SPEECH ENHANCEMENT EMPLOYING ADAPTIVE BEAMFORMER WITH RECURSIVELY UPDATED SOFT CONSTRAINTS

Hai Q. Dam[1], *Sven Nordholm*[1], *Nedelko Grbić*[2], *and Hai H. Dam*[1]

[1] Western Australian Telecommunications Research Institute (WATRI) *,
University of Western Australia, WA 6009, Australia
[2] Blekinge Institute of Technology, Department of Telecommunications
and Signal Processing, 372 25 Ronneby, Sweden

ABSTRACT

A novel adaptive beamformer employing recursively updated soft constraints for acoustic speech enhancement is proposed. The beamformer operates in a subband structure to allow a time-frequency operation for each channel. Consequently, the processing performed can be viewed as a combined weighted spatial,frequency and temporal filter. The major benefit of the new recursive soft constrained beamformer is that it allows the possibility of using the spectral information of the desired source to modify the soft constraint. This has clear benefits on the speech distortion of the source of interest. The novel adaptive beamformer involves continuous modification of the soft constraint by feeding back the spectral content of the estimated output speech signal. Evaluations of the proposed beamformer based on real car data show that the proposed algorithm significantly improves the speech quality with noise suppression levels up to 17 dB.

1. INTRODUCTION

In recent years, microphone arrays have received increasing attention for the acquisition of speech in hands-free and distant-talker scenarios [1]. Microphone arrays can be used to reduce interference in hearing aids, teleconferencing systems, hands-free microphones in automobiles, computer terminals, speaker phones and speech recognition systems [2], [3], and [4]. Based on adaptive beamforming, microphone arrays are especially promising in terms of noise and reverberation suppressions.

In [4], a soft constrained subband beamformer was proposed as a means to enhance acoustic speech signals. This soft constraint was formulated assuming that the power spectral density (PSD) of the source is constant over time and frequency range and located in a certain spatial region. However, remembering that speech is short-term stationary, which means that the spectrum change over time. Consequently, this original approach does not efficiently utilize the time-frequency information of the source of interest (SOI). The contribution in this paper is a new recursive soft constraint. The constraint is recursively updated using the current time-frequency content of the output signal from the beamformer. Since the output is an enhanced version of the SOI, this considerably improves performance and speech quality compared to the existing subband soft constrained beamformer. The PSD of the SOI is unknown at the current time instant, hence the beamformer output from one time-instant earlier is used in the next iteration as an estimation of the PSD. This results in a time-varying soft constraint that depends on the spectral content of the SOI. The spatial information is captured by using localization information and the spectral weighting is changed according to the spectral content of the SOI. The variation in the spectral content of the SOI over time will be tracked thus providing a spectrally optimized constraint at each time instant.

Evaluations using real data are performed under the same scenario as in [4]. Results show that the proposed algorithm significantly improves the speech quality. In addition absence of speech signal does not influence the performance of the algorithm and there is no need for a Voice Activity Detector (VAD).

2. SOFT CONSTRAINED BEAMFORMING ALGORITHM

Consider a linear microphone array system as shown in Fig. 1 with I microphones and a source in a near-field model. This SOI is in reality a person speaking and modelled as an area of point sources clustered closely in space. This area is assumed to be within a range of radii $[R_a, R_b]$ and angles $[\theta_a, \theta_b]$. The noise sources are assumed to be uncorrelated with the SOI. The only germane part in the modelling is to capture the SOI, since that determines the constraint.

