An Efficient GSC VSS-APA Beamformer with Integrated Log-energy Based VAD for Noise Reduction in Speech Reinforcement Systems

Marius Rotaru[1], Silviu Ciochină[1], Felix Albu[2]
[1] Dept. of Telecommunications, University Politehnica of Bucharest, Bucharest, Romania
marius.rotaru@gmail.com, silviu@comm.pub.ro
[2] Electrical Engineering Department,Valahia University of Targoviste, Targoviste, Romania
felix.albu@valahia.ro

***Abstract*—This paper presents an efficient time-domain Generalized Sidelobe Canceller (GSC) with low signal distortion capabilities using the variable step size affine projection algorithm (VSS-APA) and a log-energy based voice activity detector (VAD). The performance of the proposed VSS-APA based GSC method with integrated log-energy VAD, is illustrated in the context of speech reinforcement application using different signals with low signal-to-noise ratio and different types of noise.**

I. INTRODUCTION

The communication in vans and limousines between the passengers in the front and the rear is degraded due to the presence of the noise as well as the long distance between them [1]. This can be improved by using a speech reinforcement system. One important component of such system is the noise reduction stage. Since the positions of driver and front passenger are known, a fixed microphone array implementing a fixed beamformer has superior performances to those of single microphone noise reduction techniques [2]

Spatial filtering is a powerful technique of enhancing a signal of interest while suppressing the interference signal (e.g. feedback signal of speech reinforcement systems) and the noise at the output of an array of sensors. The array consists of a number of microphones that are spatially placed at known locations and used to simultaneously sample the data. Because of their effectiveness in enhancing the quality of signal of interest, microphone arrays have been widely studied [3-6].

By the means of an adaptive optimization algorithm, which tune the phase and the amplitude of signal wave at each sensor, it is possible to electronically steer the beam to a desired direction, usually the direction of the signal of interest and to place nulls in other directions, which corresponds to the undesired interference signal or jammer. Such adaptive beamformers have been used in a wide range of applications, e.g. antennas, radars, wireless communication, biomedical signal processing, speech processing [4-6].

By combining temporal and spatial filtering, a directive microphone array can be extended to a broadband adaptive beamformer. Such beamformer consists of a multi-input single output linear combiner and an adaptive algorithm that adjusts the weights of linear filters based on different optimization criteria [7-9]. Different beamformer characteristics may result depending on the way the filter coefficients are selected. A well studied beamformer is the linearly constrained minimum variance (LCVM) beamformer. Its weight factors are selected in order to minimize the output variance of a beamformer, and constrained such that the signal from the direction of interest is captured with a specified gain and phase.

An efficient adaptive implementation of the LCMV method was proposed in [10] and it is known as the Generalized Sidelobe Canceller (GSC). Basically the GSC is a method that changes the constrained optimization problem into an un-constrained one. This technique is very popular due to low computational complexity and simplicity in real-time system realization.

To overcome the slow convergence rate of LMS based GSC technique, which depends on the eigenvalue spread of the correlation matrix of the input data, different implementations of the variable step size affine projection algorithm (VSS-APA) proposed in [11] have been used [12-13].

It is well known that the noise reduction performance of the GSC depends on the validity of a priori assumptions about the signal model, and therefore, a VAD needs to be employed. A low complexity, log-energy based VAD has been selected and integrated with the GSC beamformer [15]. In this paper we propose an efficient way to integrate the VAD with the GSC beaformer. Our simulations have shown that the usage of VSS-APA algorithm minimizes the effect of VAD's wrong decisions, increasing the robustness of GSC system.

This paper is organized as follow: Section II is dedicated to the description of the GSC beamformer. Section III introduces the VSS APA algorithm. The Section IV presents the integration of the log-energy VAD. In Section V the experiments and simulation results are reported. Section VI concludes the work.

II. GENERALIZED SIDELOBE CANCELLER

The structure of a GSC beamformer is illustrated in Fig. 1 [10]. It consists of a fixed beamformer (FBF) characterized by a quiescent weight vector $\mathbf{w}_q$, a blocking matrix (BM) $\mathbf{C}_a$ and a multiple-input adaptive interference canceller (AIC) defined by an adaptive weight vectors $\mathbf{w}_i$.

