Efficient Multi-Channel Acoustic Echo Cancellation Using Constrained Sparse Filter Updates in the Subband Domain

Naveen Kumar Desiraju⋆*, Simon Doclo*†*, Timo Gerkmann*†*, Tobias Wolff*⋆

⋆Acoustic Speech Enhancement Research, Nuance Communications Deutschland GmbH, 89077 Ulm, Germany
†Dept. of Medical Physics and Acoustics - Signal Processing, 26111 Oldenburg, Germany
Email: naveen.desiraju@nuance.com
Web: ⋆www.nuance.com; †www.sigproc.uni-oldenburg.de

Abstract

In this paper we present constrained sparse tap-selection schemes for updating Multi-Channel Acoustic Echo Cancellation (MAEC) filters in the subband domain. At first, M-Max tap-selection on the complete multi-channel reference spectra is performed and the effect of the subsequent sparse filter update on the speed of convergence of the MAEC filters is investigated. We consider a measure to quantify signal sparsity, and use it to investigate the spectral and inter-channel sparsity present in real-world surround sound signals. A heuristic tap-selection scheme is proposed which exploits the signal sparsity properties, and performs tap-selection with much lower computational effort as compared to the M-Max and the full-update tap-selection schemes, while giving similar echo cancellation performance.

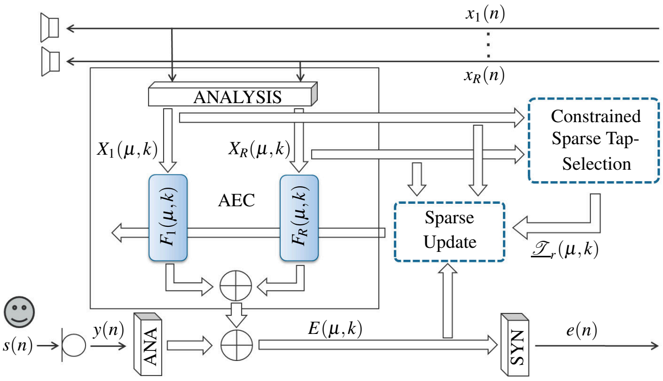

Figure 1: Block diagram of subband MAEC setup.

1 Introduction

Acoustic Echo Cancellation (AEC) is employed in many speech communication systems to reduce the undesired echoes that result from coupling between the loudspeakers and the microphones. AEC is a key technology in teleconferencing, hands-free communication as well as distant-talk voice-control systems (such as home entertainment systems). In scenarios with large reverberation times (T_{60}), very long AEC filters (several thousand taps) may be required to achieve effective echo cancellation, resulting in large computational effort for both the filter update and the filtering operation. This problem gets exacerbated when multi-channel reference signals are involved, as may be the case with home-entertainment systems with surround sound. Partial update adaptive filtering algorithms provide a potential solution as they reduce the required computational effort for the filter update but may result in slower convergence [1–3].

The M-Max NLMS [1–3] is a well-known algorithm that has been proposed for use with time-domain AEC filters that updates a constrained number of filter taps at every iteration. It uses the so-called M-Max criterion for selecting the taps corresponding to the M largest magnitude tap-inputs for updating the adaptive filter, and has been used to tackle the non-uniqueness problem for stereo AEC [2–4].

In this paper we consider subband adaptive filters and present methods which update the MAEC filters in a sparse manner, with the total number of filter taps updated in every frame being constrained. At first, we consider the M-max criterion for tap-selection on the set of all MAEC filter coefficients based on the magnitudes of the multi-channel subband reference signal (in all subbands and channels) and refer to this method as the *Full M-Max* tap-selection scheme. This tap-selection scheme maximizes the energy of the filter update at every iteration by exploiting the spectral and inter-channel sparsity of the reference signals, but requires large computational effort. To tackle the problem of computational effort, we propose a more efficient scheme for performing constrained sparse tap-selection in the MAEC filters. From here on we shall refer to inter-channel sparsity as spatial sparsity, as in an MAEC setup the channels are displayed from different spatial positions (positions of the loudspeakers).

In Section 3, we present the Full M-Max tap-selection scheme, followed by the proposed scheme in Section 4. We consider a measure to quantify signal sparsity in Section 5. In Section 6, the simulations and results for the comparison between the Full M-Max and the proposed scheme for different scenarios are presented. Some conclusions and remarks are presented in Section 7.

