Adaptive Blocking Beamformer for Speech Separation

Ngoc Thuy Tran [1], *William Cowley* [2], *André Pollok* [3]

Institute for Telecommunications Research, University of South Australia

[1]tratn014@mymail.unisa.edu.au, [2]Bill.Cowley@unisa.edu.au,
[3]Andre.Pollok@unisa.edu.au

Abstract

This paper tackles the speech separation problem in a meeting room using a new acoustic beamforming method – adaptive blocking (AB) beamformer. The proposed method is an optimum beamforming with a structure similar to the generalized sidelobe canceller (GSC) structure, but simpler. Thus, it inherits the flexibility of GSC and functions well in dynamic environments. We investigate the performance of the proposed method through different experiments and compare the results with a GSC beamformer for minimum variance distortionless response (MVDR). The experimental setups include one wanted speaker, two interferers, air conditioner noise and uncorrelated sensor noise. AB provides improvement over MVDR-GSC.

Index Terms: speech separation, beamforming, optimum beamforming

1. Introduction

Speech separation has been of interest for many researchers in the last few decades [1]. One big challenge in the field is to extract wanted signals from reverberation, nonstationary interference and noise in a dynamic environment, such as a meeting room. Addressing these problems, a number of approaches have been proposed. They can be classified into: acoustic feature based methods, such as computational auditory scene analysis (CASA); blind source separation and acoustic beamforming [2, 1, 3]. Beamforming uses microphone array input signals and exploits spatial information of sound sources to direct the separation. In contrast to the other approaches, beamforming can avoid acoustic feature diversity and speech model complexity.

Beamforming includes fixed and adaptive approaches. Fixed beamformers are data independent. An example is conventional beamforming, which aims at phased-aligning the signals from the direction of interest [4]. Adaptive beamforming, in contrast, adapts to data and is updated during the separation process and possesses advantages when dealing with nonstationary signals like speech. Among available adaptive methods, optimum beamforming with GSC structure has become a major approach [5, 4]. MVDR is an example which minimizes the variance of the output subject to a constraint to avoid a distorted signal [4]. It can achieve high performance in ideal conditions, but deteriorates in reverberant environments and becomes sensitive to mismatch problems [6]. Alternatively, MVDR is often implemented through its MVDR-GSC representation.

In general, GSC is a structure in which the input signal is processed synchronously through two paths: (1) a fixed beamformer that satisfies the constraints of the optimum beamforming and (2) an adaptive path following a signal blocking step [5]. The blocking step is to remove the wanted components of the first path, and the adaptive beamformer is to guarantee the optimum solution after subtracting the output of the second path from the first path. This structure allows a high flexibility of GSC as it can accommodate different beamformer designs or integrate with other speech separation approaches. MVDR-GSC usually has a conventional beamformer in the first path, and uses a least mean squares (LMS) or recursive least squares (RLS) approach for the adaptive part. The blocking matrix blocks signals from the wanted direction and is available with a number of options [5]. In [7], the authors propose a GSC based system that uses minimum mutual information (MMI) to separate multiple sound sources. In [8], a structure similar to GSC but employing independent component analysis (ICA) is used for a robotic application and achieves about 62% word accuracy with an automatic speech recognition system. Other research, such as [9] with GSC structure employing an eigenvalue approach, indeed show the potential of GSC and its flexibility.

However, GSC faces a number of difficulties in replicating results from laboratory to practice. One concern when designing a GSC based system is the blocking matrix. A steering vector based matrix can let signals pass through in signal mismatch situations [10, 11]. The other options for a blocking matrix, such as acoustic transfer function based and eigenvector methods [5, 9] are both highly complex and are suffering from multi-path propagation and dynamic, noisy environment. Besides, the adaptive weighting vector component also faces other problems. For example, LMS or RLS need to overcome nonstationarity, speed of convergence and divergence due to changes in the recording environment.

Inspired by the ideas of GSC and optimum beamforming, we propose a new beamformer addressing speech separation in a reverberant environment, such as a meeting room. We target improving the signal quality using a simpler structure while inheriting the flexibility of GSC. The proposed beamformer, adaptive blocking, is defined with two processing paths - the first one is to keep a unit-gain of the wanted signal and the other blocks the wanted signal and minimizes the output power of interferences within only one step. The performance is assessed with both synthetic and real recordings, focussing on situations with moving speakers and different scenarios with overlapping speech. The results are compared with MVDR-GSC performance. AB beamformer provides apparent improvement over MVDR-GSC.

