A Study on LCMV Filtering for Crosstalk Cancellation in a Distributed Microphone Environment

Frowin Derr[2], Jan Philip Janssen[1], Simon Graf[1], Markus Buck[1] and Tobias Wolff[1]

[1] *Nuance Communications Deutschland GmbH, Email: tobias.wolff@nuance.com*

[2] *Technische Hochschule Ulm, Email: derr@hs-ulm.de*

Abstract

Speech applications in automotive environments often utilize distributed microphones to capture the speech of all passengers and minimize crosstalk. A Multiple Input Multiple Output processing with minimal crosstalk is desirable to enable seat-dedicated hands-free telephony as well as seat-dedicated voice control. To this end different signal processing methods can be applied. In this paper Linearly Constrained Minimum Variance (LCMV) processing is compared to a Minimum Variance Distortionless Response design. The latter minimizes the overall crosstalk power while the LCMV approach cancels each crosstalk component explicitly. Both approaches are compared in terms of their maximum possible crosstalk cancellation gain. Furthermore, temporal adaptation effects are analyzed.

Introduction

For both, hands-free telephony and voice control, the separation of speech signals might be desired. This problem may be approached using the Minimum Variance Distortionless Response (MVDR) design. Here, a distortionless response for the desired speech is ensured while minimizing the variance of any other signal. Since interfering sources are not distinguished their cancellation may suffer in multi-talk scenarios. As opposed to this, the LCMV design [1-3] allows to represent each crosstalk component in an explicit constraint equation. This allows to memorize each interfering source and hence promises improvements over the MVDR. The present study compares the MVDR design in Generalized Sidelobe Canceller (GSC) structure [2,4] to a Direct Form (DF) LCMV design with FIR sub-band filters. Known LCMV filters use a time-invariant constraint matrix [1] in time-domain or 1-tap sub-band processing [3,4,5]. In the following a multi-tap multi-microphone LCMV-filter in the sub-band domain with a time-variant constraint matrix is investigated.

Signal Model

The speech signals of N speakers are captured by $M \geq N$ microphones. The spectrum of the m-th microphone signal can be written as

$$X_m(e^{j\Omega}) = \sum_{\mathrm{n}=1}^{N} F^*_{m,n}(e^{j\Omega}) \cdot Q_n(e^{j\Omega}), \qquad (1)$$

where $F_{m,n}(e^{j\Omega})$ is the acoustic transfer function from the n-th source to the m-th microphone. The spectrum of a source signal is denoted by $Q_n(e^{j\Omega})$. With $S_n(e^{j\Omega}) = F_{n,n}(e^{j\Omega}) \cdot Q_n(e^{j\Omega})$ above equation can be expressed by

$$X_m(e^{j\Omega}) = \sum_{n=1}^{N} G^*_{m,n}(e^{j\Omega}) \cdot S_n(e^{j\Omega}), \qquad (2)$$

where $G_{m,n}(e^{j\Omega}) = F_{m,n}(e^{j\Omega})/F_{n,n}(e^{j\Omega})$ are the relative transfer functions (RTFs). For simplicity we use the convention that the $m\!=\!n$-th microphone is the one closest to the n-th speaker. Each speaker hence has an associated dedicated microphone. Without loss of generality the microphones are considered to be spatially distributed in the vicinity of the respective speakers. In the following an overlap-add framework is considered and hence the microphone signals are modeled in the sub-band domain accordingly. For every sub-band λ the complex sub-band signal in the k-th time frame is modeled as:

$$X_m(\lambda, k) = \sum_{n=1}^{N} \sum_{\kappa=1}^{L_\mathrm{G}} G^*_{m,n}(\lambda, \kappa) \cdot S_n(\lambda, k\!-\!\kappa\!+\!1). \qquad (3)$$

Hence expressing that all sub-bands are considered truly independent. It should be noted in Eq.(3) that the relative transfer functions $G_{m,n}(e^{j\Omega})$ are modeled here as a set of (time-invariant) length L_G sub-band FIR filters:

$$\underline{G}_{m,n} = (G_{m,n}(1), G_{m,n}(2), \ldots, G_{m,n}(L_\mathrm{G}))^T. \qquad (4)$$

In Eq.(4) and from here on underlined quantities denote vectors and the sub-band index λ is omitted for brevity. In the following the goal is to cancel crosstalk from speakers with $n\!\neq\! m$ in the n-th microphone.