*WATRI is a joint venture between The University of Western Australia and Curtin University of Technology. The work has also been sponsored by ARC under grant no. A00105530

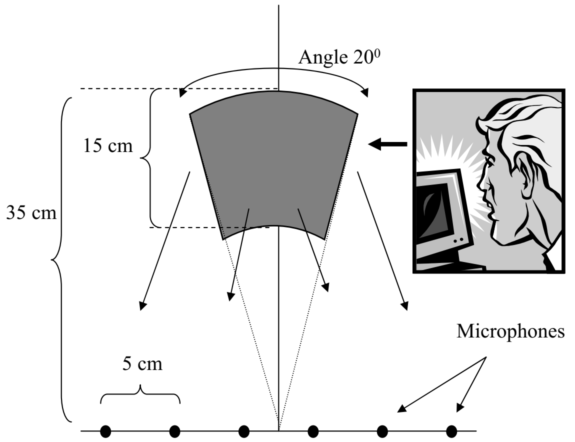

Fig. 1. Configuration of the linear microphone array and the source area.

2.1. Problem formulation

The structure of the proposed subband beamformer is shown in Fig. 2. The received data vector of the microphone array $\mathbf{s}(l) = [s_1(l), \cdots, s_I(l)]$, where l denotes the time index, is first divided into M subbands by using analysis filterbanks. The subband signals are then fed into the proposed beamformer. Finally, the estimated output signal $y(l)$ is obtained by passing the beamformer output through a synthesis filterbank.

For a frequency Ω, the covariance matrix $\mathbf{R}_s^{(\Omega)}$ and the cross-covariance vector $\mathbf{r}_s^{(\Omega)}$ of the received signals can be calculated as follows

$$\mathbf{R}_s^{(\Omega)} = \int\int_{R_a,\theta_a}^{R_b,\theta_b} S(\Omega)\mathbf{d}(R,\theta,\Omega)\mathbf{d}(R,\theta,\Omega)^H dRd\theta \tag{1}$$

and

$$\mathbf{r}_s^{(\Omega)} = \int\int_{R_a,\theta_a}^{R_b,\theta_b} S(\Omega)\mathbf{d}(R,\theta,\Omega)dRd\theta \tag{2}$$

where $(.)^H$ denotes the Hermitian transpose of a vector and $S(\Omega)$ is the PSD of the SOI. The response vector $\mathbf{d}(R,\theta,\Omega)$ is given as

$$\mathbf{d}(R,\theta,\Omega) = \left[\frac{1}{R_1}e^{-j\Omega\tau_1(R,\theta)}, \ldots, \frac{1}{R_I}e^{-j\Omega\tau_I(R,\theta)}\right]^T \tag{3}$$

where $\tau_i(R,\theta)$ and R_i, $1 \leq i \leq I$, denote the time delay from a point source of radius R and angle θ to the sensor i and the distance between the source and the sensor i, respectively.

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

$$\mathbf{w}_{opt}^{(\Omega)} = [w_1^{(\Omega)}, w_2^{(\Omega)}, \ldots, w_I^{(\Omega)}]^T \tag{4}$$

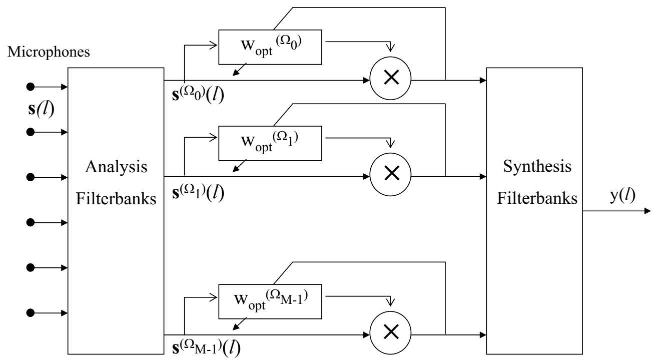

Fig. 2. Structure of the proposed subband beamformer.

