# Computationally Efficient Frequency-Domain Combination of Acoustic Echo Cancellation and Robust Adaptive Beamforming

*W. Herbordt, H. Buchner, and W. Kellermann*

Telecommunications Laboratory
University of Erlangen-Nuremberg
Cauerstrasse 7, 91058 Erlangen, Germany
{herbordt, buchner, wk}@LNT.de

## Abstract

For hands-free acoustical human/machine interfaces, e. g. for automatic speech recognition or teleconferencing systems, microphone arrays using robust Generalized Sidelobe Cancellers (GSCs) in conjunction with acoustic echo cancellation (AEC) can be efficiently applied for optimum communication. This contribution devises a new structure for combining AEC and GSC. It reduces the computational complexity by more than a factor of ten relative to a time-domain arrangement, increases convergence speed, and preserves positive synergies.

## 1. Introduction

With the need for natural and comfortable communication between the user and the personalized computing devices gaining critical significance, speech-driven application control, videoconferencing, and many other multimedia services call for high-quality hands-free acoustical interface technologies that allow the user to move freely without wearing or holding any microphone device.

Even with rapidly increasing computing power, computational complexity remains an important aspect, especially when the speech processing algorithms run directly on a PC processor in the background.

For optimum quality, the signals of interest should be free from any kind of impairment, i.e. noise, reverberation, interferences, and echoes of loudspeaker signals. Compared to single-channel temporal filtering noise-reduction schemes, multi-channel space-time filtering provides better desired signal quality and more efficient suppression of local interferences [1, 2].

Acoustic echo cancellation is desirable whenever a reference of the interference is accessible. With personalized devices, these interferers may be echoes from the loudspeakers that are part of the device.

This research aims at reconciling multi-channel noise-reduction techniques and AEC while exploiting synergy effects and keeping the computational load moderate. Three basic concepts have been presented in [3], and have been applied in [4] to a combination of the robust GSC after [5] and AEC (AEC-GSC). For this combination, it has been illustrated that optimum synergies are obtained when placing the AECs directly in the sensor channels while multiplying the computational load by the number of microphones (see Fig. 1). On the one hand, since the GSC converges much faster than the AEC, it suppresses both acoustic echoes and local interferences when the AEC has not converged. On the other hand, after convergence of the AECs, more degrees of freedom of the GSC are available for the suppression of local interferences, and consequently the local interference rejection increases.

For reducing the computational complexity relative to the time-domain realization, we devise in this contribution an efficient frequency-domain AEC-GSC that uses frequency-domain adaptive filters (FDAFs) [6]. We show that the frequency-domain arrangement (a) reduces the computational complexity by more than $90\%$ relative the time-domain implementation, and (b) increases the convergence speed of the AEC, while preserving the positive synergies of the combination. Audio examples can be fond at 'http://www.LNT.de/∼herbordt/eurospeech01.html'.

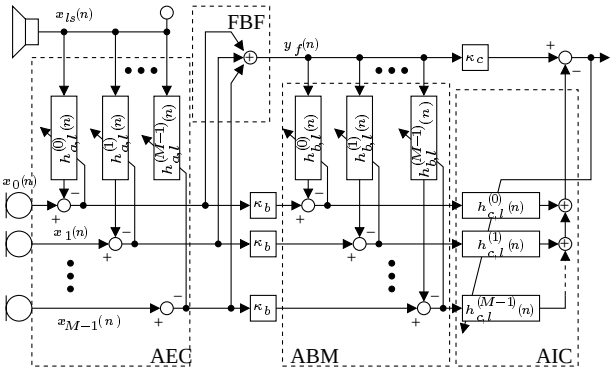


Figure 1: *Conventional robust time-domain GSC after [5] with acoustic echo cancellers (AECs) in the sensor channels*

In Section 2, we transform the time-domain AEC-GSC into the frequency-domain. Then, we compare the computational requirements of both arrangements (Section 3). In Section 4, we finally illustrate the performance with respect to interference rejection and apply the acoustical interface to a large-vocabulary continuous speech recognition system (SR).