Generalized Sidelobe Canceller

For crosstalk cancellation in the n-th microphone the MVDR principle may be implemented as GSC (Fig.1). In the particular realization considered here, the output signals of the $M\!-\!1$ blocking filters $\underline{\hat{G}}_{m,n}(k)$, $m \neq n$, are

$$U_{m,n}(k) = X_m(k) - \underline{\hat{G}}^H_{m,n}(k) \cdot \underline{X}_n(k), \qquad (5)$$

where $\underline{X}_n(k)$ is the vector of the last L_G values of the n-th microphone signal. The blocking filters are updated during single talk activity of the n-th speaker using NLMS:

$$\underline{\hat{G}}_{m,n}(k\!+\!1) = \underline{\hat{G}}_{m,n}(k) + \mu_n(k) \cdot \frac{U^*_{m,n}(k) \cdot \underline{X}_n(k)}{\underline{X}^H_n(k) \cdot \underline{X}_n(k)} \qquad (6)$$

with time variant and speaker specific stepsize $\mu_n(k)$. Upon convergence we get $\underline{\hat{G}}_{m,n}(k) = \underline{G}_{m,n}$.

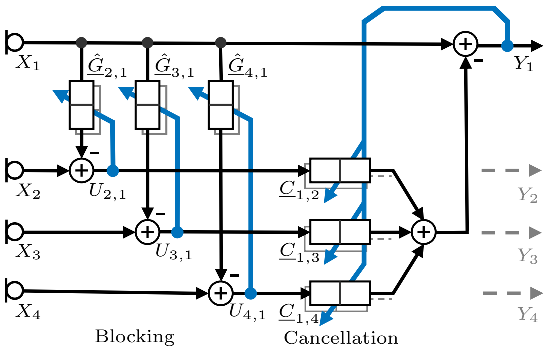

Figure 1: The GSC structure exemplary for $M=4$ microphones in a single sub-band. Here, the signal of the first speaker ($n=1$) should be captured while the cancellation filters $\underline{C}_{1,m}(k)$ cancel the crosstalk of all other speakers jointly.

As can be seen from Fig.1, the signals $U_{m,n}(k)$ are fed to the sub-band FIR filters $\underline{C}_{n,m}(k)$ which jointly minimize crosstalk in the n-th microphone. As in [11], the "fixed beamformer" which usually minimizes spatially uncorrelated noise in a GSC structure is omitted here. Still, this GSC still does not distort $S_n(e^{j\Omega})$. For both, $\underline{\hat{G}}_{m,n}(k)$ and $\underline{C}_{n,m}(k)$, the same filter length L_{G} is chosen. The filters $\underline{C}_{n,m}(k)$ are updated using the multichannel NLMS during interfering speech activity.

Multitap Direct Form Filter

Let $\underline{H}_{n,m} = (H_{n,m}(1), \ldots, H_{n,m}(L_{\mathrm{H}}))^T$ be the length L_{H} sub-band FIR filter applied in the m-th microphone channel whereas $H_{n,m}(1)$ is its first filter tap. The complete system $\underline{H}_n$ is then obtained by stacking the FIRs:

$$\underline{H}_n = (\underline{H}_{n,1}^T, \underline{H}_{n,2}^T, \ldots, \underline{H}_{n,M}^T)^T. \tag{7}$$

