamformer objective

The objective is formulated in the frequency domain as a least squares solution defined for a data set of N samples. However, since the reference source signal information is not directly available, a calibration sequence gathered in a quiet environment is used instead. This calibration signal will represent the temporal and spatial information about the source.

The objective is to calculate

$$\mathbf{w}_{ls,opt}^{(k)}(N) = \left[\hat{\mathbf{R}}_{ss}^{(k)}(N) + \hat{\mathbf{R}}_{xx}^{(k)}(N) \right]^{-1} \hat{\mathbf{r}}_s^{(k)}(N) \quad (3)$$

where the real frequency $f = F_s k / K$, with F_s the sampling frequency and K the total number of subbands, and where the array weight vector, $\mathbf{w}_{opt}^{(k)}$, for the subband k is defined as

$$\mathbf{w}_{opt}^{(k)} = [w_1^{(k)} \ w_2^{(k)} \ \dots \ w_I^{(k)}]^T. \quad (4)$$

The source correlation estimates are precalculated in the calibration phase as

$$\hat{\mathbf{R}}_{ss}^{(k)}(N) = \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{s}^{(k)}[n] \mathbf{s}^{(k)H}[n] \quad (5)$$

$$\hat{\mathbf{r}}_s^{(k)}(N) = \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{s}^{(k)}[n] s_r^{(k)*}[n] \quad (6)$$

where

$$\mathbf{s}^{(k)}[n] = [s_1^{(k)}[n], \ s_2^{(k)}[n], \ \dots \ s_I^{(k)}[n]]^T$$

are microphone observations when the calibration source signal is active alone, while the observed data correlation matrix estimate is given by

$$\hat{\mathbf{R}}_{xx}^{(k)}(N) = \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{x}^{(k)}[n] \mathbf{x}^{(k)H}[n] \quad (7)$$

where the received array data vector is

$$\mathbf{x}^{(k)}[n] = [x_1^{(k)}[n], \ x_2^{(k)}[n], \ \dots \ x_I^{(k)}[n]]^T.$$

3. A RECURSIVE ALGORITHM

In order to keep track of the surrounding environment variations, an adaptive structure is preferred. A weighting factor is introduced on the observed signal correlation matrix estimates, while using at the same time the precalculated source correlation estimates as specified in Eq. (5) and (6). It is desirable to calculate the optimal beamforming weights according to Eq. (3) based on the available data continuously in a recursive way. Therefore, a total covariance matrix, $\mathbf{R}^{(k)}$, for subband k is first introduced

$$\hat{\mathbf{R}}^{(k)}[n] = \hat{\mathbf{R}}_{ss}^{(k)}(N) + \hat{\mathbf{R}}_{xx}^{(k)}[n]. \quad (8)$$

A final aim is to update the inverse of this matrix recursively at each time index n , by using the *Matrix-Inversion-Lemma*, while maintaining the constant portion corresponding to the pre-calculated source covariance matrix, according to,

$$\begin{aligned} \hat{\mathbf{R}}^{(k)}[n] = & \hat{\mathbf{R}}_{ss}^{(k)}(N) + \lambda \hat{\mathbf{R}}_{xx}^{(k)}[n-1] + \mathbf{x}^{(k)}[n] \mathbf{x}^{(k)H}[n] = \\ & \lambda \hat{\mathbf{R}}^{(k)}[n-1] + \mathbf{x}^{(k)}[n] \mathbf{x}^{(k)H}[n] + (1-\lambda) \hat{\mathbf{R}}_{ss}^{(k)}(N). \end{aligned} \quad (9)$$

where λ is a weighting factor. The effect of the above update is that the total correlation matrix is weighted and both the rank one “correction term,” $\mathbf{x}(n)\mathbf{x}^H(n)$, and the small portion $(1 - \lambda)$, of the pre-calculated source covariance matrix, which has been reduced by the weighting factor, are added. One way to circumvent the inversion of the calibration correlation matrix, $\hat{\mathbf{R}}_{ss}^{(k)}(N)$, and thus substantially reduce the complexity, is to update the total correlation matrix by adding scaled eigenvectors of the calibration correlation matrix, which will result in several rank one updates as

$$\hat{\mathbf{R}}^{(k)}[n] = \lambda \hat{\mathbf{R}}^{(k)}[n-1] + \mathbf{x}^{(k)}[n]\mathbf{x}^{(k)H}[n] + (1 - \lambda) \sum_{p=1}^P \gamma_p^{(k)} \mathbf{q}_p^{(k)} \mathbf{q}_p^{(k)H} \quad (10)$$