## 2. Frequency-Domain AEC-GSC

In this section, we derive the combined system of AEC and robust GSC in the frequency domain. For fast linear convolution, we use the overlap-save (OLS) method for partitioning and reassembling the data. Although constrained and unconstrained frequency-domain adaptive filters [7, 8] are considered in later sections, we only describe the constrained algorithms here. The unconstrained realizations can be obtained by simply omitting the constraints in the filter update equations.

In the following, uppercase symbols denote frequency-domain variables, lowercase symbols stand for time-domain

This work was supported by a grant from Intel Corp., Hillsboro, OR.

variables, and the boldface font indicates a vector or matrix quantity. Superscripts $T$ and $H$ represent transpose and complex conjugate transpose, respectively. The number of microphones is denoted by $M$, the DFT length is $2L_b$. $\mathbf{F}$ is the $2L_b \times 2L_b$ discrete Fourier transform (DFT) matrix. The discrete time variable is $n$. We further use the time index $k = n/L_b$ that reflects the discrete time in numbers of blocks. A block overlap by a factor $\alpha > 1$ is introduced to improve the tracking behavior of the FDAF [9].

The conventional time-domain GSC with AECs in the sensor channels is depicted in Fig. 1. In Section 2.1, we describe the frequency-domain realization of the AECs. For maximum computational savings in conjunction with other GSC modules, we implement the fixed beamformer in the frequency domain and in the time domain (Section 2.2). Finally, we transform the adaptive sidelobe cancelling path, consisting of the Adaptive Blocking Matrix (ABM) and the Adaptive Interference Canceller (AIC), into the frequency domain and we show that they can be efficiently combined, which reduces the number of required DFTs considerably.

### 2.1. Acoustic Echo Canceller (AEC)

The AECs estimate the acoustic echoes by identifying the room impulse responses between the loudspeaker and the microphones and subtract these estimates from the sensor signals.

In Fig. 2, only one AEC signal path is depicted.

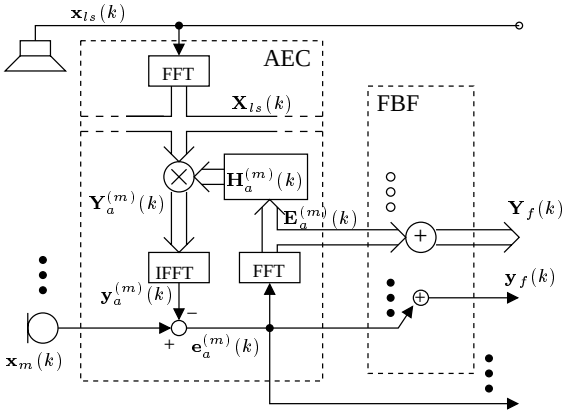


Figure 2: *AEC-GSC in the frequency-domain: Acoustic Echo Canceller (AEC) and Fixed Beamformer (FBF)*

We use a multidelay filter algorithm [7] that corresponds to the sectioned OLS method for the realization of the AECs. It allows (a) to reduce the processing delay compared to the simple OLS method, since the input signal block size $L_b$ can be chosen independently of the number of taps of the adaptive filter $L$ ($L > 1000$ is common), and (b) to choose the input signal block size $L_b$ such that optimum interference suppression and optimum target signal quality with maximum computational savings are obtained with the GSC.

