On the Performance of the Partitioned Versus Non-Partitioned Stereo Frequency Domain Adaptive Kalman Filter in In-Car Communication Systems

Inka Meyer zum Alten Borgloh, Jan Franzen, Tim Fingscheidt
Institute for Communications Technology, Technische Universität Braunschweig, 38106 Braunschweig, Germany
Email: {i.meyer-zum-alten-borgloh, j.franzen, t.fingscheidt}@tu-bs.de

Abstract

In this paper, we investigate a stereo partitioned-block approach to frequency domain adaptive Kalman filtering for usage in feedback control of an in-car communication (ICC) system. To accomplish this, we will compare the stereo partitioned-block processing to stereo non-partitioned processing in the same acoustic conditions and for a variety of scenarios which are relevant in the context of ICC systems. The results presented in this paper show that despite the possibility to improve computational complexity and overall feedback suppression, the partitioned-block approach leads to higher robustness to noise.

Introduction

The loud and noisy environment in a car is a challenging situation for conversations. Especially the passengers sitting in the rear seats of the car often struggle to understand the ones sitting in the front seats. To support the communication, an in-car communication (ICC) system records the front passenger's speech with the built-in or an additional microphone, amplifies it and plays it back through the rear loudspeakers. In this configuration, acoustic coupling between the loudspeakers' outputs and the microphone can lead to an acoustic feedback. In order to avoid howling or echoes, an acoustic echo cancellation (AEC) algorithm can be used for feedback control, to estimate the remaining amount of the loudspeaker signals, that gets acquired by the microphone again.
While in general also manual ways of feedback control are possible through microphone and loudspeaker arrangement [1], these are not further discussed here.
An overview over methods for feedback control in ICC systems is given in [2]. One possibility for feedback suppression is automatic gain reduction, so the gain is cut down over the whole frequency range or in smaller critical subbands/around critical frequencies [1]. Adaptive algorithms that can be used in feedback suppression like NLMS or RLS are described in [3] with further investigations like stepsize control mechanisms as in [4].
In hands-free systems, the usage of Kalman filtering has shown to be effective for AEC. In [5], the frequency domain adaptive Kalman filter (FDAKF) is evaluated as a stereo AEC (SAEC) for wideband signals in an automotive context. Extensive investigations on the performance of the SAEC in an ICC setup have been performed in [6].
Since the filter adaptation takes place in the frequency domain, time domain signals need to be transformed via FFT. Since the complexity of the FFT does not scale linearly with its size, it can be useful to compute multiple FFTs of a smaller size instead of one large FFT. This is taken up in [7]: *Partitioned-block* processing is presented as an approach to increase convergence speed and reduce computational complexity.
In this work, we investigate the partitioned-block processing as presented in [7] in the context of a stereo ICC system as shown in [6]. We evaluate the partitioned in comparison to the unpartitioned algorithm with regard to overall feedback suppression and computational complexity under a variety of relevant acoustic scenarios for ICC systems.
The remainder of this paper is structured as follows: In the next section the AEC algorithm is presented. Afterwards, setup and methods for the evaluation in the ICC context are described, followed by the results of our simulations. Finally, conclusions are drawn.

Algorithm

The investigated algorithm is based on the SAEC presented in [5, 8], which is based on the variationally diagonalized FDAKF as shown in [9, 10]. The performance of this SAEC in an ICC setup has been investigated in [6]. To convert the ICC SAEC into partitioned processing, the algorithms presented by Kuech et al. in [7] are used.
In the underlying system setup, four simulated loudspeakers in the car are fed with the signals $x_c(n)$, $c \in \mathscr{C} = \{FL, FR, RL, RR\}$ with discrete sample index n, e.g., from FM radio or the general infotainment system. The microphone signals contain the front passenger's speech signal and an echo of the loudspeaker signals due to the acoustic coupling by the loudspeaker-enclosure-microphone (LEM). The LEM paths are simulated through four different impulse responses that are applied to the loudspeaker signals. Additionally, the presence of car noise at the microphone is considered in the simulations.
In contrast to single-channel ICC systems, the *stereo* AEC estimates two system transfer functions modeling the room impulse responses. These represent two of the true impulse responses each. Therefore each two of the loudspeaker signals are combined as reference signals $x_j(n), j \in \mathscr{J} = \{1, 2\}$, for the SAEC, where j denotes the channel index. In general one is free to combine either the two rear and front or alternatively the two right and left signals. In the following we use the first option, so the rear signals are combined to one reference signal $x_R(n) = x_{RL}(n) + x_{RR}(n)$ as well as the front