2. Signal Model

We address the separation problem in a meeting room with P nonstationary sound sources. Speech from a wanted speaker s_1 needs to be extracted from the mixed input signal recorded by an M element microphone array. At microphone m, the received signal at sample n is:

$$x_m(n) = \sum_{p=1}^{P} h_{p,m}(n) * s_p(n) + \nu_m(n), \qquad (1)$$

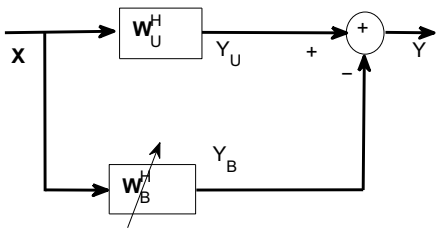


Figure 1: *Adaptive Blocking beamformer structure*

where $*$ is the convolution, $s_2, .., s_P$ are interferers and correlated noises. The uncorrelated noise at microphone m is denoted ν_m, the room impulse response (RIR) between source p and microphone m is $h_{p,m}$. This RIR is dynamic, depending on various factors in the recording environment, such as speaker movement, air temperature and room configuration.

The input signal is transformed into the frequency domain using a fast fourier transform (FFT) and we apply beamforming for each frequency bin. The final output is transferred back into the time domain using an inverse FFT [12]. The proposed beamformer is derived in the frequency domain. For the sake of more compact notation, we will omit the frequency bin index. The column vector of input signal across the microphones for frame $K \geq 1$ is $\boldsymbol{X}(K) = \begin{bmatrix} X_1(K) & \cdots & X_M(K) \end{bmatrix}^T$, where X_m is the FFT of x_m and $(.)^T$ is the transpose operation.

3. Adaptive Blocking Beamformer

Figure 1 presents the proposed beamformer structure, which aims at minimizing the output power of the interference plus noise. The top path of the structure (beamformer $\boldsymbol{W_U}$) maintains a unit-gain at the wanted direction to avoid signal distortion. The bottom path (adaptive weighting vector $\boldsymbol{W_B}$) blocks signals from this direction and minimizes the final output power of unwanted signals in $|Y|^2$.

$$Y_U(K) = \boldsymbol{W_U}^H(K)\boldsymbol{X}(K) \quad , Y_B(K) = \boldsymbol{W_B}^H(K)\boldsymbol{X}(K),$$
$$Y(K) = Y_U(K) - Y_B(K),$$

where the weighting vectors are given by

$$\boldsymbol{W_U}(K) = \begin{bmatrix} W_{U1}(K) & \cdots & W_{UM}(K) \end{bmatrix}^T$$
$$\boldsymbol{W_B}(K) = \begin{bmatrix} W_{B1}(K) & \cdots & W_{BM}(K) \end{bmatrix}^T,$$

and $(.)^H$ denotes the conjugate transpose operation.

We assume a pre-knowledge of the wanted speaker's location, which can be identified by a steering vector $\boldsymbol{A}$ [4]. The structure design is satisfied by having a unit-gain weighting vector for the look direction $\boldsymbol{W_U}^H\boldsymbol{A} = 1$, and an optimum beamformer $\boldsymbol{W_B}$ that is formalized as follows.

In each frequency bin, the blocking weighting vector $\boldsymbol{W_B}(K)$ is the solution of the optimization problem:

$$\min_{\boldsymbol{W_B}(K)} \sum_{k=K-T+1}^{K} \mu^{K-k} |Y_U(k) - \boldsymbol{W_B}(K)^H \boldsymbol{X}(k)|^2 \quad (2)$$
$$\text{subject to} \quad \boldsymbol{W_B}(K)^H \boldsymbol{A} = 0,$$

where $0 < \mu < 1$ is the forgetting factor, T is the number of samples that contribute to the cost function. To simplify notations, the frame index K will be dropped.