The AEC adaptive filter impulse responses $h_{a,l}^{(m)}(n)$, $m = 0, 1 \ldots, M-1$, which consist of $L$ taps each, are thus segmented into $P$ small partitions of length $L_b$. That is,

$$\mathbf{h}_{a,p}^{(m)T}(k) = \left(h_{a,pL_b}^{(m)}(k), \ldots, h_{a,pL_b+L_b-1}^{(m)}(k)\right), \quad (1)$$

with $p = 0, \ldots, P-1$. The discrete time variable $n$ is replaced by the block time variable $k$. Then, these adaptive filter partitions are transformed into the frequency domain according to

$$\mathbf{H}_{a,p}^{(m)}(k) = \mathbf{F}\left(\mathbf{h}_{a,p}^{(m)}(k), \quad \mathbf{0}_{L_b\times 1}\right)^T. \quad (2)$$

This step is not required in a realization, but it illustrates the time-domain counterpart. These filter sections are independently convolved in the frequency domain with the corresponding blocks $\mathbf{X}_{ls}(k)$ of the loudspeaker signal $x_{ls}(n)$, where the matrix $\mathbf{X}_{ls}(k)$ is defined as:

$$\mathbf{X}_{ls}(k) = \mathrm{diag}\Big\{\mathbf{F}\big(x_{ls}(k\frac{L_b}{\alpha} - L_b), \ldots, x_{ls}(k\frac{L_b}{\alpha}), \ldots, x_{ls}(k\frac{L_b}{\alpha} + L_b - 1)\big)^T\Big\}, \quad (3)$$

and they are summed up in order to obtain the filter outputs $\mathbf{Y}_a^{(m)}(k)$:

$$\mathbf{Y}_a^{(m)}(k) = \sum_{p=0}^{P-1} \mathbf{X}_{ls}(k-p)\mathbf{H}_{a,p}^{(m)T}(k). \quad (4)$$

In order to provide error signals $\mathbf{E}_a^{(m)}(k)$ that are free of circular convolution effects for the adaptation algorithm, it is necessary to constrain the time-domain signals $\mathbf{e}_a^{(m)}(k)$ as follows:

$$\mathbf{e}_a^{(m)}(k) = \mathbf{x}_m(k) - \mathbf{w}\,\mathbf{F}^{-1}\mathbf{Y}_a^{(m)}(k). \quad (5)$$

The matrix $\mathbf{w} = \mathrm{diag}\big\{(\mathbf{0}_{1\times L_b}, \quad \mathbf{1}_{1\times L_b})\big\}$ is the diagonal matrix with zeroes on the upper half of the main diagonal and with ones on the lower half of the main diagonal. The vector $\mathbf{x}_m(k)$ is given by

$$\mathbf{x}_m(k) = \big(\mathbf{0}_{1\times L_b}, \; x_m(kL_b), \ldots, x_m(kL_b + L_b - 1)\big)^T. \quad (6)$$

With these definitions, the update equation for the $p$-th filter block reads:

$$\mathbf{H}_{a,p}^{(m)}(k+1) = \mathbf{H}_{a,p}^{(m)}(k) + \mathbf{G}\boldsymbol{\mu}(k)\mathbf{X}_{ls}^H(k-p)\mathbf{E}_a^{(m)}(k). \quad (7)$$

The matrix $\mathbf{G} = \mathbf{F}\mathbf{g}\mathbf{F}^{-1}$ with $\mathbf{g} = \mathrm{diag}\big\{(\mathbf{1}_{1\times L_b}, \mathbf{0}_{1\times L_b})\big\}$ constrains the gradient and ensures linear correlation.

The matrix with normalized step sizes $\boldsymbol{\mu}(k)$ is defined as

$$\boldsymbol{\mu}(k) = 2\mu\mathrm{diag}\big\{\big(P_0^{-1}, \ldots, P_{2L_b-1}^{-1}\big)\big\}, \quad (8)$$

where the estimate of the input power of the $i$-th frequency bin is given by

$$P_l(k) = \lambda P_l(k-1) + (1-\lambda)\sum_{p=0}^{P-1} |X_{ls,l}(k-p)|^2. \quad (9)$$

$X_{ls,l}(k), l = 0, \ldots, 2L_b - 1$ denotes the $l$-th frequency bin of $\mathbf{X}_{ls}(k)$.

### 2.2. Fixed Beamforming

The FBF (Fig. 2) is usually a simple delay&sum beamformer. It enhances target signal components that arrive from the array look-direction. The output is used as reference for the adaptation of the adaptive filters in the sidelobe cancelling path (see Fig. 3).