Fig.(2) shows a block diagram of the DF-system. The n-th output signal of this system can be written as:

$$Y_n(k) = \underline{H}_n^H \underline{X}(k) = \underline{H}_n^H \mathbf{G}^H \underline{S}(k), \tag{8}$$

where $\underline{S}(k)$ is defined in the same fashion as $\underline{H}_n$ in Eq.(7). The term $\underline{X}(k) = \mathbf{G}^H \underline{S}(k)$ is the vector of stacked input signals. The $NL_c \times ML_{\mathrm{H}}$ matrix $\mathbf{G}$ finally, is the convolution matrix of the relative transfer functions $\underline{G}_{m,n}$. For $M=N=2$ and $L_{\mathrm{H}}=L_{\mathrm{G}}=2$ this matrix looks like:

$$\mathbf{G} = \begin{pmatrix} 1 & 0 & G_{2,1}(1) & 0 \\ 0 & 1 & G_{2,1}(2) & G_{2,1}(1) \\ 0 & 0 & 0 & G_{2,1}(2) \\ G_{1,2}(1) & 0 & 1 & 0 \\ G_{1,2}(2) & G_{1,2}(1) & 0 & 1 \\ 0 & G_{1,2}(2) & 0 & 0 \end{pmatrix}. \tag{9}$$

The parameter $L_c = L_{\mathrm{G}} + L_{\mathrm{H}} - 1$ is the length of the overall system and is referred to as the convolution length.

The system $\underline{H}_n$ in Eq.(7) shall now be chosen such that $Y_n(k) = S_n(k)$, which is equivalent to satisfying:

$$\mathbf{G} \cdot \underline{H}_n = \underline{c}_n. \tag{10}$$

Here, $\underline{c}_n$ is a selection vector being 1 at the entry $(n-1)L_c+1$ and zero otherwise (to select $S_n(k)$ out of $\underline{S}(k)$).

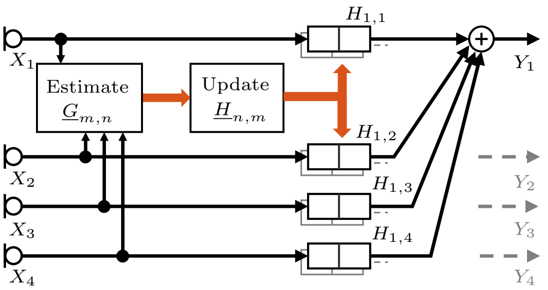

Figure 2: The Direct Form structure including RTF-Estimation and filter update.

For $NL_c = ML_{\mathrm{H}}$, $\mathbf{G}$ can be inverted as it is square and assumed to have full rank (determined case). In the underdetermined case ($NL_c < ML_{\mathrm{H}}$) the Moore-Penrose pseudo-inverse provides the minimum norm solution [6, 7, 8]

$$\underline{H}_n = \mathbf{G}^H (\mathbf{G}\mathbf{G}^H)^{-1} \underline{c}_n \tag{11}$$

which solves Eq.(10) exactly. Note that for $L_{\mathrm{H}} = L_{\mathrm{G}} = 1$, Eq.(11) is identical to the well known LCMV solution for the uncorrelated and homogenous noise field [2, 3]. In the context of dereverberation Eq.(11) is known as the "MINT" solution [6], whereas the matrix $\mathbf{G}$ in Eq.(9) carries a sub-band FIR filter representation of the RTFs rather than time domain impulse responses.

The third possible case is the overdetermined case with $NL_c > ML_{\mathrm{H}}$. Here, Eq.(10) can only be solved approximately. The least-squares solution is also given by the Moore-Penrose pseudo-inverse but this time as the so-called "left-inverse":

$$\underline{H}_n = (\mathbf{G}^H \mathbf{G})^{-1} \mathbf{G}^H \underline{c}_n. \tag{12}$$

The overdetermined case occurs once the number of equations NL_c in Eq.(10) exceeds the available degrees of freedom ML_{H}. This is true if:

$$\frac{M}{N} < \frac{L_{\mathrm{G}} + L_{\mathrm{H}} - 1}{L_{\mathrm{H}}}. \tag{13}$$

Interestingly, if the number of microphones M equals that of the sources N and the order of the acoustic model L_{G} exceeds 1 the system will always be over-determined prohibiting an exact solution. Furthermore, the system length L_{H} cannot be used to reach an under-determined case. As a rule of thumb: For $L_{\mathrm{G}} > 1$ and independent of L_{H} an exact solution can be guaranteed once the number of microphones is at least twice the number of sources.