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

$$\mathbf{w}_{opt}^{(\Omega)} = \left[\mathbf{R}_s^{(\Omega)} + \mathbf{R}_n^{(\Omega)}\right]^{-1} \mathbf{r}_s^{(\Omega)} \tag{5}$$

where $\mathbf{R}_n^{(\Omega)}$ is the noise covariance matrix. It follows from (5) that

$$\begin{aligned}\mathbf{w}_{opt}^{(\Omega)} &= \left[\mathbf{R}_s^{(\Omega)}/S(\Omega) + \mathbf{R}_n^{(\Omega)}/S(\Omega)\right]^{-1}\left(\mathbf{r}_s^{(\Omega)}/S(\Omega)\right)\\ &= \left[\bar{\mathbf{R}}_s^{(\Omega)} + \bar{\mathbf{R}}_n^{(\Omega)}\right]^{-1}\bar{\mathbf{r}}_s^{(\Omega)}\end{aligned} \tag{6}$$

where

$$\bar{\mathbf{R}}_s^{(\Omega)} = \int\int_{R_a,\theta_a}^{R_b,\theta_b} \mathbf{d}(R,\theta,\Omega)\mathbf{d}(R,\theta,\Omega)^H dRd\theta, \tag{7}$$

$$\bar{\mathbf{R}}_n^{(\Omega)} = \mathbf{R}_n^{(\Omega)}/S(\Omega) \tag{8}$$

and

$$\bar{\mathbf{r}}_s^{(\Omega)} = \int\int_{R_a,\theta_a}^{R_b,\theta_b} \mathbf{d}(R,\theta,\Omega)dRd\theta. \tag{9}$$

From equations (7) and (9), $\bar{\mathbf{R}}_s^{(\Omega)}$ and $\bar{\mathbf{r}}_s^{(\Omega)}$ can be calculated for a given constraint region without the knowledge of the PSD of the source. Thus, we only need to recursively estimate $\bar{\mathbf{R}}_n^{(\Omega)}$. Since data containing only the active noise is not available, the noise covariance matrix $\mathbf{R}_n^{(\Omega)}$ is estimated by using K samples of received data $\mathbf{s}^{(\Omega)}(l)$, where K is a fixed positive number. Moreover, the exact PSD of the source $S(\Omega)$ is not possible to obtain especially in the car environment where strong speech masking components of noise exists. Thus, we propose to estimate $S(\Omega)$ by using the previous beamformer output $\mathbf{w}_{opt}^{(\Omega)}(l-1)^H\mathbf{s}^{(\Omega)}(l-1)$. For any $k > 0$, let

$$\mathbf{z}^{(\Omega)}(k) = \frac{\mathbf{s}^{(\Omega)}(k)}{|\mathbf{w}_{opt}^{(\Omega)}(k-1)^H\mathbf{s}^{(\Omega)}(k-1)|} \tag{10}$$

where $|.|$ is the amplitude of a complex number. At iteration l, $\bar{\mathbf{R}}_n^{(\Omega)}(l)$ can be estimated based on $\mathbf{z}^{(\Omega)}(k)$ where $\max(0, l-K) \leq k \leq l$ as follows.

- If $l \leq K$ then

$$\bar{\mathbf{R}}_n^{(\Omega)}(l) = \frac{1}{l}\sum_{k=1}^{l} \mathbf{z}^{(\Omega)}(k)\mathbf{z}^{(\Omega)}(k)^H. \qquad (11)$$

- If $l > K$ then

$$\bar{\mathbf{R}}_n^{(\Omega)}(l) = \frac{1}{K}\sum_{k=l-K+1}^{l} \mathbf{z}^{(\Omega)}(k)\mathbf{z}^{(\Omega)}(k)^H. \qquad (12)$$

In the next section, a recursive algorithm is developed to efficiently update the beamforming weights according to (6), (11) and (12) based on the received data at the microphones.