We assume that the microphone signals $x_m(n)$ are steered into the assumed target direction-of-arrival (DOA). The fractional time delays that are required in the discrete time-domain for the required time-alignment are realized by short fractional delay filters. Without loss of computational efficiency, they can thus be realized in the time domain.

The fixed beamforming is then reduced to the simple summation of the AEC output signals. Since it is computationally more efficient to provide a frequency-domain output $\mathbf{Y}_f(k) = \sum_{m=0}^{M-1}\mathbf{E}_a^{(m)}(k)$ as ABM reference and, on the other hand, a time-domain output $\mathbf{y}_f(k) = \sum_{m=0}^{M-1}\mathbf{e}_a^{(m)}(k)$ is preferable as a AIC reference, the summation is performed in both domains.

### 2.3. Adaptive Blocking Matrix (ABM)

The adaptive sidelobe cancelling path with one ABM path and one AIC path is depicted in Fig. 3. The adaptive filters are usually short [5], such that the partitioning of the filter impulse responses of ABM and AIC is not required.

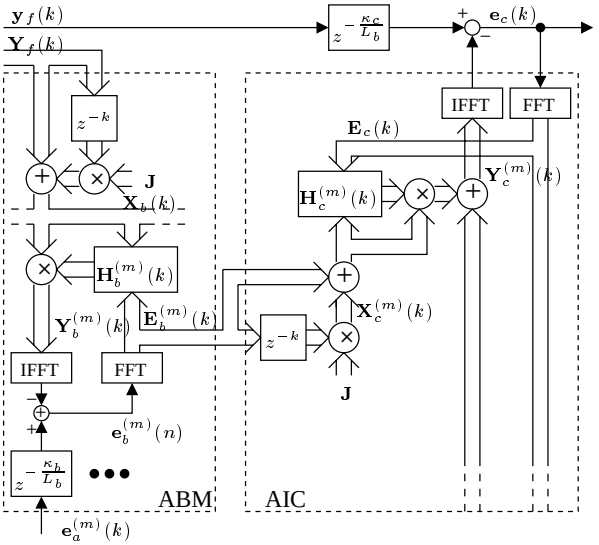


Figure 3: *AEC-GSC in the frequency-domain: Adaptive Blocking Matrix (ABM) and Adaptive Interference Canceller (AIC)*

The ABM prevents target signal cancellation that would be caused by the spatially unconstrained adaptation of the AIC: It subtracts signal components from the sidelobe cancelling path that arrive from the actually assumed target DOA. The implementation of robustness constraints and adaptation control aspects [5] are not addressed within this framework.

The ABM filter tap weights $h_{b,l}^{(m)}(n)$ are transformed into the frequency domain by

$$\mathbf{H}_b^{(m)}(k) = \mathbf{F}\left(h_{b,0}^{(m)}(k),\, \ldots,\, h_{b,L_b-1}^{(m)}(k),\, \mathbf{0}_{1\times L_b}\right)^T . \quad (10)$$

The block-based convolution using the OLS method requires the DFT of one actual block of $L_b$ FBF output samples combined with the previous block of $L_b$ output samples. Given the DFT of a block of zeroes in front of the $k$-th block of $L_b$ FBF output samples, we obtain the signal in the required format as follows [6]:

$$\mathbf{X}_b(k) = \mathrm{diag}\big\{\mathbf{Y}_f(k) + \mathbf{J}\,\mathbf{Y}_f(k-1)\big\}\,, \quad (11)$$

where $\mathbf{J}$ is defined as

$$\mathbf{J} = \mathrm{diag}\{(1,\, -1,\, 1,\, \ldots,\, -1)_{1\times 2L_b}\}\,. \quad (12)$$

Note that the operator $\mathbf{J}$, which realizes a circular shift of $L_b$ samples in the frequency domain, does not require any multiplications in a DSP realization.