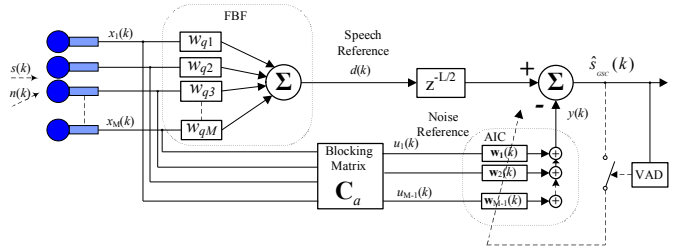


Figure 1. Generalized sidelobe canceller (GSC)

The FBF, which is usually a delay-sum beamformer, enhances the desired signal component $d(k)$ in the look direction, being obtained from the outputs of M sensors $x_i(k)$, i=1...M. The FBF sums up the steered sensor signals:

$$d(k) = \sum_{i=1}^{M} w_{qi} x_i(k), \tag{1}$$

where

$$x_i(k) = s(k) + n_i(k), \tag{2}$$

with $s(k)$ the desired target signal and $n_i(k)$ a background noise/interferences. Because the source (position of the driver) is directly in front of the microphone array, the fix delay-sum beam is obtained by a simple averaging of microphone signals $x_i(k)$.

The blocking matrix (BM), with M-1 linearly independent rows which sum-up to zero, (e.g. for M=4 [10]):

$$BM = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \tag{3}$$

rejects the desired target signal by allowing only the interferences arriving from other direction than the look direction to pass through. The output of the BM consisting of interference signal only $u_j(k)$, j = 1…M-1, will be adaptively cancelled from the reference signal $d(k)$, which is the output of FBF. In order to respect the causality of the system, a delay line of L/2 samples needs to be added on the FBF path, where L is the length of adaptive filters.

III. VSS-APA Algorithm for GSC Beamformer

To overcome both the slow convergence of LMS based algorithm and VAD failure rate (when speech and noise frames are classify as noise only frames) issues, the variable step size affine projection algorithm (VSS-APA) proposed in [11], developed in context of echo cancellation, was selected for AIC's adaptive filters. The AIC subsystem can be seen as a MISO (multiple-input single output) system composed of a bank of adaptive filter. The output of each adaptive filter $y_i(k)$ is cancelled from the speech reference signal $d(k)$, yields an estimate of beamformer output $\hat{s}_{GSC}(k)$. Considering $\mathbf{y}_i(k) = \left[y_i(k), y_i(k-1), ..., y_i(k-P+1)\right]^T$ the output vector of the last P values corresponding to the channel i of AIC subsystem:

$$\mathbf{y}_i(k) = \frac{1}{M-1}\left(\mathbf{U}_i^T(k)\mathbf{w}_i(k)\right), \tag{4}$$

with P the projection order, M the number of microphones, i a channel of AIC subsystem ($i = 1...M-1$), $\mathbf{w}_i(k) = \left[w_{i0}, w_{i1}, ..., w_{i(L-1)}\right]^T$ the coefficient vector, $\mathbf{U}_i(k)$ is a $L \times P$ reference noise matrix, $\mathbf{U}_i(k) = \left[\mathbf{u}_i(k), \mathbf{u}_i(k-1), ..., \mathbf{u}_i(k-P+1)\right]$, with $\mathbf{u}_i(k) = \left[u_i(k), u_i(k-1), ..., u_i(k-L+1)\right]^T$, the VSS-APA [11] for a channel i based approach is defined by the following equations:

$$\hat{\mathbf{s}}_{i,GSC}(k) = \mathbf{e}_i(k) = \mathbf{d}(k) - \mathbf{y}_i(k), \tag{5}$$

$$\mathbf{w}_i(k+1) = \mathbf{w}_i(k) - \mathbf{U}_i(k)\left[\mathbf{U}_i^T(k)\mathbf{U}_i(k) + \delta\mathbf{I}\right]^{-1}\boldsymbol{\mu}_i(k)\mathbf{e}_i(k), \tag{6}$$

$$\boldsymbol{\mu}_i(k) = diag\left\{\mu_{i0}(k), \mu_{i1}(k), ..., \mu_{iP-1}(k)\right\}, \tag{7}$$

$$\mu_{il}(k) = \left| 1 - \frac{\sqrt{\left|\hat{\sigma}_d^2(k-l) - \hat{\sigma}_{y_i}^2(k-l)\right|}}{\hat{\sigma}_{e_{i(l+1)}}(k) + \xi} \right|. \tag{8}$$

The parameter $\hat{\sigma}_\alpha^2(k)$ denotes the power estimate of the sequence $\alpha(k)$, and can be computed as

$$\hat{\sigma}_\alpha^2(k) = \lambda\hat{\sigma}_\alpha^2(k-1) + (1-\lambda)\alpha^2(k), \tag{9}$$

where λ is a weighting factor chosen as $\lambda = 1 - 1/(KL)$, with $K > 1$ and ξ is a small positive regularization constant added to the denominator of $\mu_{il}(k)$ to avoid division by zero.