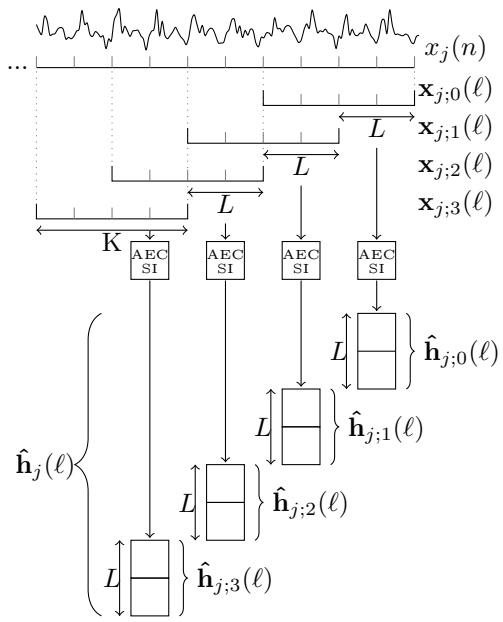

Figure 1: Partitioned usage of one reference signal $x_j(n)$ in a specific frame ℓ with $B = 4$. In principle, each AEC system identification (AEC SI) step estimates a short impulse response $\hat{\mathbf{h}}_{j;b}(\ell)$.

signals to provide $x_F(n) = x_{FL}(n) + x_{FR}(n)$. Those reference signals include media as well as the echo-free and re-amplified enhanced microphone signal that is added to the rear loudspeaker signals.

The system functions estimated by the SAEC can be divided into a number of B partitions, each of length L. Moving along with this partitioned treatment, also the framing of the reference signals needs to be fitted to the partitioning. Contrary to the unpartitioned SAEC, where one K-point FFT is performed with one large consecutive set of samples, we need a number of B overlapping smaller sets from the latest frame ℓ of the reference signals. Since the sets of samples for the partitioned calculations can be chosen smaller, also a smaller FFT size K can be used.

For each partition $b \in \mathscr{B} = \{0, ..., B-1\}$ this set is shifted backwards in time by $b \cdot L$ samples, as visualized in Figure 1, whereas L denotes the partition size and thereby defines the number of coefficients contributed by each partition. This leads to the following expression:

$$\begin{aligned}\mathbf{x}_{j;b}(\ell) = \big[&x_j\big((\ell-1)\cdot R - b\cdot L + R - K\big), \ldots,\\ &x_j\big((\ell-1)\cdot R - b\cdot L + R - 1\big)\big]^T,\\ &j \in \mathscr{J}, b \in \mathscr{B}\end{aligned} \tag{1}$$

with frame index $\ell \in \{1, 2, \ldots\}$ and frame shift R. The b-th partition of the reference signals can be transformed to the frequency domain by multiplication with the K-point DFT matrix $\mathbf{F}_{K\times K}$:

$$\mathbf{X}_{j;b}(\ell) = \mathrm{diag}\{\mathbf{F}_{\mathrm{K\times K}} \cdot \mathbf{x}_{\mathrm{j;b}}(\ell)\}. \tag{2}$$

Furthermore we use the latest frame of the microphone signal given by

$$\begin{aligned}\mathbf{y}(\ell) = \big[&\mathbf{0}_{K-R}^T,\ y\big((\ell-1)\cdot R\big), \ldots,\\ &y\big((\ell-1)\cdot R + R - 1\big)\big]^T,\end{aligned} \tag{3}$$

also transferred to the frequency domain as $\mathbf{Y}(\ell)$ analog to (2). The coefficient adaptation is also widely partition-specific. The partitions of the estimated system functions are predicted from the previous frame's estimation

