SPEECH ENHANCEMENT USING MULTIPLE SOFT CONSTRAINED SUBBAND BEAMFORMERS AND NON-COHERENT TECHNIQUE

Siow Yong Low, Nedelko Grbić and Sven Nordholm

Western Australian Telecommunications Research Institute (WATRI)
The University of Western Australia, WA 6009, Australia

ABSTRACT

This paper presents a new robust microphone array processing technique to enhance speech signal under the influence of noise and jammer(s). The new structure comprises of two soft constrained subband beamformers and a non-coherent processing technique. Essentially, the first beamformer enhances the desired speech signal in a specified constrained region. The residual interference in the beamformer's output is then spectral subtracted using the estimated interference from the second beamformer. Evaluations in a real office environment show higher interference suppression compared to those obtained using the soft constrained beamformer only. Most importantly, this is achieved with negligible expense on target signal distortion.

1. INTRODUCTION

Microphone array has had a long-standing achievement as far as speech enhancement is concerned [1]. It makes use of both spatial and temporal information to enhance target signal by suppressing interference whether due to reverberation, background noise or jammer (e.g. loudspeaker). Such capability paves the way for microphone array in applications like hands-free communications, speech recognition devices and hearing aids. There are a great deal of literatures which explain the various noise reduction techniques employing the microphone array. Among them, the generalized sidelobe canceller (GSC) prominently stands out [1, 2]. The scheme offers good interference suppression but succumbs to target signal cancellation in a reverberant environment.

This paper proposes a novel robust subband adaptive microphone array incorporating multiple soft constrained beamformers [3] and a non-linear technique. The structure aims at reducing the interference effects whilst maintaining smallest target signal cancellation even in reverberant environment. Basically, one of the beamformers extracts the target signal whilst the other extracts interference. Following that, an improved spectral subtraction is performed on the outputs of both the beamformers to give the desired target signal. The idea of using the non-coherent method offers less signal distortion compared to conventional coherent approach such as adaptive noise canceller, since the interference extraction process cannot be made perfect. Simply, the interference is suppressed in stages by the soft constrained beamformer and spectral subtraction. Another advantage is the fact that all processing is made in subbands [4]. This means that wideband signals can be decomposed into a number of narrower band signals which yields a more efficient processing system.

Evaluations in a real office hands-free environment are presented. Various setups were tested including the performance in both diffuse and directional noise fields. Results show the proposed structure achieves higher noise and jammer suppressions compared to that of employing the soft constrained beamformer only.

2. PROBLEM FORMULATION

2.1. Objective

Consider a wideband source located in the near field of a uniform linear array with L microphones. The speech source is modeled as an infinite number of point sources clustered closely within a range of radius $[R_a, R_b]$ and inside the range of arrival angles $[\theta_a, \theta_b]$ (See Figure 1). Our objective is to construct the beamformer such that it passes the speech signal in the specified constrained region and rejects all interference outside this region.

The response vector of the array is given as

$$\mathbf{d}(R,\theta,\Omega_m) = \left[\frac{1}{R_1}e^{-j\Omega_m\tau_1(R,\theta)},\ldots,\frac{1}{R_L}e^{-j\Omega_m\tau_L(R,\theta)}\right]^T \quad (1)$$

where $\tau_l(R,\theta)$ denotes the time delay from a point source at radius R from the origin and angle θ to sensor l, R_l is the distance between the source and sensor l, and Ω_m denotes the real angular center frequency in the mth band. The reference point for the beamformer response is defined at the

WATRI is a joint venture between The University of Western Australia and Curtin University of Technology. This work was supported by the Australian Research Council (ARC) grant no. A00105530.

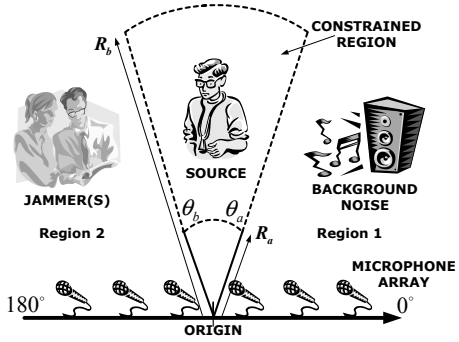

Fig. 1. *The constrained region contains the speech source as defined by the angles [θ_a, θ_b] and the radii [R_a, R_b].*

origin of coordinates. The interference statistics and arrival angles are assumed unknown.