A low computational complexity implementation of VSS-APA that uses dichotomous coordinate descent (DCD) iterations, called VSS-DCD-APA, is a suitable choice for practical implementations, especially for high projection orders [14]. If some performance losses are acceptable, the block exact implementations could be considered [12-13].

IV. Voice Activity Detector

Even if a variable step size adaptive algorithm is used, the adaptation during speech activity considerably degrades the performance of GSC beamformer. In this case a voice activity detection (VAD) algorithm with the role of classifying the signal as either noise only or speech & noise is required.

An efficient VAD with good performances at lower SNR's and reliable for strongly nonstationary signals has been proposed in [15]. It is based on short-time smoothed log-energy distribution estimation. Based on statistics of the signal (mean and variance), the instantaneous short-time log-energy of frame signal can be distinguished from the smoothed noise log-energy

distribution using two different thresholds, for speech (*speech onset*) T_S and noise (*speech offset*) T_N, defined as:

$$\begin{aligned} T_S(m) &= \hat{\mu}_N(m) + \alpha\hat{\sigma}_N(m) \\ T_N(m) &= \hat{\mu}_N(m) + \beta\hat{\sigma}_N(m) \quad , \end{aligned} \tag{10}$$

where, $\hat{\mu}_N(m)$ and $\hat{\sigma}_N^2(m)$ represent an estimate of the noise mean and noise variance respectively, at the frame instance *r*, being continuously updated only during non-speech periods. $\alpha = 4$ and $\beta = 1.2$ [15] are factors used to define the "upper" and "lower" thresholds for speech onset and offset, respectively.

Defining the frame-based log energy of a signal *x*(*k*) as:

$$E_x(m) = \log_{10}\left(\frac{1}{L_F}\sum_{i=0}^{L_F-1} x(L_F m + i)^2\right), \tag{11}$$

the VAD algorithm works as follow:

TABLE I. LOG-ENERGY VAD ALGORITHM

1. compute $E_x(m)$
2. if ($E_x(m) \geq T_S(m-1)$), speech onset detected => VAD=1
   else VAD=1
3. if ($E_x(m) \leq T_N(m-1)$)
   speech offset detected => update noise statistics and thresholds.

The VAD can be connected either at the output of the FBF, just before the delay line or at the output of the GSC beamformer. In the first situation, the detection is made on speech reference signal *d*(*k*).

In the second case the following constrains have to be considered:

- during the first frames (until the adaptive algorithm reaches to steady state), the VAD could make wrong decisions, slowing down the adaptation process. In order to avoid this situation, during the start-up phase, the VAD has to make the decision based on the output of FBF.
- the delay introduced by the VAD could not fulfill the system requirements, e.g. to not affect the speech intelligibility the max delay of speech reinforcement systems shall not exceed 10ms [16].

The adopted solution was to connect the VAD at the GSC beamformer. Also the usage of VSS-APA algorithm, minimizes the VAD's wrong decisions, increasing the robustness of GSC beamformer. The log-energy of the signal was computed with an overlap on 160 samples. The delay introduced by VAD was only 2.5ms at 16kHz.

V. SIMULATION RESULTS

The linear microphone array used in this work was composed of 4 omni-directional microphones. The distance between the microphones is set to 2.5cm. The interference source started from the right at an angle of 30° while the speaker signal angle is 0°. The system is implemented under a sampling rate of 16kHz. To simulate the environment, Matlab Phase Array System Toolbox™ has been used. The GSC beamformer has been tested using four types of interferences: white Gaussian noise, bubble noise, non-stationary noise and car engine noise with different signal to noise ratio SNR.

The performance was measured using the output SNR metric defined as:

$$SNR = 10\log\left[\frac{\sum_k s^2(k)}{\sum_k \left(s(k)-\hat{s}(k)\right)^2}\right], \tag{12}$$

where $s^2(k)$ represents the generic clean input signal energy while $\left(s(k)-\hat{s}(k)\right)^2$ is the noise energy. Also the quality of the output signal $\hat{s}^2(k)$, expressed in mean opinion scores (MOS), has been evaluated using the ITU-T Recommendation P.862, known as the Perceptual Evaluation of Speech Quality [17].