$$\hat{\mathbf{H}}^+_{j;b}(\ell) = a\hat{\mathbf{H}}_{j;b}(\ell-1). \tag{4}$$

In the next step, the process-noise covariance matrices are calculated as

$$\begin{aligned}\mathbf{\Psi}^{\Delta}_{j,j;b}(\ell-1) = (1-a^2)\big[&\hat{\mathbf{H}}_{j;b}(\ell-1)\hat{\mathbf{H}}^H_{j;b}(\ell-1) + \\ &\mathbf{P}_{j,j;b}(\ell-1)\big],\\ \text{and } \mathbf{\Psi}^{\Delta}_{j,i\neq j;b}(\ell-1) &= \mathbf{0}_{K\times K},\end{aligned} \tag{5}$$

with the predicted state covariance matrices

$$\mathbf{P}^+_{j,i;b}(\ell) = a^2\mathbf{P}_{j,i;b}(\ell-1) + \mathbf{\Psi}^{\Delta}_{j,i;b}(\ell-1). \tag{6}$$

The DFT of the preliminary error signal itself is non-partitioned but summed up over all B partitions,

$$\tilde{\mathbf{E}}(\ell) = \mathbf{Y}(\ell) - \sum_{b\in\mathscr{B}}\sum_{j\in\mathscr{J}} \mathbf{G}\cdot\big(\mathbf{X}_{j;b}(\ell)\cdot\hat{\mathbf{H}}^+_{j;b}(\ell)\big) \tag{7}$$

as well as the measurement noise covariance matrix

$$\begin{aligned}\mathbf{\Psi}^S(\ell) =&(1-\beta)\cdot\Big(\tilde{\mathbf{E}}(\ell)\tilde{\mathbf{E}}^H(\ell)\\ &+ \tfrac{K-L+1}{K}\big(\sum_{b\in\mathscr{B}}\sum_{j\in\mathscr{J}}\sum_{i\in\mathscr{J}}\mathbf{X}_j(\ell)\mathbf{P}^+_{j,i}(\ell)\mathbf{X}^H_i(\ell)\big)\Big)\\ &+ \beta\cdot\mathbf{\Psi}^S(\ell-1)\\ &\text{with smoothing factor } \beta = 0.5\end{aligned} \tag{8}$$

Next, the partition-specific stepsizes are calculated

$$\boldsymbol{\mu}_{j,i;b}(\ell) = \tfrac{R}{K}\mathbf{P}^+_{j,i;b}(\ell)\mathbf{D}^{-1}(\ell) \tag{9}$$

with

$$\begin{aligned}\mathbf{D}(\ell) =&\tfrac{K-L+1}{K}\big(\sum_{b\in\mathscr{B}}\sum_{j\in\mathscr{J}}\sum_{i\in\mathscr{J}}\mathbf{X}_{j;b}(\ell)\mathbf{P}^+_{j,i;b}(\ell)\mathbf{X}^H_{i;b}(\ell)\big)\\ &+ \mathbf{\Psi}^S(\ell).\end{aligned} \tag{10}$$

The Kalman gains are computed as

$$\mathbf{K}_{j;b}(\ell) = \sum_{i\in\mathscr{J}}\boldsymbol{\mu}_{j,i;b}(\ell)\mathbf{X}^H_{i;b}(\ell). \tag{11}$$

and the state-error covariances are updated:

$$\begin{aligned}\mathbf{P}_{j,i;b}(\ell) =&\mathbf{P}^+_{j,i;b}(\ell)\\ &- \tfrac{K-L+1}{K}\mathbf{K}_{j;b}(\ell)\big(\sum_{j\in\mathscr{J}}\mathbf{X}_{j;b}(\ell)\mathbf{P}^+_{j,i;b}(\ell)\big).\end{aligned} \tag{12}$$

Afterwards, the K-dimensional partitions of the estimated system functions can be updated

$$\hat{\mathbf{H}}_{j;b}(\ell) = \hat{\mathbf{H}}^+_{j;b}(\ell) + \mathbf{K}_{j;b}(\ell)\cdot\tilde{\mathbf{E}}(\ell) \tag{13}$$