Remarks. The filters $\underline{G}_{m,n}$ are estimated by $\underline{\hat{G}}_{m,n}$ using the NLMS algorithm as in Eq.(6). Changing the RTF-Estimates $\underline{\hat{G}}_{m,n}$ for one speaker means a rank-L_c change in $\mathbf{G}$ and a rank-$2L_c$ change in $\mathbf{G}^H\mathbf{G}$, respectively $\mathbf{G}\mathbf{G}^H$. Although the efficient implementation was beyond the scope of this study, efficient solutions for updating the pseudo-inverse exist [8-10]. For instance, rank-L_c changes may be split into rank-1 or rank-2 changes and applied iteratively.

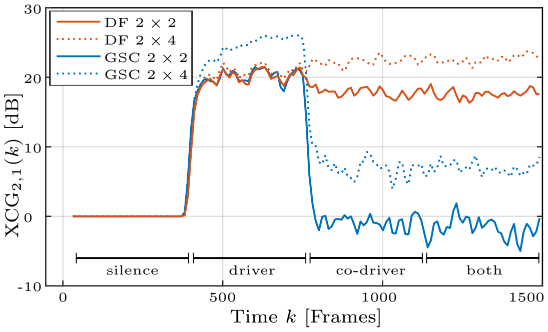

Figure 3: Cancellation gain $\mathrm{XCG}_{2,1}(k)$ of driver into the co-driver's output for 2 × 2 (overdetermined), 2 × 4 (undertermined) GSC and DF system.

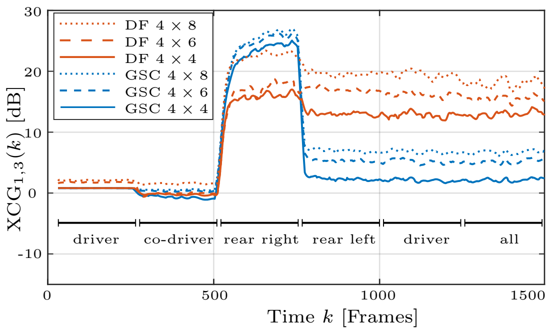

Figure 4: Cancellation gain XCG of rear-right passenger into the driver's output for $4 \times M$ systems, GSC, DF.

Experimental Setup

The two systems, GSC and DF, have been MATLAB simulated with different $(N, M, L_\mathrm{G}, L_\mathrm{H})$-configurations. The Mouth-Enclosure-Microphone impulse responses stem from real in-car measurements (960 taps, 60 ms, 16 kHz). After convolving the source signals and impulse responses, spatially and spectrally white background noise has been added to the signals to obtain an SNR of ≈ 40 dB. The sub-band processing uses an FFT of size $N_\mathrm{FFT} = 512$ with Hann window and a frameshift of $R = 128$ samples. The estimation of $\underline{G}_{m,n}$ uses the NLMS algorithm (Eq.6) with a common base step size of $\mu_0 = 0.1$ for all sub-bands (except Fig.(3) where $\mu_0 = 1$ was used). It is assumed that the speech activity periods are known for each speaker so the stepsize $\mu(k)$ can be controlled to update the respective adaptive filters during single talk periods only. Diagonal loading has been applied prior to any matrix inversion. For this study $L_\mathrm{G} = L_\mathrm{H}$ was chosen[1].

The performance metric to evaluate the described systems is defined as the ratio of speaker related signal terms in input $x_{m,n}(l)$ and output $y_{m,n}(l)$, averaged over all time samples l of one frame k.

$$\mathrm{XCG}_{m,n}(k) = 10 \cdot \lg \left(\frac{\sum_{l=0}^{N_\mathrm{FFT}-1} \|x_{m,n}(l+kR)\|^2}{\sum_{l=0}^{N_\mathrm{FFT}-1} \|y_{m,n}(l+kR)\|^2} \right) \tag{14}$$

E.g. $\mathrm{XCG}_{2,3}(k)$ corresponds to speaker 3 in microphone 2 and output 2. Best case values are $\mathrm{XCG}_{m,n}(k) = 0$ dB for an undistorted signal and ∞ for perfect crosstalk cancellation. Averaging of 20 frames is used to smooth the curves. Systems are characterized as $N \times M$ systems, e.g. 2 × 4 equals 2 speakers into 4 mics.