2 Signal Model

Figure 1 shows the block diagram of a subband MAEC setup, where $x_r(n)$ denotes the time domain reference signal in channel r. $X_r(\mu,k)$ represents the subband reference signal in the μ-th subband and k-th frame, obtained by applying the Short-time Fourier Transform (STFT) on $x_r(n)$. Here, $\mu \in \{0,\dots,N-1\}$ and $r \in \{0,\dots,R-1\}$, where N denotes the number of subbands: $N = \frac{N_{\mathrm{FFT}}}{2}+1$, with N_{FFT} being the DFT order. Let

$$\underline{X}_r(\mu,k) = [\ X_r(\mu,k), \quad \dots, \quad X_r(\mu,k-L+1)\ ]^H \quad (1)$$

represent the vector containing the values of $X_r(\mu,k)$ for L frames, where L is the filter length. By stacking $\underline{X}_r(\mu,k)$ together for all subbands and channels, we obtain an $(N \cdot R \cdot L \times 1)$-dimensional vector referred to here as the *multi-*

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) project DREAMS under grant agreement ITN-GA-2012-316969.

channel reference signal buffer:

$$\underline{\chi}(k) = [\,\underline{X}_0^H(0,k),\, \underline{X}_1^H(0,k),\, \ldots,\, \underline{X}_{R-1}^H(0,k),$$
$$\underline{X}_0^H(1,k),\, \underline{X}_1^H(1,k),\, \ldots,\, \underline{X}_{R-1}^H(1,k),$$
$$\ldots,\, \underline{X}_0^H(N-1,k),\, \ldots,\, \underline{X}_{R-1}^H(N-1,k)\,]^H. \quad (2)$$

The vector $\underline{X}_r(\mu,k)$ is used to update the L-tap long subband AEC filter $\underline{F}_r(\mu,k)$ using the NLMS algorithm [5] in every frame:

$$\underline{F}_r(\mu,k+1) = \underline{F}_r(\mu,k) + \frac{\lambda \cdot E^*(\mu,k)}{\sum_{r=0}^{R-1} \underline{X}_r^H(\mu,k)\,\underline{X}_r(\mu,k)} \cdot \{\, \underline{\mathscr{T}}_r(\mu,k) \odot \underline{X}_r(\mu,k)\,\}. \quad (3)$$

Here λ is the step-size, $E(\mu,k)$ is the error signal after echo-cancellation, $\underline{\mathscr{T}}_r(\mu,k)$ (with entries $\{0,1\}$) is the $(L\times 1)$-dimensional tap-selection vector, $*$ denotes the complex conjugate and $\odot$ denotes element-wise multiplication. To represent the fraction of the total $N\cdot R\cdot L$ taps in the MAEC filters that are updated in every frame, we use:

$$Q = \frac{M}{N\cdot R\cdot L}, \quad (4)$$

where M represents the number of taps updated in every frame. In the following sections, we consider different schemes for constructing $\underline{\mathscr{T}}_r(\mu,k)$ and investigate their effect on AEC performance.

3 Full M-Max Tap-Selection

The Full M-Max scheme performs tap-selection in the MAEC filters by selecting those M tap-inputs from $\underline{\chi}(k)$ (from (2)) which have the largest magnitudes. The selected spectro-spatial regions in the MAEC filters correspond to the largest energy concentration in the multi-channel reference signal. Updating the MAEC filters using these selected taps will ensure that the energy of the sparse filter update vector in (3) is closest to the energy of the full filter update vector, with the expectation being that this would result in the smallest difference in AEC performance as compared to full filter update. Thus, the Full M-Max tap-selection may be taken as a reference for comparing the AEC performance of sparse tap-selection schemes as, for a given M, it maximizes the *closeness* of the sparse filter update to the full filter update. However, performing this selection on $\underline{\chi}(k)$ results in large computational complexity because of the large sorting effort involved, as $N\cdot R\cdot L$ elements need to be sorted in every frame.