3. THE PROPOSED STRUCTURE

3.1. Overview of the Scheme

The structure of the proposed robust microphone array is shown in Figure 2. Firstly, the soft constrained beamformer (source) suppresses all sidelobes simultaneously in each subband. Consequently, interference outside the constrained region is greatly suppressed at this stage. The second beamformer (interference) on the other hand passes the interference and suppresses the target signal. With this in mind, the residual of the interference in the first beamformer outputs can be further spectral subtracted using the estimated interference spectrum from the second beamformer. Put simply, the scheme performs double interference suppressions whilst maintaining negligible target signal distortion. Each of the blocks is explained in the following subsections.

3.2. Analysis & Synthesis Filter Banks

A uniform over-sampled analysis DFT filter bank is employed to decompose each of the L microphone input signals into M subbands with a decimation factor of $\frac{M}{2}$. Likewise, a synthesis filter bank is used to reconstruct the subband signals into fullband representation. Both filter banks are designed with the methodology described in [4], where transformation and reconstruction aliasing effects are minimized.

3.3. Soft Constrained Beamformer (Source)

The soft constrained beamformer is based on the idea proposed by Grbić and Nordholm [3]. It makes use of the Wiener solution where the source covariance matrix is obtained from the specified constrained region shown in Figure 1. The constraint mentioned is calculated from known source position(s) and the predefined array geometry. The

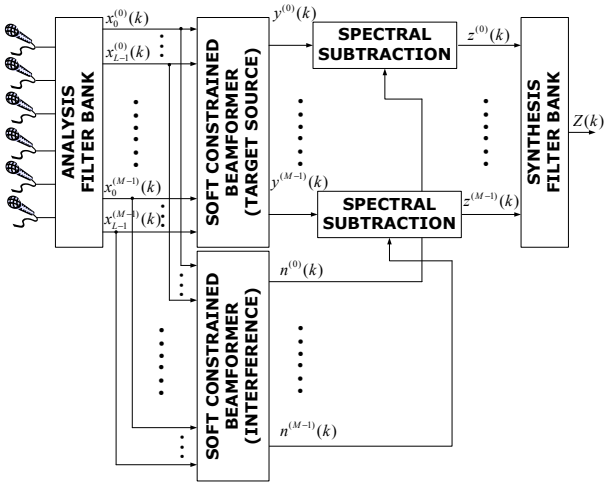

Fig. 2. *Structure of the proposed robust microphone array.*

interference (noise and jammer) covariance matrix on the other hand is estimated from the received data.

Mathematically, given the known array geometry and a corresponding constrained region, our goal is to calculate the set of optimal weights

$$\mathbf{w}_{opt(s)}^{(m)} = \left[\mathbf{R}_s^{(m)} + \hat{\mathbf{R}}_i^{(m)}\right]^{-1} \mathbf{r}_s^{(m)} \tag{2}$$

where the array weight vector, $\mathbf{w}_{opt}^{(m)}$, for the mth frequency band is

$$\mathbf{w}_{opt(s)}^{(m)} = [w_1^{(m)} \; w_2^{(m)} \; \ldots \; w_L^{(m)}]^T. \tag{3}$$

The source covariance matrix is given by

$$\mathbf{R}_s^{(m)} = \int \int_{R_a,\theta_a}^{R_b,\theta_b} S(\Omega_m)\mathbf{d}(R,\theta,\Omega_m)\mathbf{d}(R,\theta,\Omega_m)^H dR d\theta \tag{4}$$

where $S(\Omega_m)$ is the source power spectral density (PSD) of the mth subband. The interference covariance matrix, $\hat{\mathbf{R}}_i^{(m)}$ for mth subband are estimates from K samples of received data during source "silence" periods i.e. when the interference is active,

$$\hat{\mathbf{R}}_i^{(m)} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{x}_i^{(m)}(k)\mathbf{x}_i^{(m)}(k)^H. \tag{5}$$

The cross covariance vector, $\mathbf{r}_s^{(m)}$, is given by the response vector and the source PSD

$$\mathbf{r}_s^{(m)} = \int \int_{R_a,\theta_a}^{R_b,\theta_b} S(\Omega_m)\mathbf{d}(R,\theta,\Omega_m) dR d\theta. \tag{6}$$

The beamformer output for the mth subband is then

$$y^{(m)}(k) = \mathbf{w}_{opt(s)}^{(m)}(k)^H \mathbf{x}^{(m)}(k), \tag{7}$$

where $\mathbf{x}^{(m)}(k)$ is the received array data vector in the mth frequency band.