Applying the constraint that ensures linear convolution, the $m$-th error signal can then be expressed according to Eq. 5 as

$$\mathbf{e}_b^{(m)}(k) = \mathbf{e}_a^{(m)}(k - \frac{\kappa_b}{L_b}) - \mathbf{w}\,\mathbf{F}^{-1}\mathbf{X}_b(k)\mathbf{H}_b^{(m)}(k)\,. \quad (13)$$

The delay $\kappa_b$ is required for causality reasons of the adaptive filters. Finally, the $m$-th update equation can be written as

$$\mathbf{H}_b^{(m)}(k+1) = \mathbf{H}_b^{(m)}(k) + \mathbf{G}\boldsymbol{\mu}(k)\mathbf{Y}_f^H(k)\mathbf{E}_b^{(m)}(k)\,, \quad (14)$$

where the matrix with normalized step sizes $\boldsymbol{\mu}(k)$ is defined as in Eq. 8. The power estimates $P_l(k)$ are determined by

$$P_l(k) = \lambda P_l(k-1) + (1-\lambda)|X_{b,l}(k)|^2\,. \quad (15)$$

$X_{b,l}(k)$, $l = 0,\, \ldots,\, 2L_b - 1$ denotes the $l$-th frequency bin of $\mathbf{X}_b(k)$.

### 2.4. Adaptive Interference Canceller (AIC)

The ABM outputs are estimates of the interferences. They are combined by the AIC adaptive filters to subtract interference components from the reference path.

According to Eq. 11, the filter inputs $\mathbf{X}_c^{(m)}(k)$ are obtained from the ABM error signals $\mathbf{E}_b^{(m)}(k)$ by

$$\mathbf{X}_c^{(m)}(k) = \mathrm{diag}\big\{\mathbf{E}_b^{(m)}(k) + \mathbf{J}\,\mathbf{E}_b^{(m)}(k-1)\big\}\,. \quad (16)$$

With the time-domain FBF output $\mathbf{y}_f(k)$, with the AIC filters $\mathbf{H}_a^{(m)}(k)$, and with the constrained output of AIC filters $\mathbf{y}_c(k)$, which reads

$$\mathbf{y}_c(k) = \mathbf{w}\,\mathbf{F}^{-1}\sum_{m=0}^{M-1}\mathbf{X}_c^{(m)}(k)\mathbf{H}_c^{(m)}(k)\,, \quad (17)$$

the AIC error signal $\mathbf{e}_c(k)$ can be expressed as

$$\mathbf{e}_c(k) = \mathbf{y}_f(k) - \mathbf{y}_c(k)\,, \quad (18)$$

The filter update equation finally is given by

$$\mathbf{H}_c^{(m)}(k+1) = \mathbf{H}_c^{(m)}(k) + \mathbf{G}\mu(k)\mathbf{X}_c^{(m)\,H}(k)\mathbf{E}_c(k)\,, \quad (19)$$

where $\boldsymbol{\mu}(k)$ is defined by Eq. 8 with

$$P_l(k) = \lambda P_l(k-1) + (1-\lambda)\sum_{m=0}^{M-1}|X_{c,l}(k)|^2\,. \quad (20)$$

$X_{c,l}(k)$ denotes the $l$-th frequency bin of $\mathbf{X}_c(k)$. Finally, the GSC output is obtained by saving the last $\frac{L_b}{\alpha}$ samples of $\mathbf{e}_c(k)$.

Thus far, we have only considered the frequency-domain AEC-GSC implementation with constraints assuring linear correlation in the update equations. In the following, we also reconsider an unconstrained version, where the constraining matrices $\mathbf{G}$ in the update equations Eq. 7,14,19 are omitted.