In (2), we introduced two parameters to control the forgetting process – the exponential weight μ and the memory window length T. μ is used with the same purpose of the forgetting factor in RLS. The contribution of earlier data to the minimization shrinks over time. However, in order to reduce the memory length, we need to reduce μ, which also means the memory window is steeper. To give more freedom to control this process, we use the window length T. All data outside the window is erased, independently of the forgetting speed within the window. For example, if the samples within T frames are highly correlated but the changing is fast, we can have short T and big μ. Short memory length becomes an advantage when data for beamformer adaptation is limited. The disadvantage of using T is programming complexity and computational memory.

The problem (2) can be solved by a Lagrange multiplier method in the complex-valued space [13]. Let

$$\boldsymbol{R}_{XX} = \sum_{k=K-T+1}^{K} \mu^{K-k} \boldsymbol{X}(k)\boldsymbol{X}(k)^H, \quad (3)$$

$$\boldsymbol{D}_{YX} = \sum_{k=K-T+1}^{K} \mu^{K-k} Y_U(k)^* \boldsymbol{X}(k), \quad (4)$$

where $(.)^*$ denotes the complex conjugate. Omitting detailed derivations, we obtain the closed form solution for the Adaptive Blocking weighting vector:

$$\boldsymbol{W_B} = \boldsymbol{R}_{XX}^{-1} \left(\boldsymbol{D}_{YX} - \frac{\boldsymbol{A}^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{D}_{YX}}{\boldsymbol{A}^H \boldsymbol{R}_{XX}^{-1} \boldsymbol{A}} \boldsymbol{A} \right). \quad (5)$$

Notice that the unit-gain weighting vector $\boldsymbol{W_U}$ can be any weighting vector satisfying $\boldsymbol{W_U}^H\boldsymbol{A} = 1$. It can be either a fixed beamformer, such as a conventional beamformer, or an adaptive beamformer with additional constraints. The blocking weighting vector $\boldsymbol{W_B}$ is calculated using (5) and diagonal loading technique can be applied for $\boldsymbol{R}_{XX}^{-1}$ estimation [4]. The overall beamformer of AB is closely related to MVDR. If $\boldsymbol{W_U}$ is a conventional beamformer and $\boldsymbol{W_B}$ defined in (5) has a growing memory (infinite T, $\mu = 1$, and a normalizing factor), AB will be equal to MVDR.

Deriving from (3) and (4), the matrix $\boldsymbol{R}_{XX}$ and the vector $\boldsymbol{D}_{YX}$ can be calculated recursively as follows:

$\forall 1 \leq K \leq T$

$$\boldsymbol{R}_{XX}(K) = \mu \boldsymbol{R}_{XX}(K-1) + \boldsymbol{X}(K)\boldsymbol{X}^H(K) \quad (6)$$
$$\boldsymbol{D}_{YX}(K) = \mu \boldsymbol{D}_{YX}(K-1) + Y_U(K)^* \boldsymbol{X}(K) \quad (7)$$

$\forall K > T$

$$\boldsymbol{R}_{XX}(K) = \mu \boldsymbol{R}_{XX}(K-1) + \boldsymbol{X}(K)\boldsymbol{X}^H(K) - \mu^T \boldsymbol{X}(K-T)\boldsymbol{X}^H(K-T) \quad (8)$$
$$\boldsymbol{D}_{YX}(K) = \mu \boldsymbol{D}_{YX}(K-1) + Y_U^*(K)\boldsymbol{X}(K) - \mu^T Y_U^*(K-T)\boldsymbol{X}(K-T). \quad (9)$$

Ideally, the adaptation on $\boldsymbol{W_B}$ should be implemented continually, using (5)-(9), to keep the advantage of adaptive beamforming in dealing with nonstationary signals and dynamic environments. Considering the correlation between signals, however, data used for training and adaptation must not include the wanted signal. Otherwise, the power output of interference plus noise will not be minimized [11]. This leads to the challenge of training process in practice. We will show that when the beamformer is not updated in a long period, the cancelation will be seriously affected. Specific training schemes used in this paper are described in the next section.