The VSS-APA algorithm is compared against APA and NLMS algorithms. The results are shown in Fig. 3 and Fig.4.

Also, Fig. 5 shows the performances of VSS-APA algorithm when VAD is connected either to the output of FBF or to the output of GSG beamformer. The following common parameters of adaptive algorithms have been used: L = 128 (length of adaptive filter), μ = 0.2 (fix step size of APA and NLMS algorithms), K = 6 (constant for the power estimation of signals), P = 8 (projection order).

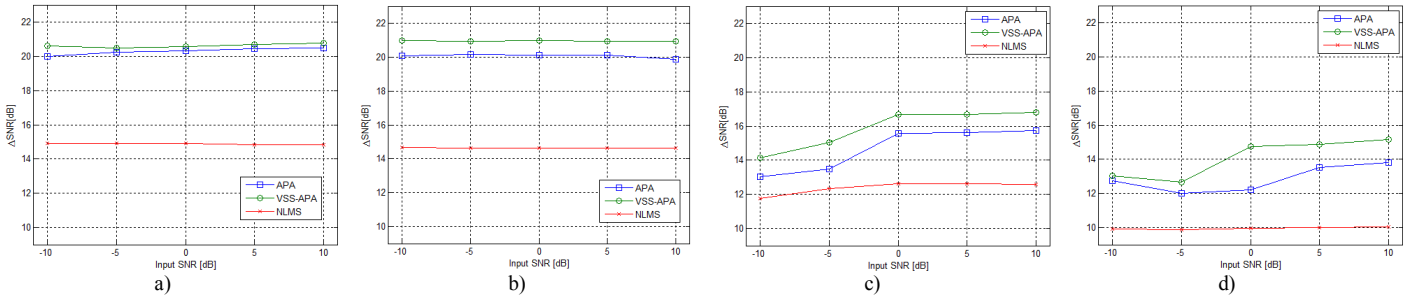
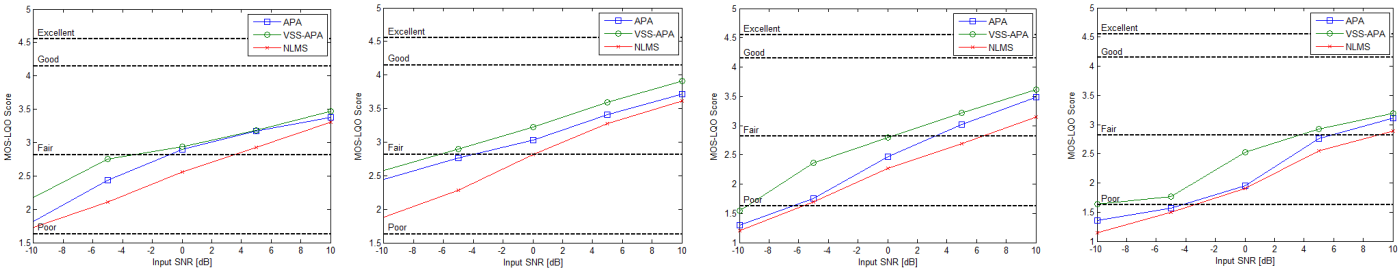
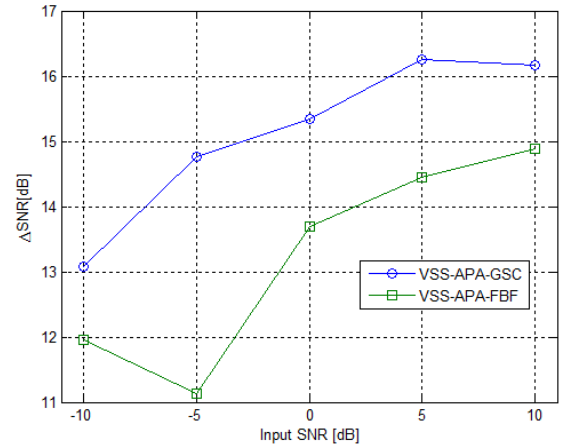


Figure 2. Performance comparison (SNR) under different input noise level and time among different algorithms. (a) white gaussian noise, (b) non-stationary noise, (c) bubble noise and (d) car engine noise