## 3. Computational Complexity

We examine the computational complexity of the frequency-domain AEC-GSC in comparison with the time-domain realization. The total numbers of real multiplications (NRM) per output sample as well as the memory requirements are illustrated. We assume that the DFTs are carried out by the radix-2 algorithm for real valued time-domain sequences. Then, we obtain for the NRM per output sample of the unconstrained frequency-domain GSC

$$\begin{aligned} NRM_u &= \frac{\alpha}{L_b}\big((4M+3)2L_b\log_2 2L_b + 4MP(2L_b-1) \\ &\quad +6M(3L_b-1) + 10(L_b+1)\big)\,. \end{aligned} \quad (21)$$

The constrained frequency-domain GSC requires 2 additional DFTs per adaptation unit, or

$$NRM_c = NRM_u + 4\alpha M(P+2)\log_2 2L_b\,. \quad (22)$$

With increasing block size $L_b$, the number of partitions $P$ decreases, which leads to maximum computational savings, and to larger processing delay. For comparison, the time-domain GSC using NLMS algorithms requires

$$NRM_t = 2MPL_b + M(4L_b+1) + 5 \quad (23)$$

real multiplications per output sample. The complexity ratios,

$$CR_x = NRM_x/NRM_t, \quad \text{with } x \in \{c,\, u\} \quad (24)$$

are depicted for $M = 8$, $L = 3072$, $\alpha = 1$ for variable block lengths $L_b$ in Fig. 4 (a).

The memory requirements of the unconstrained and the constrained frequency-domain GSC are assumed to be identical. Assuming IEEE double precision format, we find

$$\begin{aligned} S_t &= 8L_b(MP + 2P + 5.5M + 5)\ \text{Bytes} \\ S_f &= 8L_b(2MP + 2P + 8.5M + 11)\ \text{Bytes} \end{aligned}$$

for the memory that is required for the time-domain GSC and for the frequency-domain GSC, respectively. Fig. 4 (b) illustrates these values for variable block lengths $L_b$, $M = 8,\ L = 3072,\ \alpha = 1$.

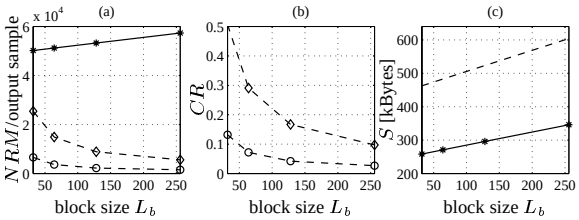


Figure 4: *(a) Number of real multiplications $NRM_t$ (∗), $NRM_u$ (∘), and $NRM_c$ (⋄) per output sample ; (b) Complexity ratios $CR_u$ (∘) and $CR_c$ (⋄); (c) Required memory of time-domain GSC $S_t$ (solid), and of frequency-domain GSC $S_f$ (dashed) ; (d) ($M = 8,\ L = 3072,\ \alpha = 1$)*

We see that for a block size $L = 128$, $97\%$ computational complexity can be saved at the cost of doubling the memory when using the unconstrained frequency-domain GSC.

## 4. Experimental Evaluation

We first examine interference rejection, expressed by the interference-energy ratio between the microphones and the beamformer output. Second, we apply the system to a large-vocabulary continuous speech recogntion system (dictation system "Dragon Naturally Speaking Preferred 3.52").

A linear microphone array with $M = 8$ equally spaced ($d = 4$ cm) sensors has been placed in both a chamber with low reverberation ($T_{60} = 50$ ms) and in an office room ($T_{60} = 300$ ms). The male target (loud-)speaker and the female interferer (loud-)speaker are located broadside at a distance of $60$ cm and $30$ degrees off-axis at a distance of $1.3$ m, respectively. The (male) acoustic echoes arrive from a loudspeaker placed along the array axis at a distance of $60$ cm from the array center. The frequency band is $360-5000$ Hz at a sampling rate $T_s = 11025$ Hz. The signal-to-interference ratio and the signal-to-acoustic-echo ratio at the speaker positions are $0$ dB. The GSC parameters were chosen for maximum IR. For the frequency-domain arrangement $\mu = 0.25$, $\lambda = 0.9$ was used, and for the time-domain scheme we used as step size for NLMS adaptation: $\mu = 0.7$, and as number of ABM/AIC filter taps: $N = 32$.