4 Proposed Tap-Selection Scheme

We propose a heuristic method which exploits signal sparsity properties for performing constrained sparse tap-selection and tackles the problem of computational complexity by avoiding the large sorting effort required for Full M-Max tap-selection as described in Section 3. For every frame, the tap-selection vector $\underline{\mathscr{T}}_r(\mu,k)$ for updating the filter $\underline{F}_r(\mu,k)$ in (3) is obtained by performing the M-Max operation on $\underline{X}_r(\mu,k)$ (from (1)). Let $\underline{\psi}_r(\mu,k)$ represent the $(L\times 1)$-dimensional vector containing the magnitudes of the elements in $\underline{X}_r(\mu,k)$, then

$$\phi_r(\mu,k) = \underline{\psi}_r^T(\mu,k)\,\underline{1}_{(L\times 1)} \quad (5)$$

is the sum of the magnitudes of the elements in $\underline{X}_r(\mu,k)$. To find the important regions of the multi-channel reference spectra, we refer $\phi_r(\mu,k)$ to its mean across all subbands and channels to compute:

$$H_r(\mu,k) = \min\left(\frac{\phi_r(\mu,k)}{\sum_{\mu=0}^{N-1}\sum_{r=0}^{R-1}\phi_r(\mu,k)} \cdot N\cdot R, 1 \right), \quad (6)$$

with its sum across all subbands and channels denoted by:

$$h(k) = \sum_{\mu=0}^{N-1}\sum_{r=0}^{R-1} H_r(\mu,k). \quad (7)$$

In (6), the ratio is limited to 1 as $H_r(\mu,k)$ is used to compute the number of taps to be selected when updating the filter $\underline{F}_r(\mu,k)$, according to:

$$\mathscr{L}_r(\mu,k) = \lfloor \mathscr{F}\{H_r(\mu,k),\gamma(k)\}\cdot L \rfloor, \quad (8)$$

where the $\lfloor\cdot\rfloor$ operator denotes rounding downwards, while the function $\mathscr{F}$ is defined as:

$$\mathscr{F}\{H_r(\mu,k),\gamma(k)\} = \begin{cases} \gamma(k) + (1-\gamma(k))\cdot H_r(\mu,k), & \text{if } h(k) < Q\cdot N\cdot R \\ \gamma(k)\cdot H_r(\mu,k), & \text{else.} \end{cases} \quad (9)$$

Here, $\gamma(k)$ is used to satisfy the constraint:

$$\sum_{\mu=0}^{N-1}\sum_{r=0}^{R-1} \mathscr{F}\{H_r(\mu,k),\gamma(k)\} \overset{!}{=} Q\cdot N\cdot R, \quad (10)$$

which is imposed to restrict the total number of taps updated in the MAEC filters in every frame. Plugging (9) into (10) and solving for $\gamma(k)$ yields:

$$\gamma(k) = \begin{cases} \frac{Q\cdot N\cdot R - h(k)}{N\cdot R - h(k)}, & \text{if } h(k) < Q\cdot N\cdot R \\ \frac{Q\cdot N\cdot R}{h(k)}, & \text{else.} \end{cases} \quad (11)$$

The function $\mathscr{F}$ in (9) is designed such that when $h(k)$ (from (7)) is less than the constraint, we allocate the majority of effort to the important regions of the multi-channel reference spectra (through $H_r(\mu,k)$), with the leftover effort being distributed equally across all spectro-spatial regions (through $\gamma(k)$). In the event that $h(k)$ exceeds the constraint, $H_r(\mu,k)$ is simply scaled down using $\gamma(k)$.

Finally, the number of taps used to perform M-Max tap-selection is given by $\mathscr{L}_r(\mu,k)$ (from (8)). The computational effort is lower as compared to Full M-Max tap-selection, as only a vector with L elements needs to be sorted in each subband and channel, which is implemented efficiently using the SORTLINE algorithm [6].