Temporal Behavior

At first the GSC and the DF system are compared in terms of their temporal cancellation performance. Here, the filter length $L_\mathrm{G} = L_\mathrm{H} = 6$ is used.

2×2 and 2×4 Systems. In Fig. (3) $\mathrm{XCG}_{2,1}(k)$ (cancellation of the driver speech in the co-driver's output) is depicted for both systems. It can be seen that the cancellation gain in the DF system (orange curve) rises with the driver activity and remains stable for the rest of the simulation. Once a speaker and its related RTFs have been estimated, this information is stored in the constraint matrix: "what has been learned is kept in mind". In the GSC system, $\mathrm{XCG}_{2,1}(k)$ (blue curve) shows a similar behavior during driver activity. Once the co-driver becomes active, however, the performance degrades because the blocking filters change which makes the interference canceler filters lose optimality.

4×M Systems. From the 12 cancellation gains, that have to be tracked in systems with 4 speakers and M microphones, $\mathrm{XCG}_{1,3}(k)$ (rear-right passenger cancellation in driver output) is chosen exemplary in Fig.4. As in $2 \times M$ systems, XCG is highly time-varying for the GSC, although with temporarily better values than the DF system. In this example the degradation is caused by the interference canceler filters that "switch" from one interferer to another. As can be seen, the DF systems also show a slight dependency on the speaker situation ("steps") but do yield a cancellation gain greater than 0 dB. However, the best overall values are achieved in the case of the underdetermined 4×8 system.

Filter Length

To analyze the effect of the filter length on the cancellation performance the average cancellation gain after full convergence of all filters is used for comparison. The curves in Fig. (5) demonstrate this influence for $L_\mathrm{G} = L_\mathrm{H}$ in a 2×2 system.

It can be seen that the achieved cancellation gain increases with filter lengths smaller than 10 and saturates afterwards. A filter length of around 10 taps corresponds well with the equivalent acoustic filter length $L_\mathrm{G} = 7.5$ and the number of blocks per frame (512/128 = 4). The Direct Form system achieves similar gains as the GSC. In this particular example even somewhat higher gains can be observed compared to GSC.

Finally, the same simulated speaker sequence as in the previous section is used (every passenger speaks one after

[1] Note that this must not be the case. In particular, it may be considered to choose $L_\mathrm{H} < L_\mathrm{G}$ in order to provide accurate acoustical modeling and to implement these constraints by means of a large M.

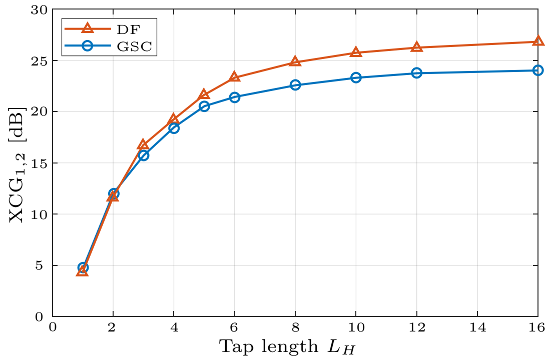

Figure 5: Influence of the filter length L_H for DF and GSC System on the cancellation gain.

another and finally all 4 speakers speak simultaneously) to give an overview on the differences for 4×M systems. Here, each cancellation gain ($\mathrm{XCG}_{1,2}(k) \ldots \mathrm{XCG}_{4,3}(k)$) is averaged during the multi-talk situation at the end of the sequence. The resulting 12 values are concentrated in a box-and-whisker plot in Fig.(6) (min, 1-st quartile, median, 3-rd quartile, max).