3.4. Soft Constrained Beamformer (Interference)

The principle of the second beamformer is the same as those mentioned in the previous section. The primary difference is that this beamformer passes the interference and blocks the target signal. Here, two set of weights are calculated. Each set corresponds to the regions 1 and 2 outside the constrained area shown in Figure 1. The first set is given as

$$\mathbf{w}_{set1}^{(m)} = \left[\mathbf{R}_i^{(m)} + \mathbf{R}_s^{(m)}\right]^{-1} \mathbf{r}_i^{(m)}, \tag{8}$$

where the interference covariance matrix is obtained by integrating from angle $[0^\circ\ \theta_a]$ defined as

$$\mathbf{R}_i^{(m)} = \int \int_{R_a,0^\circ}^{R_b,\theta_a} S(\Omega_m)\mathbf{d}(R,\theta,\Omega_m)\mathbf{d}(R,\theta,\Omega_m)^H dR d\theta. \tag{9}$$

The source covariance matrix $\mathbf{R}_s^{(m)}$ is defined in Eq. (4) and the cross covariance vector $\mathbf{r}_i^{(m)}$ is calculated using Eq. (6) but with area of integration corresponds to region 1. Likewise, the second set of weights is calculated in the same manner but it covers region 2 as specified by angle $[\theta_b\ 180^\circ]$.

Both sets of weights are then added to form the optimal weights for the mth band as

$$\mathbf{w}_{opt(i)}^{(m)} = \mathbf{w}_{set1}^{(m)} + \mathbf{w}_{set2}^{(m)}. \tag{10}$$

Therefore, the output for the mth subband from this beamformer is

$$n^{(m)}(k) = \mathbf{w}_{opt(i)}^{(m)}(k)^H \mathbf{x}^{(m)}(k). \tag{11}$$

Clearly, this beamformer relies solely on the precalculated covariance and cross covariance information. The output contains mainly the interference and is used to enhance the target signal further in the next section.

3.5. Spectral Subtraction

Since the implementation is in the frequency domain, spectral subtraction can be readily performed. Each of the $n^{(m)}(k)$ signals from the second beamformer is first partitioned into P sub-blocks as

$$\mathbf{n}_p^{(m)} = \Big[n^{(m)}(K/P \cdot p), n^{(m)}(K/P \cdot p - 1), \ldots, n^{(m)}(K/P \cdot (p-1) + 1)\Big] \tag{12}$$

where $p = 1, 2, \ldots, P$ and K is the data length. Similarly, the outputs from the first beamformer is partitioned in the manner defined in Eq. (12). The gain function of the pth sub-block is given as

$$\mathbf{G}_p^{(m)} = \left(1 - g_p^{(m)} \left(|\mathbf{n}_p^{(m)}| \oslash |\mathbf{y}_p^{(m)}|\right)\right) e^{-j\pi m(1+M/2)}, \tag{13}$$

where $\oslash$ denotes elementwise division and $|\cdot|$ represents the absolute value of each element in the vector. The vectors $\mathbf{n}_p^{(m)}$ and $\mathbf{y}_p^{(m)}$ are the mth subband of the output signal of the second beamformer and the first beamformer output of the pth sub-block respectively. The exponential function is included to introduce a phase to the gain function for causality.

The parameter $g_p^{(m)}$ in Eq. (13) adjusts the desired interference reduction in each pth block of the mth subband signal. Here, different values of $g_p^{(m)}$ are estimated for each sub-block in each subband. The novel method estimates each "$g_p^{(m)}$" during periods of silence by dividing the pth block of the mth subband first beamformer output to that of the second beamformer. As such, different levels of interference in each subband can be determined for maximum interference reduction with minimum target signal distortion. A simple exponential averaging described by

$$\bar{\mathbf{G}}_p^{(m)} = (1-\alpha)\bar{\mathbf{G}}_{p-1}^{(m)} + \alpha \mathbf{G}_p^{(m)}, \tag{14}$$

is then used to reduce the variance of the calculated gain function. $\bar{\mathbf{G}}_p^{(m)}$ is the exponential average for the current block and α controls the length of the exponential memory. Finally, the pth sub-block in the mth subband of the spectral subtraction output is

$$\mathbf{Z}_p^{(m)} = \bar{\mathbf{G}}_p^{(m)} \odot \mathbf{y}_p^{(m)}, \tag{15}$$

where $\odot$ denotes elementwise multiplication of vectors. All the P sub-blocks in the mth subband are then recombined to form the K length block.

4. SIMULATIONS

4.1. Office Environment

The performance evaluation of the proposed structure was made in an office ($313 \times 345 \times 301$ cm) with a six element microphone array (1/2" free field Larson-Davis) and sampled at 8 kHz. The inter-element distance was 4 cm and the speech source was located 50 cm from the centre of the array at an angle 90°. Two experimental setups for the interference were considered. The first setup consisted of a directional noise source at an angle 40° and a jammer at an angle 140°. Both were at radii 60 cm and 50 cm from the centre of array, respectively. The second setup was made in a diffuse noise environment and jammer at the position mentioned above. The reverberation time of the room was measured to be in the order of 400 ms. The purpose of having such difficult experimental setups was to test the robustness of the structure as opposed to having it in a highly unlikely ideal environment.