5 Closeness Measure

To analyze the amount of spectral and spatial sparsity in the reference signals we have used the so-called Closeness Measure [2, 3]:

$$\xi(\underline{\alpha},\underline{\beta},k) = \frac{\underline{\chi}^H(k)\ \mathrm{diag}\{\underline{\alpha}(k)\}\ \underline{\chi}(k)}{\underline{\chi}^H(k)\ \mathrm{diag}\{\underline{\beta}(k)\}\ \underline{\chi}(k)}, \quad (12)$$

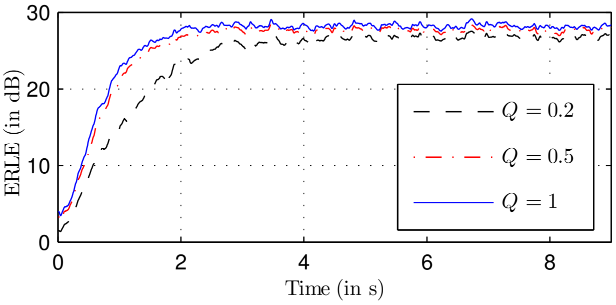

Figure 2: ERLE Convergence curves for single-channel White Gaussian Noise (WGN) signal for different values of Q (4), with Full M-Max tap-selection.

where $\underline{\alpha}(k)$ and $\underline{\beta}(k)$ are tap-selection vectors, stacked similarly as in (2), and $\mathrm{diag}\{\underline{z}\}$ generates a diagonal matrix with $\underline{z}$ as the main diagonal. When using $\underline{\beta}(k) = \underline{1}_{(N\cdot R\cdot L\times 1)}$, representing full tap-selection, the Closeness Measure indicates the efficiency of tap-selection $\underline{\alpha}(k)$ to maximize the energy of the update as compared to full tap-selection. Here, $\underline{1}_{(N\cdot R\cdot L\times 1)}$ is a column vector with all $N\cdot R\cdot L$ elements equal to 1.

6 Simulations and Results

In this section we present the experimental setup, procedures and results of our simulations. For our experimental setup we consider time domain reference signals at a sampling frequency of 16 kHz, which are recorded in a room with $T_{60} \approx 200$ ms. A single microphone has been considered, which captures the acoustic echo at an Echo to Noise Ratio of 30 dB. The reference signals are analyzed using a Hanning window with $N_{\mathrm{FFT}} = 512$ and 75% overlap. For the subband domain MAEC filters, a length of $L = 22$ taps has been chosen which corresponds to $N_{\mathrm{FFT}} \cdot (1 + 0.25\cdot(L-1))$ samples, or 200 ms. We measure the AEC performance by evaluating the Echo Return Loss Enhancement (ERLE) [7] and by observing its speed of convergence. The speed of convergence of the ERLE is assessed using T_{20}, which is the time required to first reach an ERLE of 20 dB.

6.1 White Gaussian Noise (WGN) signal

This section presents the AEC performance when a single-channel WGN is used as reference signal. Figure 2 shows the ERLE convergence behaviour for different values of Q, with Full M-Max scheme used to perform tap-selection. We see that for $Q = 0.5$, meaning that only 50% of the total taps are updated in every frame, the ERLE convergence behaviour is very similar to that for $Q = 1$ (full tap-selection). Even for $Q = 0.2$, the deterioration in ERLE at convergence is relatively small (~1–2 dB). Figure 3 shows the T_{20} and $\xi(\underline{\alpha}_{\text{M-Max}}, \underline{1}, k)$ values for different values of Q. We see that T_{20} for $Q = 0.5$ is almost identical to that for $Q = 1$, which corresponds with the results shown in [2, 3] for time domain AEC. Hence, our investigation shows that AEC performance obtained using the M-Max criterion in the time domain [2, 3] translates similarly into the subband domain. From both figures we can say that the AEC performance for $Q = 0.5$ is very similar to $Q = 1$. As $\xi = 0.85$ for $Q = 0.5$ from Figure 3(b), we can consider it as a sufficient condition to achieve good AEC performance.

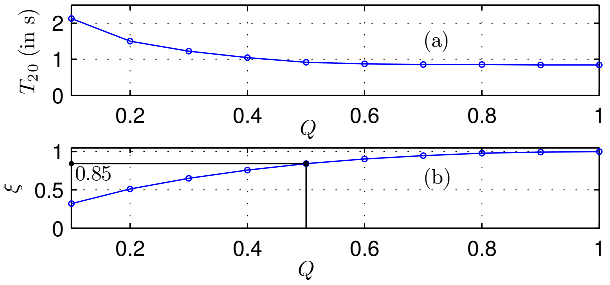

Figure 3: (a) Comparison of T_{20} values, and (b) Closeness measure ξ (from 12) for different values of Q for single-channel WGN signal, with Full M-Max tap-selection.