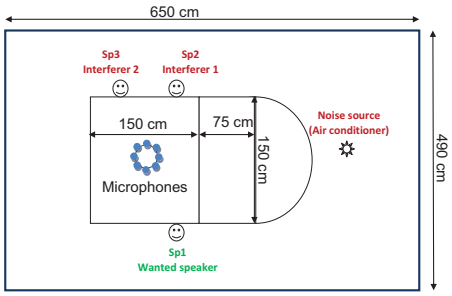


Figure 2: *Meeting room*

4. Performance

4.1. Scenarios and Setup

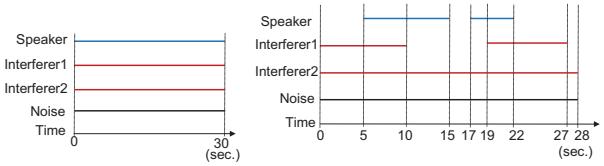
To evaluate the performance of the proposed beamformer, a large number of test cases for both synthetic and real recordings were performed and compared with an MVDR-GSC beamformer. Results are evaluated through (1) subjective listening method and (2) objective methods: signal to interference plus noise ratio (SINR) and log spectral distance (LSD) measure [14, Eq.(42)]. In general, AB shows improvement over GSC. This paper presents some key results using two interferers. We examine the effect of overlapping scenarios, the environment's dynamic level and training data availability on the performance of the beamformer. The input and output signals in this paper and more test cases are available at [15].

Speech signals used for simulations are originally from ES2009b-AMI corpus [16]. The recording setup is shown in Figure 2. The microphone array is an 8 element circular array with a radius of 10cm. Headset recordings are used as reference signals. In synthetic scenarios, we simulate the same room setting and imitate the reverberation process. RIRs with reverberation time of 0.3sec (T_{60}) were generated by a simulator available in [17]. The RIRs change depending on the movement of the speakers, causing a dynamic recording environment. Different movement scenarios for interferer 1 are listed in Table 1. Interferer 2 and the wanted speaker do not move. Signals and noises are then convolved with the corresponding RIRs. The final input signal is the summation of the wanted speech, interferences, the air-conditioner noise, and additional uncorrelated sensor noise (with signal to noise ratio of 33dB). We use two overlapping scenarios as shown in Figure 3. Overlapping occurs during the whole recording in scenario OVL1. Scenario OVL2 is more realistic as overlapping between the wanted speaker and interferer 1 happens only in some segments $[5\cdots10; 19\cdots22]$sec. Noise and interferer 2 always appear. In real recording scenarios, the AMI's single speaker signals recorded from the real microphone array are summed, following either OVL1 or OVL2.

The training and adaptation schemes for $\boldsymbol{W_B}$ and RLS vary across the overlapping scenarios. $\boldsymbol{W_B}$ is updated using (5)-(9). In OVL1, overlapping occurs during the whole recording, and we use the whole interference plus noise sig-

Table 1: *Interferer* 1 *movement scenarios*

ID	Motion Type	Motion Description
NS1	Uniform circular	Radius: .2m, speed: 1rad/s
NS2	Uniform circular	Radius .5m, speed .1rad/s
NS3	Non-uniform linear	Furthest distance: .5m, speed: .3m/s, acceleration: .05m/s^2



(a) OVL1 (b) OVL2

Figure 3: *Overlapping scenarios showing periods of signal activity*

nal to adapt the weighting vector in every frame. This is equivalent to study the beamformers' performance with full knowledge of the correlation matrix of the signals. However, this ideal condition is not available in reality. In OVL2, instead, a feasible adaptation is employed by using only segments without active wanted speaker. We presume that information about periods of signal activity is known. Thus, only segments $[0\cdots5; 15\cdots17; 22\cdots28]$sec are used for training.

The input and output SINRs are measured for synthetic data following the methods described in [9]:

$$SINR_I = 10\log_{10}\frac{\sum_k\sum_{l\in \mathrm{FR}}|S_W(k,l)|^2}{\sum_k\sum_{l\in \mathrm{FR}}|S_{IpN}(k,l)|^2}, \quad (10)$$

$$SINR_O = 10\log_{10}\frac{\sum_k\sum_{l\in \mathrm{FR}}|Y_W(k,l)|^2}{\sum_k\sum_{l\in \mathrm{FR}}|Y_{IpN}(k,l)|^2}, \quad (11)$$

where k and l are frame index and frequency index respectively, FR indicates the frequency range used for estimating SINR. We use k over the whole recording and FR $=[200\cdots7000](Hz)$. Y_W, Y_{IpN} are the desired signal component and the interference plus noise component in the output respectively. S_W, S_{IpN} are the desired signal component and the interference plus noise component in the input of a microphone (microphone 1). These input components consist of multi-path propagation signals.