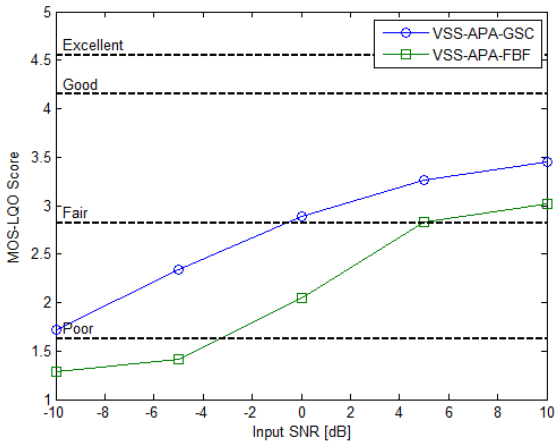


a) b) c) d)

Figure 3. Performance comparison (MOS-LQO scores) under different input noise level and time among different algorithms. (a) white gaussian noise, (b) non-stationary noise, (c) bubble noise and (d) car engine noise



a)



b)

Figure 4. Performance of VSS-APA. vs input signal used for VAD decision for a combined bubble and car engine noise. (a) SNR variation and (b) MOS-LQO score

VI. CONCLUSIONS

An efficient GSC beamformer using VSS-APA with an integrated log-energy based VAD has been proposed. It is shown that an improved performance in term of noise reduction and speech quality has been obtained, especially for non-stationary perturbations. Also our simulations have revealed that the best performance is obtained when the VAD is connected at the output of GSC beamformer.

ACKNOWLEDGMENT

This work was supported under the Grant POSDRU/107/1.5/S/76813 and Grant CNCS-UEFISCDI PN-II-ID-PCE-2011-3-0097.

REFERENCES

[1] E. Hänsler, G. Schmidt, "Acoustic Echo and Noise Control: A Practical Approach," John Wiley & Sons, New York, NY, USA, 2004.

[2] P. C. Loizou, "Speech enhancement: theory and practice," CRC Press, 2007.

[3] R. J. Compton, "Adaptive Antennas: Concepts and Applications," Englewood Cliffs, NJ: Prentice-Hall, 1988.

[4] J. Hudson, "Adaptive Array Principles," IEE Press, London 1991.

[5] S. Haykin, "Adaptive filter theory," 4th edition, Prentice Hall, Englewood Cliffs, N.J., 2001.

[6] J. Benesty, J. Chen, and Y. Huang, "Microphone Array Signal Processing," Springer-Verlag, Berlin, Germany, 2008.

[7] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," Proc. IEEE, Vol. 60, No. 8, pp. 926–935, 1972.

[8] K. Buckley, "Spacial/spectral filtering with linearly constrained minimum variance beamformers," IEEE Trans. Acoust. Speech Signal Process., Vol. 35, No. 3, pp. 249–266, 1987.

[9] B.D. Van Veen, K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," IEEE ASSP Magazine, Vol. 5, No. 2, pp. 4-24, 1988.

[10] L.J. Griffiths, C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas Propag., Vol. 30, No. 1, pp. 27–34, 1982.

[11] C. Paleologu, J. Benesty, S. Ciochina, "A Variable Step-Size Affine Projection Algorithm Designed for Acoustic Echo Cancellation," IEEE Transactions on Audio, Speech & Language Processing Vol. 16, No. 8, pp. 1466-1478, 2008.

[12] D. Comminiello, M. Scarpiniti, R. Parisi and A. Uncini, "A novel affine projection algorithm for superdirective microphone array beamforming", in Proc. of ISCAS 2010, Paris, France, pp. 2127-2130, 2010.

[13] F. Albu, D. Coltuc, D. Comminiello, M. Scarpiniti, "The variable step size regularized block exact affine projection algorithm", " in Proc. of ISETC 2012, Timisoara, Romania, 15-16 November 2012, pp. 283-286.

[14] F. Albu, C. Paleologu, J. Benesty, Y. V. Zakharov, "Variable Step Size Dichotomous Coordinate Descent Affine Projection Algorithm," Proc. IEEE EUROCON, pp. 1366-1371, St. Petersburg, Rusia, 2009.

[15] S. Van Gerven and F. Xie, "A Comparative Study of Speech Detection Methods," in Proc. EUROSPEECH, vol. 3, pp. 1095–1098, Sept. 1997.

[16] Siemens, Acceptable Delay in Digital Hearing Aids, http://hearing.siemens.com.

[17] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Feb. 2001.