where $\gamma_p^{(k)}$ is the p -th eigenvalue, $\mathbf{q}_p^{(k)}$ is the p -th eigenvector of the calibration correlation matrix, $\hat{\mathbf{R}}_{ss}^{(k)}$, and P is the dimension of the signal space, i.e. the effective rank of the matrix. One way to further reduce the complexity while only affecting the scale of the problem is by sequentially adding one scaled eigenvector at each sample instant. The weighted optimal recursive least squares solution at sample instant, n , is now given by

$$\mathbf{w}_{ls}^{(k)}[n] = [\hat{\mathbf{R}}^{(k)}[n]]^{-1} \hat{\mathbf{r}}_s^{(k)}(N) \quad (11)$$

where the calibration correlation vector, $\hat{\mathbf{r}}_s^{(k)}(N)$, is gathered in advance and it is assumed to be uncorrelated with the observed data.

When the statistical properties of the environmental noise is changing abruptly, a smoothing of the weights may be appropriate.

3.1. Summary of the Algorithm

We assume that the source signal and the additional disturbing and known sources with fixed location are available during the acquisition phase. The estimated correlation matrix, $\hat{\mathbf{R}}_{ss}^{(k)}$, and the estimated source signal cross correlation vector, $\hat{\mathbf{r}}_s^{(k)}$, is made available from an initial acquisition, for each subband. The source cross correlation estimates must be acquired when only the source signal of interest is active.

The algorithm is stated as an iterative procedure, individually for each subband, indexed $k = 0, 1, \dots, K - 1$. The algorithm is run sequentially with the steps in the operation phase.

Initialization phase:

- Calculate the source correlation matrix and the cross correlation vector according to Eqs. (5) and (6) when the source of interest is active alone.

- Calculate the observed data correlation matrix according to Eqs. (7) when known disturbing sources are active.

- The correlation matrices are memorized in a diagonalized form:

$$(\hat{\mathbf{R}}_{ss}^{(k)}(N) + \hat{\mathbf{R}}_{xx}^{(k)}(N)) = \mathbf{Q}^{(k)H} \mathbf{\Gamma}^{(k)} \mathbf{Q}^{(k)}$$

The eigenvectors are denoted:

$$\mathbf{Q}^{(k)} = [\mathbf{q}_1^{(k)}, \mathbf{q}_2^{(k)}, \dots, \mathbf{q}_I^{(k)}]$$

The eigenvalues are denoted:

$$\mathbf{\Gamma}^{(k)} = \text{diag}([\gamma_1^{(k)}, \gamma_2^{(k)}, \dots, \gamma_I^{(k)}])$$

- Initialize the weight vector from Eq. (4) as a zero vector
- Initialize the inverse total correlation matrix as $\mathbf{P}^{(\Omega)}(0) = \sum_{p=1}^P \gamma_p^{-1} \mathbf{q}_p \mathbf{q}_p^H$, and define the same size dummy variable matrix, $\mathbf{D}$.
- Choose a weighting factor λ and a weight smoothing factor α

Operation phase:

for $n = 1, 2, \dots$

- When any of the sources are active simultaneously, update the inverse total correlation matrix, as

$$\mathbf{D} = \lambda \mathbf{P}^{(k)}(n-1) - \frac{\lambda^{-2} \mathbf{P}^{(k)}(n-1) \mathbf{x}(n) \mathbf{x}^H(n) \mathbf{P}^{(k)}(n-1)}{1 + \lambda^{-1} \mathbf{x}^H(n) \mathbf{P}^{(k)}(n-1) \mathbf{x}(n)}$$

$$\mathbf{P}^{(k)}(n) = \mathbf{D} - \frac{\gamma_p(1-\lambda) \mathbf{D} \mathbf{q}_p \mathbf{q}_p^H \mathbf{D}}{1 + \gamma_p(1-\lambda) \mathbf{q}_p^H \mathbf{D} \mathbf{q}_p}$$

where index $p = k \pmod{P}$ denotes the index of the eigenvalues and eigenvectors given in Eq. (10).

- Calculate the weights, for each sample instant, as

$$\mathbf{w}^{(k)}(n) = \alpha \mathbf{w}^{(k)}(n-1) + (1 - \alpha) \mathbf{P}^{(k)}(n) \mathbf{r}_s^{(k)}$$

where α is a smoothing parameter for the weights and it corresponds to the real valued pole of a first order AR-model.

- Calculate the output for the subband k as

$$y^{(k)}(n) = \mathbf{w}^{(k)}(n)^H \mathbf{x}(n).$$

4. SIMULATIONS

The subband adaptive algorithm is evaluated on real data recordings conducted in different hands-free telephony environments. A linear microphone array was used and data was gathered on a multichannel DAT-recorder with a sampling rate of 12 KHz, and with 5 KHz bandwidth. The simulations were made with speech sequences of both male and female speakers. The algorithm was run with a number of 64 subbands.