The output SINR for each frequency bin computed as:

$$SINR_B(l) = 10\log_{10}\frac{\sum_k|Y_W(k,l)|^2}{\sum_k|Y_{IpN}(k,l)|^2} \quad (12)$$

is shown in Figure 4.

AB and GSC are designed with a conventional beamformer in the top path. Steering vector based blocking matrix [5] and RLS adaptive beamforming [4] complete the other path of MVDR-GSC. In synthetic cases, the location of the wanted speaker is known, avoiding localization errors. Based on the given speaker locations, steering vectors for each frequency bin are calculated (eg. [10, Eq.(11.1.16)]). In the real recordings, the locations are estimated to have an error up to 0.5m. The forgetting factors for RLS and AB are $\mu = 0.85$, and a window of 50 frames (1.6sec) in synthetic cases, 100 frames (3.2sec) in real recordings for the forgetting limit T of AB. We also ran simulations with infinite T for AB to compare with the standard RLS. AB still gave better results. Signals are block-wise processed with FFT length of about 64msec, 50% overlapping between frames and using Hanning windows. We apply diagonal loading for $\boldsymbol{R}_{XX}^{-1}$, with a relatively small factor α:

$$\alpha = \begin{cases} 10^{-6}, & \min_i(|\boldsymbol{R}_{XX}(i,i)|) = 0 \\ 10^{-4}\min_i(|\boldsymbol{R}_{XX}(i,i)|), & \text{otherwise.} \end{cases}$$

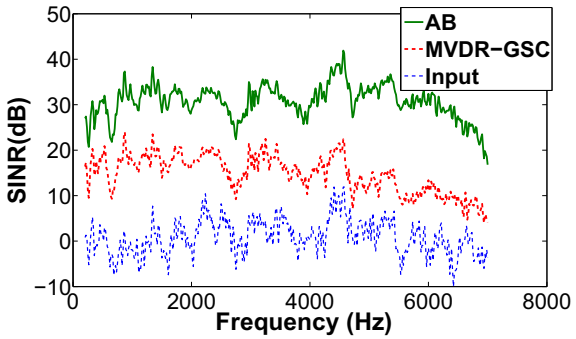


Figure 4: *$SINR_B$ across frequencies, OVL1, row* 1 *Table 2*

4.2. Results and Discussion

Table 2 shows the objective results for each test case. In general, AB gives better results than GSC in all scenarios. AB achieves high SINR - from 12 to 27dB increase, 12dB higher than MVDR-GSC in the genie-aided synthetic cases (OVL1). Furthermore, the distortion of AB measured in term of LSD is lower by about 0.2. This result is also subjectively supported in listening tests. The SINR improvement of AB happens consistently in all frequency bins, not only at high frequencies, as illustrated in Figure 4 for the test case 1, row 1 Table 2. The improved performance of AB can be explained as the positive effect of the simpler design in the bottom path of AB. Both AB and MVDR-GSC use the bottom path to block the wanted signal coming from the wanted direction. However, GSC becomes more vulnerable as it relies on both the blocking matrix and RLS. Either signal leaking in the blocking matrix or changes in the non-stationary input signal can lead to unstable performance of RLS and therefore suboptimum interference cancellation.

Secondly, from Table 2 we observe the impact of the dynamic environment on the beamformers. In the tests using signal NS1,2,3, the environment changes faster. This results in greater nonstationarity. Thus, adaptive weighting vectors need to converge faster with more stability, which is shown to be better with AB. In the synthetic recordings, AB keeps a quite consistent SINR gain over GSC, about 2 to 12dB varying across different overlapping scenarios. In real recording tests, the speakers move in smaller areas, but the RIRs changing is more complicated and longer than the synthetic case. Thus, the performance is impacted more by the reverberation. Consequently, the outputs have higher distortion, yet AB still performs better than MVDR-GSC.

Lastly, we observe the impact of training data availability by comparing results between scenarios using OVL1 and OVL2. OVL2 uses a practical training strategy, which allows vector adaption if the wanted speech is inactive. In contrast, OVL1, we have genie-aided full training for every frame. Consequently, the SINR improvement drops from about 26dB to about 12dB, and the LSD increases from around 0.7 up to more than 1. This drop is due to the lack of adaptation in the long period - about 15sec - in the middle of the recording, whilst the interference is highly nonstationary.