6.2 Real-World Signals

For our simulations, we have considered surround sound signals (action movie scenes and pop concert recordings) in Dolby Digital 5.1 format and mono speech signals (audiobook) as reference. The pop concert recordings are active in all frames, with significant surround content, and can hence be considered as the worst case scenario for the scope of our investigation. Figure 4 plots the percentage of frames with $\xi(\underline{\alpha}_{\text{M-Max}}, \underline{1}, k) > 0.85$ for the different real-world signals, and suggests that at least 99% of frames for all signals achieve $\xi > 0.85$ at a value of Q as low as 0.2. This figure clearly highlights the potential that exists for exploiting spectral and spatial sparsity in real-world signals for the purpose of AEC.

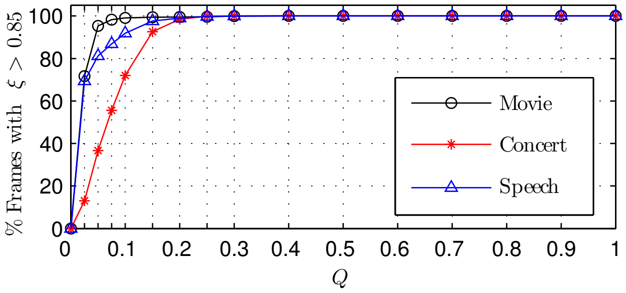

Figure 4: Percentage of frames in real-world signals for which $\xi > 0.85$ vs. Q. Action movie and pop concert are 5 channel surround signals, while the speech signal is mono.

As the Full M-Max scheme maximizes $\xi(\underline{\alpha}_{\text{M-Max}}, \underline{1}, k)$ in every frame, we propose to analyze the efficiency of the proposed tap selection $\underline{\alpha}_{\text{Prop}}$ to the Full M-Max by computing the average closeness measure

$$\delta = \frac{1}{K}\sum_{k=1}^{K} \xi(\underline{\alpha}_{\text{Prop}}, \underline{\alpha}_{\text{M-Max}}, k). \qquad (13)$$

Figure 5 shows the values of δ for the different real-world signals for different values of Q. From this we can conclude that for $Q \geq 0.5$, the proposed scheme approximates the Full M-Max scheme in terms of efficiency of tap-selection for real-world signals.

Figure 6 shows the ERLE curves for an action movie signal, obtained by updating the MAEC filters using the full tap-selection scheme ($Q = 1$) as well as using the two sparse tap-selection schemes for $Q = 0.2$. We observe that

Operation	Full Selection + NLMS	Full M-Max + NLMS	Proposed + NLMS
# Additions	$4NRL+6NR$	$4QNRL+6NR$	$4QNRL+NRL+8NR+1$
# Comparisons	0	$NRL\log_2(NRL)$	$2NR\log_2(L)+3NR+1$
# Multiplications	$4NRL+4NR+2N$	$4QNRL+4NR+2N$	$4QNRL+7NR+2N$
# Divisions	N	N	$N+2$

Table 1: Computational effort for different tap-selection schemes and filter update using NLMS algorithm.

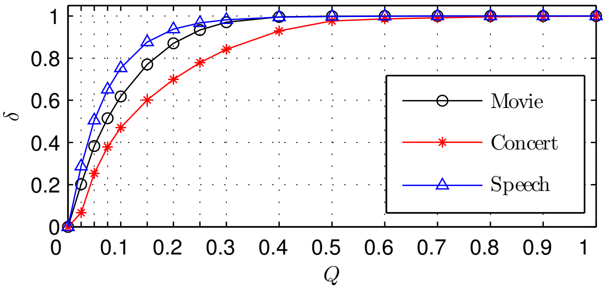

Figure 5: Efficiency of proposed tap-selection to Full M-Max tap-selection vs. Q for different real-world signals.

by just updating only 20% of the total taps in the MAEC filters, the AEC performance for the sparse schemes is almost identical to the performance for the full tap-selection scheme (∼1–2 dB deterioration). Hence, we see that using a sparse tap-selection scheme, for a certain range of Q, results in almost identical AEC performance as compared to the full tap-selection scheme for real-world surround signals.