The estimated feedback signals therefore result in

$$\hat{\mathbf{D}}_j(\ell) = \mathbf{G}\cdot\sum_{b\in\mathscr{B}}\big(\mathbf{X}_{j;b}(\ell)\cdot\hat{\mathbf{H}}_{j;b}(\ell)\big). \tag{14}$$

with the overlap-save constraint matrix

$$\mathbf{G} = \mathbf{F}_{K\times K}\mathbf{Q}\mathbf{Q}^T\mathbf{F}^{-1}_{K\times K}. \tag{15}$$

These are used to calculate the DFT of the error signal

$$\mathbf{E}(\ell) = \mathbf{Y}(\ell) - \sum_{j\in\mathscr{J}}\hat{\mathbf{D}}_j(\ell), \tag{16}$$

resulting in the time domain error signal vector of length K:

$$\mathbf{e}(\ell) = \mathbf{F}^{-1}_{K\times K}\cdot\mathbf{E}(\ell). \tag{17}$$

Only the last R samples of this time domain error signal are used as actual output $e(n)$, due to the underlying overlap-save structure.

As a postfilter for residual echo suppression (RES) the structure as proposed in [11] is used and modified to fit the partitioned processing. This leads to the following

expression:

$$G_{\mathrm{PF}}(\ell,k) = 1 - \sum_{b\in\mathscr{B}} \sum_{j\in\mathscr{J}} X_{j;b}(\ell,k)\mu_{j;b}(\ell,k)X^*_{j;b}(\ell,k) \tag{18}$$

with

$$\mu_{j;b}(\ell,k) = \Big(\Psi_{ss}(\ell,k) \; + \; \tfrac{K-L+1}{K} \cdot \Big(\sum_{b\in\mathscr{B}} \sum_{j\in\mathscr{J}} X_{j;b}(\ell,k)P^+_{j,j;b}(\ell,k)X^*_{j;b}(\ell,k)\Big)\Big)^{-1} \cdot \tfrac{K-L+1}{K} P^+_{j,j;b}(\ell,k). \tag{19}$$

Evaluation

In order to evaluate the algorithm under realistic circumstances, the simulation is divided into 4 different scenarios, each with and without additional in-car noise, denoted as variants A and B, simulating the real surroundings of an ICC system as done in [6]. For the signals audio files of length 45 s from ITU-T Recommendation P.501 [12] are used.

In scenario 1, the enhanced microphone signal $e(n)$ is amplified and played via the rear loudspeakers without any mixing of signals from the infotainment system. The amplification gain for the enhanced microphone signal is chosen as 12 dB for all scenarios.

In scenario 2, an additional mono speech signal is played by the infotainment system on all four loudspeakers of the car. Additionally, the rear speakers play back the 12 dB re-amplified enhanced microphone signal.

In Scenario 3, a stereo speech signal is added to the loudspeakers by the infotainment system, representing, e.g., a radio play. Since this added signal differs between right and left, all four loudspeaker signals are different.

In scenario 4, the infotainment's stereo speech signal is replaced by stereo pop music, which still causes all four loudspeaker signals to differ.

In [13] a delay of 10...20 ms is proposed for ICC systems. With a sample rate of 16 kHz and an algorithmic delay of $3R$ including the postfilter, using a frame shift of $R = 64$ resembles a delay of 12 ms, which we will presume from now on.

For the unpartitioned processing, the length of the estimated filter is directly derived by the FFT size K and frame shift R as $K-R$ due to overlap-save processing. In this (unpartitioned) reference we use an FFT size of $K = 1024$, so the length of the estimated filter is $K - R = 960$ coefficients. To maintain comparability we ensure that the partitioned algorithm holds $B \cdot L = 960$ coefficients throughout each tested configuration. With regard to all other circumstances (reference signals, noise, filter length) the simulation for both, the unpartitioned reference and the partitioned algorithm, are the same. Thus, differences in results are based on the partitioned processing only.

Since the estimated filter length results from the product $B \cdot L$, there are several configurations to achieve a filter length of 960 coefficients, and also various FFT sizes that fit the partition sizes. A larger FFT size allows larger partition sizes, which means less partitions need to be calculated. Smaller FFT sizes require less computational complexity, but also have in consequence the need to calculate more partitions. Since most steps in the filter adaptation algorithm have to be computed for each partition separately, the computational complexity then heavily depends on the number of used partitions.