Table 2: *Beamformer results*

ID	OVL	Interfe rer1	$SINR_I$ (dB)	$SINR_O$(dB)		LSD	
				GSC	AB	GSC	AB
1	OVL1	NS1	1	16	28	0.98	0.75
2	OVL1	NS2	6	22	33	0.84	0.71
3	OVL1	NS3	2	15	27	1.04	0.78
4	OVL2	NS1	0	12	12	1.17	1.07
5	OVL2	NS2	6	16	18	1.05	1.03
6	OVL2	NS3	1	12	14	1.18	1.06
7	OVL1	real	–	–	–	2.17	1.9
8	OVL2	real	–	–	–	1.9	1.64

5. Conclusions

The new AB beamformer has been proposed and examined under a number of synthetic and real recording scenarios. AB has the advantage of a simpler design in the blocking path and shows considerable improvement over MVDR-GSC when we choose the conventional fixed beamformer for the top path. Notice that AB allows any unit-gain beamformer for this path, including an adaptive beamformer. This mechanism can be exploited to help the system adapt to movement in the look direction. Besides, the forgetting process of AB helps it to adapt more efficiently with changes of environment.

Similarly to other adaptive beamformers, AB is challenged by the dynamic environment and training data problem. Experimental results show that if we have enough data for the adaptation process, the beamformer can claim high SINR improvement and good cancellation.

Future research regarding AB approach will explore an adaptive training strategy that handles varying speakers locations and overlapping speech.

6. References

[1] S. Markovich, S. Gannot, and I. Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 17, no. 6, pp. 1071–1086, 2009.

[2] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 1st ed. Addison-Wesley Publishing Company, 1987.

[3] J. Benesty, M. M. Sondhi, and Y. A. Huang, Eds., *Springer Handbook of Speech Processing*. Springer, 2008.

[4] H. L. V. Trees, *Optimum Array Processing Part IV of Detection, Estimation, and Modulation Theory*, 1st ed. Wiley-Interscience, 2002.

[5] S. Gannot and I. Cohen, "Adaptive beamforming and postfiltering," in *Springer Handbook of Speech Processing*. Springer, 2008.

[6] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *IEEE Trans. Sig. Proc.*, vol. 51, no. 7, pp. 1702–1715, 2003.

[7] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wolfel, "Adaptive beamforming with a minimum mutual information criterion," *IEEE Trans. on Audio, Speech, and Language Proc.*, vol. 15, no. 8, pp. 2527–2541, 2007.

[8] Y. Takahashi, K. Osaka, H. Saruwatari, and K. Shikano, "Blind source extraction for Hands-Free speech recognition based on wiener filtering and ICA-Based noise estimation," in *Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 164–167.

[9] A. Pezeshki, B. V. Veen, L. Scharf, H. Cox, and M. Nordenvaad, "Eigenvalue beamforming using a multirank MVDR beamformer and subspace selection," *IEEE Trans. Sig. Proc.*, vol. 56, no. 5, pp. 1954–1967, 2008.

[10] D. G. Manolakis, V. K. Ingle, and S. M. Kogon, *Statistical and Adaptive Signal Processing: Spectral Estimation, Signal Modeling, Adaptive Filtering and Array Processing*. McGraw-Hill Science/Engineering/Math, 1999.

[11] H. Cox, "Resolving power and sensitivity to mismatch of optimum array processors," *The Journal of the Acoustical Society of America*, vol. 54, no. 3, pp. 771–785, 1973.

[12] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice Hall, 1999.

[13] D. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proceedings on Communications, Radar and Signal Processing*, vol. 130, no. 1, pp. 11 –16, 1983.

[14] I. Cohen, "Relaxed statistical model for speech enhancement and a priori SNR estimation," *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 5, pp. 870–881, 2005.

[15] "Adaptive blocking beamforming," http://www.itr.unisa.edu.au/∼tratn014.

[16] "AMI corpus MMM server - MultiModal media fileserver," http://www.idiap.ch/mmm/corpora/ami.

[17] E. Habets, "Room impulse response generator," http://home.tiscali.nl/ehabets/rir_generator.html.