6.3 Computational Effort

Table 1 shows the computational effort for the different tap-selection schemes as a function of Q, N, L and R. The number of additions, multiplications, divisions and comparison operations for tap-selection and filter update are shown separately. If we assume that comparison, multiplication and division operations are 1, 5 and 50 times as complex as additions respectively, then for $\{N,L,R\} = \{257,22,5\}$, the Full M-Max and the proposed scheme require less computational effort than the full tap-selection scheme for $Q < 0.384$ and $Q < 0.903$ respectively. For $Q = 0.2$, the Full M-Max scheme saves ∼17% and the proposed scheme saves ∼65% respectively in computational effort as compared to the full tap-selection scheme.

7 Conclusions

We presented sparse update techniques in the subband domain (Full M-Max and proposed) which constrain the number of taps updated in the MAEC filters. Spectral and spatial sparsity present in real-world surround sound signals was exploited for updating the MAEC filters. For real-world surround sound signals, the proposed scheme was found to approach the Full M-Max scheme in terms of efficiency of tap-selection when the number of taps updated exceeded 50% of the total taps in the MAEC filters. The AEC performance for the sparse schemes, when only 20% of the filter taps were updated, was found to be almost identical to the performance when all the taps were updated. The proposed scheme achieves comparable results with lower computational effort as compared to the Full M-Max and the full tap-selection schemes.

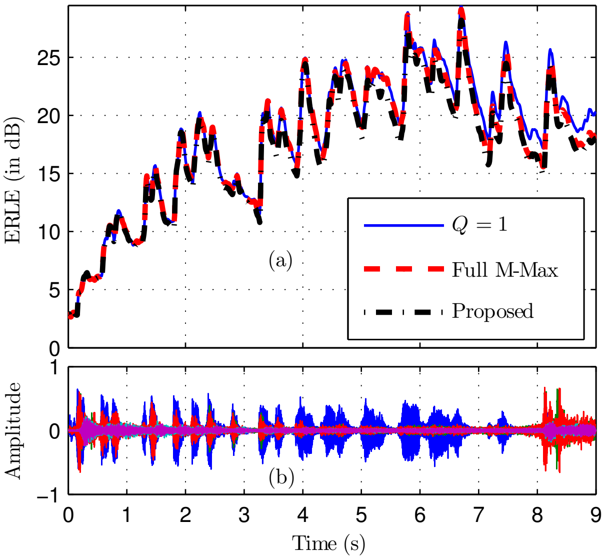

Figure 6: (a) ERLE curves for a 5-channel action movie signal with full tap-selection ($Q = 1$) and sparse tap-selection (Full M-Max and proposed) for $Q = 0.2$, (b) Waveform of 5-channel action movie signal, with different channels distinguished by colour.

References

[1] K. Doğançay, P.A. Naylor, “Recent advances in partial update and sparse adaptive filters”, in *Proc. 13th European Signal Processing Conference*, Antalya, Turkey, Sep. 2005.

[2] A.W.H. Khong, P.A. Naylor, “A family of selective-tap algorithms for stereo acoustic echo cancellation”, in *Proc. IEEE International Conference Acoustics, Speech, and Signal Processing*, vol. 3, pp. iii/133–iii/136, Mar. 2005.

[3] A.W.H. Khong, P.A. Naylor, “Stereophonic acoustic echo cancellation employing selective-tap adaptive algorithms”, *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 785–796, May 2006.

[4] A.W.H. Khong, P.A. Naylor, “Reducing inter-channel coherence in stereophonic acoustic echo cancellation using partial update adaptive filters”, in *Proc. of European Signal Processing Conference*, 2004, pp. 405–408.

[5] S.O. Haykin, *Adaptive filter theory, 4th edition*. Upper Saddle River, NJ: Prentice Hall, 2002.

[6] I. Pitas, “Fast algorithms for running ordering and max/min calculation”, *IEEE Trans. Circuits and Systems*, vol. 36, no. 6, pp. 795–804, Jun. 1989.

[7] E. Hänsler, G. Schmidt, *Acoustic echo and noise control - a practical approach*. Hoboken, NJ: Wiley and Sons, 2004.