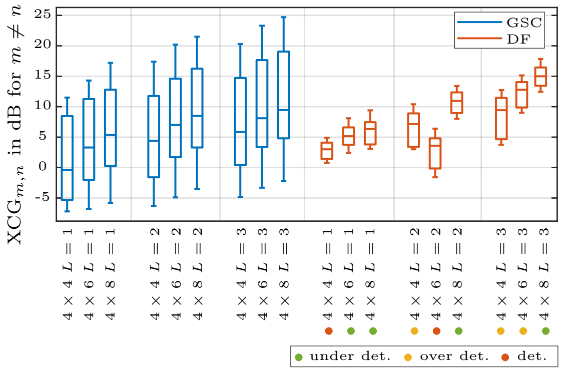

Figure 6: Box-and-whisker plot of the achievable cancellation gain for different 4×M systems, $L = L_\mathrm{G} = L_\mathrm{H} = 1 \ldots 3$.

As can be seen from Fig.(6), increasing the filter-length is beneficial for both GSC and DF. The main difference between both systems becomes obvious when comparing their median values. Here, the DF-system achieves higher values than the GSC. The only exception is the determined 4×6-system with $L = 2$. The reason for this effect is a worse matrix condition as for the under- or overdetermined systems[2]. The DF-system also achieves better results in terms of variance indicating that the ability to memorize RTFs pays off and leads to more consistent results. The GSC on the other hand achieves higher maximum cancellation gains. This can be attributed to the fact that this system always uses all interference canceler filters to minimize the *current* crosstalk. Fig.(6) also shows that underdetermined LCMV systems, e.g. 4×8, $L = 2$ and $L = 3$ show the best performance.

Conclusion

The study addressed speech signal separation in a multi-speaker and multi-microphone scenario. The achievable crosstalk cancellation gain of an MVDR system in GSC structure has been compared to that of a Direct Form system with explicit constraints. The GSC system achieves considerable cancellation gains but crosstalk cancellation depends strongly on the temporal speaker sequence. The system adapts to a "local" minimum. The Direct Form system has proven its potential for signal separation. It allows for a wide range of constraint configurations (number M of microphones or tap lengths L_G and L_H). It optimizes an "overall" cancellation gain. The Direct Form system memorizes the acoustic environment and thereby leads to more consistent results than the GSC.

References

[1] Frost O.L.: An Algorithm for Linearly Constrained Adaptive Array Processing. Proceedings of the IEEE, vol. 60 (1972), 926 - 935.

[2] Van Veen B.D.; Buckley, K.M.: Beamforming: A Versatile Approach to Spatial Filtering. IEEE ASSP Magazine, vol. 5-2 (1988), 4 - 24.

[3] Habets E.A.P., Benesty J.: A Perspective on Frequency-Domain Beamformers in Room Acoustics. IEEE Transactions on Audio, Speech, and Language Processing, vol. 20-3 (2012), 947 - 960.

[4] Gannot S., Burshtein D., Weinstein E.: Signal Enhancement Using Beamforming and Nonstationary with Applications to Speech, IEEE Transactions on Signal Processing, vol. 49-8 (2001).

[5] Markovich S., Gannot S., Cohen I.: A Comparison between alternative Beamforming Strategies for Interference Cancellation in Noisy and Reverberant Environment, Proc. IEEE 25th Convention of Electrical and Electronics Engineers in Israel, 2008.

[6] P.A.Naylor and N.D.Gaubitch, Eds.: Speech Dereverberation, New York, NY, USA, Springer 2010.

[7] Voigt C., Adamy J.: Formelsammlung der Matrizenrechnung, Oldenbourg Wissenschaftsverlag, 2007.

[8] Petersen K.B., Pedersen M.S.: The Matrix Cookbook, May 2012, URL: `http://matrixcookbook.com`

[9] Meyer C.D.: Generalized Inversion of Modified Matrices, SIAM Journal on Applied Mathematics., vol. 24-3 (1973), 315 - 323.

[10] Zielke G.: Inversion of Modified Symmetric Matrices, Journal of the ACM (JACM), vol. 15-3 (1968), 402 - 408.

[11] T. Matheja, M. Buck and T. Fingscheidt: A dynamic multi-channel speech enhancement system for distributed microphones in a car environment. EURASIP Journal on Advances in Signal Processing, 2013, 2013:191.

[2] As can be shown experimentally, depending on the RTFs, a high condition number is more likely to occur for determined systems.