2.2. Proposed Recursive Algorithm

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

$$\bar{\mathbf{R}}^{(\Omega)}(l) = \bar{\mathbf{R}}_s^{(\Omega)} + \bar{\mathbf{R}}_n^{(\Omega)}(l) \qquad (13)$$

and

$$\mathbf{P}^{(\Omega)}(l) = [\bar{\mathbf{R}}^{(\Omega)}(l)]^{-1}. \qquad (14)$$

The optimal weight vector (6) for the iteration l is then reduced to

$$\mathbf{w}_{opt}^{(\Omega)}(l) = \mathbf{P}^{(\Omega)}(l)\bar{\mathbf{r}}_s^{(\Omega)}. \qquad (15)$$

It follows from (12) that for $l > K$, $\bar{\mathbf{R}}^{(\Omega)}(l)$ can be obtained from the previous estimate as

$$\bar{\mathbf{R}}^{(\Omega)}(l) = \bar{\mathbf{R}}^{(\Omega)}(l-1) + \tfrac{1}{K}\mathbf{z}^{(\Omega)}(l)\mathbf{z}^{(\Omega)}(l)^H - \tfrac{1}{K}\mathbf{z}^{(\Omega)}(l-K)\mathbf{z}^{(\Omega)}(l-K)^H. \qquad (16)$$

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

$$\mathbf{P}^{(\Omega)}(l) = \mathbf{D} + \frac{\mathbf{D}\mathbf{z}^{(\Omega)}(l-K)\mathbf{z}^{(\Omega)}(l-K)^H\mathbf{D}}{K\left(1 + \mathbf{z}^{(\Omega)}(l-K)^H\mathbf{D}\mathbf{z}^{(\Omega)}(l-K)\right)} \qquad (17)$$

where

$$\mathbf{D} = \mathbf{P}^{(\Omega)}(l-1) - \frac{\mathbf{P}^{(\Omega)}(l-1)\mathbf{z}^{(\Omega)}(l)\mathbf{z}^{(\Omega)}(l)^H\mathbf{P}^{(\Omega)}(l-1)}{K\left(1 + \mathbf{z}^{(\Omega)}(l)^H\mathbf{P}^{(\Omega)}(l-1)\mathbf{z}^{(\Omega)}(l)\right)}. \qquad (18)$$

The recursive algorithm is now given in the following steps.

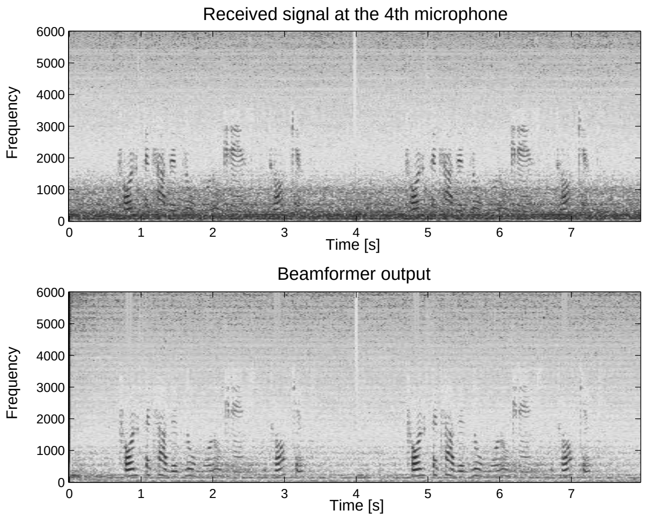

Fig. 3. Spectrograms of the noisy signal and beamformer output.

- *Step 1: Choose a number of subbands M, a block size K and a weight smoothing factor λ[1].*

- *Step 2: Initialize $l = 1$ 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 (7) and (9), respectively.*

- *Step 4: If $l \leq K$, the matrix $\mathbf{P}^{(\Omega)}(l)$ is calculated according to (10), (11) and (14) by using pseudo-inverse operation instead of the conventional matrix inverse operation due to rank deficiency. Otherwise, the matrix $\mathbf{P}^{(\Omega)}(l)$ is updated recursively by using (17) and (18). The weight vector is then updated as*

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

and the output is given by

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

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

3. SIMULATION RESULTS

The performance of the beamformer is evaluated in a hands-free situation in a car with six sensor microphone array

[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.