The AECs were adapted until an Echo-Return-Loss Enhancement (ERLE), [1] of $ERLE_a = 25$ dB is reached. The ABM filters are frozen after adaptation for $50000$ samples with only the target present. The AIC is then adapted for $50000$ samples. The average IR is calculated over the last $20000$ samples. In Table 1, we see that the average IR, ERLE are almost identical for all three AEC-GSC realizations. The AIC tracks the time-variance of the interferer power spectral density (PSD) and thus cancels the non-stationary interference efficiently. Since the block sizes of the frequency-domain arrangement are kept small relative to the short-time stationarity of speech signals ($L_b = 64$, corresponding to $5.8$ ms with $T_s = 11025$ Hz), the reduced tracking of block adaptive algorithms does not influence the average IR. Due to faster convergence, the average IR of the unconstrained frequency-domain GSC is slightly higher than the average IR of the constrained frequency-domain GSC.

For the evaluation with the SR, the SR is first trained with the proposed system in the actual receiving room. Texts for evaluation and for training are non-overlapping, however, the vocabulary of the dictated text for the evaluation is in known context. This is meaningful for large vocabulary dictation systems and real-life situations. The word recognition accuracy that is obtained with the "Dragon System" headset is $100\%$ for the dictated text. The results for the hands-free situation are given in Table 2.

[1] ERLE expresses the energy ratio of the acoustic echo within the microphone signal relative to the AEC output

| $T_{60} = 50/300$ ms | $IR$ | $ERLE_{tot}$ | $ERLE_{gsc}$ |
|---|---|---|---|
| time-domain | 25.6/14.3 | 36.3/31.3 | 11.3/6.3 |
| unconstrained | 25.4/14.1 | 36.1/31.0 | 11.1/6.0 |
| constrained | 24.7/12.9 | 35.5/30.8 | 10.5/5.8 |

Table 1: *Interference rejection (IR) and ERLE in dB*

| Environment | Single Mic. | FBF | GSC | AEC-GSC |
|---|---|---|---|---|
| Chamber w. low rev. | 32% | 60% | 92% | 97% |
| Office room | 30% | 50% | 86% | 91% |

Table 2: *Word recognition accuracies*

## 5. Conclusions

In this paper, we have devised a computationally efficient frequency-domain combination of AEC and robust GSC exploiting optimum positive synergies, which converges faster than a conventional time-domain scheme. The arrangement has been applied to a continuous speech recognition and interference rejection characteristics have been illustrated in both anechoic and echoic environments.

## 6. Acknowledgment

The authors wish to thank David Graumann, Intel Corp., Hillsboro, OR for his encouragement and stimulating discussions.

## 7. References

[1] M.S. Brandstein and D.B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*, Springer Verlag, to appear in2001.

[2] B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, pp. 4–24, April 1988.

[3] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," *IEEE Proc. ICASSP*, vol. 1, pp. 219–222, April 1997.

[4] W. Herbordt and W. Kellermann, "Acoustic echo cancellation embedded into the generalized sidelobe canceller," *Proc. EUSIPCO*, vol. 3, September 2000.

[5] O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. on SP*, vol. 47, no. 10, October 1999.

[6] J.J. Shynk, "Frequency-domain and multirate adaptive filtering," *IEEE SP Magazine*, pp. 14–37, January 1992.

[7] J. Soo and K.K. Pang, "Multidelay frequency domain adaptive filter," *IEEE Trans. on ASSP*, vol. 38, no. 2, pp. 373–376, 1990.

[8] D. Mansour and A.H. Gray, "Unconstrained frequency-domain adaptive filter," *IEEE Trans. on ASSP*, vol. ASSP-30, no. 5, pp. 726–734, 1982.

[9] E. Moulines, O.A. Amrane, and Y. Grenier, "The generalized multidelay adaptive filter: Structure and convergence analysis," *IEEE Trans. on SP*, vol. 43, pp. 14–28, 1995.