4.2. Results

We now compare the results obtained using the proposed structure to those obtained employing only the soft con-

SNR	Soft Constrained			Proposed Structure		
	Noise Supp.	Jam. Supp.	Dist.	Noise Supp.	Jam. Supp.	Dist.
-5	9.1	13.4	-31.9	13.4	16.1	-31.5
0	9.0	14.9	-32.2	12.8	17.7	-31.8
5	9.0	13.0	-32.5	13.0	18.7	-32.2
10	8.4	16.3	-32.6	11.9	18.7	-32.3
15	7.1	16.7	-32.5	9.7	19.6	-32.2
	Soft Constrained			Proposed Structure		
-5	4.0	10.1	-32.9	7.0	12.7	-32.3
0	4.0	11.8	-32.6	6.8	14.6	-32.3
5	4.2	13.7	-33.8	6.1	16.2	-33.2
10	3.8	15.1	-33.8	6.0	15.1	-33.5
15	2.5	16.0	-33.3	4.2	19.0	-33.0
	dB	dB	dB	dB	dB	dB

Table 1. *Suppression and distortion levels for noise, jammer and source signals respectively with different SNRs of the directional noise (top) and diffuse noise (bottom).*

strained beamformer. All simulations were performed with 64 subbands. Table 1 shows the structure's suppression and distortion levels by varying the signal to noise ratio (SNR) in both setups. The signal to jammer ratio (SJR) in this case was fixed at 0 dB. Evidently, the proposed structure outperforms the soft constrained beamformer by as much as 3 to 4 dB for noise and jammer suppressions in both scenarios with negligible expense on target signal distortion. For completeness, Figure 3 shows the normalized output powers of a single sensor observation, the soft constrained beamformer and the proposed structure for noise, jammer and source respectively in diffuse noise field. The SNR and SJR are 10 and 0 dBs respectively.

5. CONCLUSIONS

A new robust microphone array has been presented. The structure utilizes two soft constrained beamformers and spectral subtraction. Results show thats the new scheme has better noise and jammer suppressions compared to use of the optimum soft constrained beamformer alone. The incorporation of spectral subtraction allows the amount of suppression to be traded off against signal integrity. All in all, the proposed scheme is robust against error and achieves very good interference suppression with low complexity.

6. REFERENCES

[1] 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, June 1999.

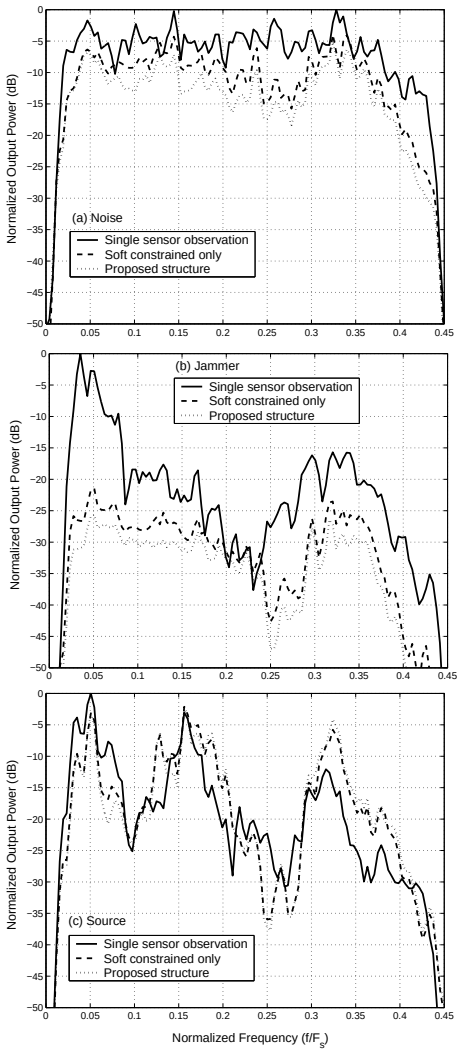

Fig. 3. *The normalized output powers of the unprocessed sensor observation, soft constrained beamformer and proposed structure for (a) noise, (b) jammer and (c) source.*

[2] J. Bitzer, K. U. Simmer, and K. D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (gsc) for speech enhancement," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, vol. 5, pp. 2965–2968, 1999.

[3] N. Grbić and S. Nordholm, "Soft constrained subband beamforming for handsfree speech enhancement," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, vol. 1, pp. 885–888, 2002.

[4] J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, "Design of oversampled uniform dft filter banks with delay specifications using quadratic optimization," *IEEE Int. Conf. on Acoust., Speech and Signal Process.*, vol. VI, pp. 3633–3636, 2001.