4.1. Car Environment

4.1.1. Conditions

The performance of the beamformer was evaluated in a car handsfree situation, with a six sensor high quality Sennheiser microphone array mounted on the visor at the passenger side. The distance between the speaker's position and the microphone array was 35 cm, perpendicular to the array axis, while the spacing between adjacent elements in the array is 50 mm. The car was running at the speed of 110 km/h on a paved road. The speech originating from the passenger position constitutes the desired source signal, while the ambient noise originating from the motion of the car constitutes the background noise.

4.1.2. Results

The performance evaluation of the proposed beamformer includes source speech distortion and suppression of both background noise and hands-free loudspeaker interference. Recordings with real speech signals, from both the artificial talker and the hands-free loudspeaker, were recorded both individually and while driving, to be used to measure the performance. The algorithm was also run when all sources were active as in a normal conversation. The performance was compared to the least squares beamformer in the same situation, as well as for different lengths of the subband filters, see Table 1.

A major reduction of ambient car noise and interfering hands-free loudspeaker signal (denoted also as echo) is obtained, as shown in Figure 2.

4.2. Restaurant Environment

4.2.1. Conditions

Measurements were conducted in a restaurant room environment where the impact on the beamformer performance, of a different set-up is investigated. An array of four low quality microphones was installed in a fixed position, while a number of scenarios have been simulated with different positions of the target source and the interfering loudspeaker, relative to the microphone array (see Figure 3), and where

Performance	Speech dist.	Noise supp.	Echo supp.
LS 1-tap	-28.6	12.9	13.6
CW-RLS			
1-tap	-28.0	12.9	13.1
2-taps	-28.8	13.4	14.4
3-taps	-30.0	13.8	15.2
4-taps	-30.4	14.2	15.4
5-taps	-30.5	14.3	15.7
6-taps	-30.7	14.3	15.8

Table 1. Performance measures in dB for Least Square (LS) beamformer and Calibrated weighted Recursive Least Squares (CW-RLS) beamformer outputs in relation to the reference sensor input, when each signal component, speech-, noise- and echo- components are active individually.

Fig. 2. Time signal of unprocessed single microphone observation followed by the beamformer output signal in short time (20 ms) power estimates.

the ambient noise generated by the activity in the restaurant constituted the background noise. The algorithm was run for a one tap subband filters.

4.2.2. Results

Results show a significant noise- and hands-free- interference reduction within the actual bandwidth, for the different scenarios presented in Figure 3 (see table 2). However the performance of the beamformer is reduced compared to the results given in Sec. 4.1.2, mainly due to the fact that less number of microphones were used. Also, since spatial separation between the source and the hands-free loudspeaker (i.e. interference signals), as well as their position to the array, controls the level of interference suppression, the placement of the hands-free loudspeaker should be made

Scenarios	Speech dist.	Noise supp.	Echo supp.
(a)	-25.5	7.4	7.2
(b)	-25.5	7.1	10.1
(c)	-25.7	7.5	18.3
(d)	-25.7	7.3	12.8
(e)	-25.7	7.4	14.6

Table 2. Performance measures in dB of CW-RLS beamformer outputs achieved for the scenarios illustrated in Figure 3.

with this fact in mind.

Fig. 3. Data acquisition scenarios: The target speaker is at a distance of 50 cm from the microphone array, and has a fixed position for all scenarios. The Interfering speaker in (a), (b), (c) and (d) is also at a fixed distance from the array but makes a different angle with the target source position at each scenario, while in scenario (e) it is positioned further away from both the array and the target source. The ambient noise of the room is included in all scenarios.

5. CONCLUSIONS

A new calibrated adaptive frequency domain beamforming has been presented. The solution is based on the principal of a soft constraint, formed from calibrated data. The algorithm recursively estimates the spatial information of the received data, while the initial precalculated source correlation estimates constitutes a soft constraint in the solution. A real hands-free implementation with a linear array shows good noise and interference suppression as well as low speech distortion.

6. REFERENCES

- [1] D. A. Florêncio and H. S. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *IEEE International Conference on Acoustics,*

Speech and Signal Processing, May 2001, vol. 1, pp. 197–200.

- [2] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 5, no. 5, pp. 425–437, Sep. 1997.
- [3] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 8, no. 5, pp. 497 – 507, Sep. 2000.
- [4] D. Johnson and D. Dudgeon, *Array Signal Processing - Concepts and Techniques*, Prentice Hall, 1993.
- [5] N. Grbić and S. Nordholm, "Soft constrained subband beamforming for hands-free speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2002, in press.
- [6] J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, "Design of oversampled uniform dft filter banks with delay specifications using quadratic optimization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2001, vol. VI, pp. 3633–3636.