Scenario		Ref	L192	L160	L120	L96
1	**clean**	12.73	−0.38	+0.13	+0.31	+0.24
	noisy	12.11	−0.25	+0.27	+0.64	+0.34
2	**clean**	16.13	+0.40	+0.74	+0.14	+0.07
	noisy	13.78	+0.13	+0.34	+0.60	+0.46
3	**clean**	12.32	−0.16	+0.38	+0.31	+0.36
	noisy	11.95	−0.00	+0.26	+0.49	+0.35
4	**clean**	15.02	+0.23	+0.48	+0.44	+0.31
	noisy	13.72	+0.28	+0.28	+0.63	+0.62

Table 1: Mean ERLE in [dB] of the partitioned processing **in relation to** the unpartitioned reference with a fixed FFT size $K = 256$, frame shift $R = 64$, without (clean) and with (noisy) in-car noise at 15 dB SNR.

K	L320	L240	L192	L160	L120	L96
512	0.36	0.42	0.47	0.53	0.64	0.74
256	−−	−−	0.31	0.34	0.42	0.50

Table 2: Relative computational complexity c_{rel} of various partitioned configurations compared to the unpartitioned reference algorithm.

To make a statement about the computational complexity, we use a relative complexity score c_{rel} as [8], being the processing time of one partitioned configuration compared to the processing time of the unpartitioned reference, both averaged over five runs. To evaluate the overall echo suppression performance of the algorithms, we provide the mean echo return loss enhancement (ERLE) over the complete audio file length of 45 s.

Measurement Results

Each scenario is tested in a version without (clean) and with (noisy) additional car noise. We tested all possible combinations of FFT sizes $K \in \{1024, 512, 256\}$ with $L \in \{480, 320, 240, 192, 160, 120, 96\}$ at SNR $\in \{15, 10, 5, 0\}$ dB. Due to the limited space, we cannot display all results, but the achieved ERLE values of the reference and a number of exemplary partitioned configurations at SNR of 15 dB are shown in Table 1: The shown partitioned configurations make use of FFT size $K = 256$. It should be noted that the other tested setups show a comparable behavior. In nearly all configurations and scenarios the partitioned processing leads to higher ERLE than the unpartitioned reference, here, best with use of $L = 160$ or $L = 120$ samples partition size. Only with use of $L = 192$ the mean ERLE in scenarios 1 and 3 slightly decreases. Over all scenarios, the maximum improvement is achieved with $L = 120$ in this set. Table 2 shows the relative computational complexity c_{rel} of partitioned processing configurations in relation to the complexity of the unpartitioned reference. It does not depend on the scenario or SNR, therefore the values apply to all tested setups. $K = 1024$ is left out here, since the reference uses the same FFT size and using it for multiple partitions does not aim for decreasing the com-

Scenario		Clean	Noisy: 15 dB	10 dB	SNR 5 dB	0 dB
1	**Ref**	12.73	12.11	11.87	11.42	11.06
	Diff	+0.31	+0.64	+0.66	+0.44	+0.46
2	**Ref**	16.13	13.78	13.09	12.34	11.59
	Diff	+0.14	+0.60	+0.70	+0.85	+0.97
3	**Ref**	12.23	11.95	11.77	11.39	11.08
	Diff	+0.31	+0.49	+0.54	+0.76	+0.82
4	**Ref**	15.02	13.72	13.12	12.47	11.80
	Diff	+0.44	+0.63	+0.73	+0.57	+0.78

Table 3: Difference in mean ERLE in [dB] between the partitioned and unpartitioned processing at different noise levels. For the partitioned processing an FFT size of $K = 256$ and a partition size of $L = 120$ is chosen here.

putational complexity. The table shows that the computational complexity can be reduced to 36% with FFT size $K = 512$ or to 31% with $K = 256$ while still improving the ERLE as shown before. These minimum values are achieved each with the largest partition size possible for this FFT size, therefore with the smallest number of partitions calculated. When decreasing the partition size and increasing the number of partitions, the computational complexity constantly increases.