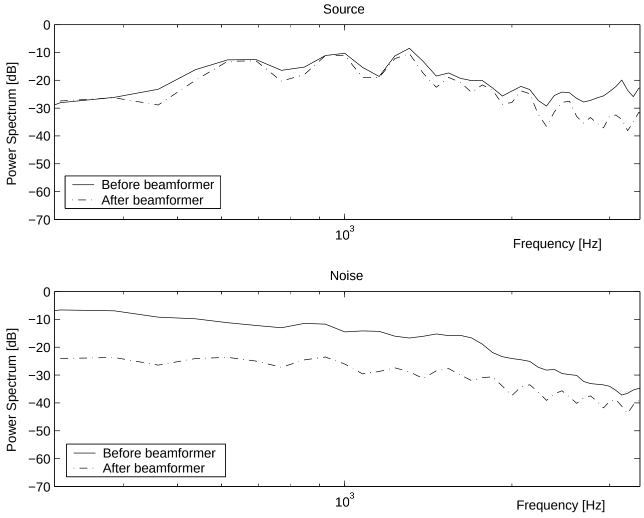

Fig. 4. PSD of the source and the noise before and after the beamformer.

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 a 300-3400 Hz bandwidth. The car was running at the speed of 110 km/h on a paved road.

A uniform factor two over-sampled DFT filterbank is used to decompose the received array signals into M subbands. The analysis and synthesis filterbanks are designed by using a Hamming window with the cut off frequency $\omega_c = \pi/M$.

Simulation is performed with 64 subbands with the noise covariance matrix estimated using $K = 64$ samples. The weight smoothing factor λ is chosen as $\lambda = 0.99$ and the length of the speech signal is 8 seconds. The matrix (7) and the vector (9) are calculated by using numerical integration with the constrained region given in Fig. 1.

Fig. 3 shows the spectrogram of the received signal at the 4^{th} microphone and the beamformer output. The noise level of the signal at other microphones is approximately the same as the 4^{th} microphone. Clearly, the noise is significantly suppressed by passing the received signal through the beamformer.

Fig. 4 plots the PSD of the source and the noise before and after the beamformer. The PSD of the source after the beamformer is approximately the same as before the beamformer, especially for low frequencies that are important to human hearing. The total noise suppression is more than 15 dB and the suppression is almost the same for all the frequencies.

The noise suppression levels for different frequencies in the first and the last four seconds are given in Table 1. The suppression is approximately the same for both time periods. Thus, the recursive estimate for the noise covariance matrix does not result in error propagation. Moreover, the absence of speech signal does not have significant influence on the performance of the algorithm. This can be seen from the fact that there exists a short silence between the two periods. Informal listening tests show a very low signal distortion.

Frequency (Hz)	Noise suppression [dB]	
	The first 4 secs	The last 4 secs
300-1000	17	17.2
1000-2000	15	15
2000-3400	13.5	13.2
Total suppression	15.2	15.1

Table 1. The noise suppression level for different frequency bands.

4. CONCLUSIONS

In this paper, a new soft constrained beamformer is developed for acoustic speech enhancement. The PSD of received speech signal is recursively estimated in order to weigh the output efficiently in the temporal domain. This is done by including the output signal power estimate into the soft constraint formulation. The advantage of this novel approach over the earlier suggested soft constrained beamformer is that the proposed adaptive beamformer significantly improves the speech quality while maintaining high noise suppression levels up to 17 dB for real car data.

5. REFERENCES

[1] M. Brandstein and D. Ward Eds., *Microphone Arrays - Techniques and Applications,* Springer Verlag, Berlin, January 2001,

[2] C. Kyriakakis, P. Tsakalides, and T. Holman, “Surrounded by Sound,” *IEEE Signal Processing Magazine,* pp. 55-66, Jan. 1999.

[3] 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 Processing,* vol. 47, no. 10, pp. 2677–2684, Jun. 1999.

[4] N. Grbić and S. Nordholm, “Soft Constrained Subband Beamforming for Hands-Free Speech Enhancement,” *Proc. ICASSP-02,* vol. 1, pp. 885–888, May 2002