An interesting observation can be made concerning the effect of different noise levels on the performance of partitioned processing. In Table 3, a well-performing setup with $K = 256$ and $L = 120$ is again compared to the unpartitioned reference at different SNRs. This setup is exemplary chosen, but all examined cases reveal a similar behavior. Whilst ERLE generally decreases with higher noise levels, it can be seen that the relative improvement through partitioning towards the reference even increases for the more challenging scenarios 2 to 4. The improvement in mean ERLE is larger in all cases that include noise, if compared to the improvement without noise. Thus we conclude that the partitioned processing shows a higher robustness to noise than the unpartitioned reference.

Conclusions

We compared partitioned processing to unpartitioned processing of the frequency domain adaptive Kalman filter in an ICC system. Under the very same acoustic conditions, we could determine that partitioned processing provides a set of parameters to fit the algorithm to the respective operating conditions. Echo suppression performance can be improved, while computational complexity can be reduced at the same time depending on requirements. On top of that, the partitioning shows a higher robustness towards noise, which can be very useful—especially for an ICC system, where noisy conditions are typically present.

References

[1] T. van Waterschoot and M. Moonen, "Fifty Years of Acoustic Feedback Control: State of the Art and Future Challenges," *Proceedings of the IEEE*, vol. 99, pp. 288–327, 2011.

[2] G. Schmidt and T. Haulick, "Signal Processing for In-Car Communication Systems," in *Topics in Acoustic Echo and Noise Control* (E. Hänsler and G. Schmidt, eds.), pp. 547–598, Springer, 2006.

[3] E. Hänsler and G. Schmidt, *Topics in Acoustic Echo and Noise Control: Selected Methods for the Cancellation of Acoustical Echoes, the Reduction of Background Noise, and Speech Processing.* Signals and Communication Technology, Springer, 2006.

[4] P. Bulling, K. Linhard, A. Wolf, and G. Schmidt, "Stepsize Control for Acoustic Feedback Cancellation Based on the Detection of Reverberant Signal Periods and the Estimated System Distance," in *Proc. of INTERSPEECH*, (Stockholm, Sweden), pp. 176–180, Aug. 2017.

[5] M. A. Jung, S. Elshamy, and T. Fingscheidt, "An Automotive Wideband Stereo Acoustic Echo Canceler using Frequency-Domain Adaptive Filtering," in *Proc. of EUSIPCO*, (Lisbon, Portugal), pp. 1452–1456, Sept. 2014.

[6] J. Franzen, I. Meyer zum Alten Borgloh, and T. Fingscheidt, "On the Benefit of a Stereo Acoustic Echo Cancellation in an In-Car Communication System," in *Proc. of 13. ITG Conference on Speech Communication*, (Oldenburg, Germany), pp. 41–45, 2018.

[7] F. Kuech, E. Mabande, and G. Enzner, "State-Space Architecture of the Partitioned-Block-Based Acoustic Echo Controller," in *Proc. of ICASSP*, (Florence, Italy), pp. 1295–1299, May 2014.

[8] J. Franzen and T. Fingscheidt, "A Delay-Flexible Stereo Acoustic Echo Cancellation for DFT-Based In-Car Communication (ICC) Systems," in *Proc. of INTERSPEECH*, (Stockholm, Sweden), pp. 181–185, Aug. 2017.

[9] S. Malik and J. Benesty, "Variationally Diagonalized Multichannel State-Space Frequency-Domain Adaptive Filtering for Acoustic Echo Cancellation," in *Proc. of ICASSP*, (Vancouver, Canada), pp. 595–599, May 2013.

[10] S. Malik and G. Enzner, "Recursive Bayesian Control of Multichannel Acoustic Echo Cancellation," *IEEE Signal Processing Letters*, vol. 18, pp. 619–622, Nov. 2011.

[11] J. Franzen and T. Fingscheidt, "An Efficient Residual Echo Suppression for Multi-Channel Acoustic Echo Cancellation Based on the Frequency-Domain Adaptive Kalman Filter," in *Proc. of ICASSP*, (Calgary, Canada), pp. 226–230, Apr. 2018.

[12] "ITU-T Recommendation P.501, Test signals for use in telephonometry." ITU, Jan. 2012.

[13] A. Theiß, G. Schmidt, J. Withopf, and C. Lüke, "Instrumental Evaluation of In-Car Communication Systems," in *Proc. of ITG Conference on Speech Communication*, (Erlangen, Germany), pp. 1–4, Sept. 2014.