Robust Adaptive Cancellation of Interfering Speakers for Distributed Microphone Systems in Cars

Timo Matheja, Markus Buck, Tobias Wolff

Nuance Communications, 89077 Ulm, Germany, Email: {timo.matheja, markus.buck, tobias.wolff}@nuance.com

Introduction

Hands-free systems in cars aim to capture the speech of different speakers at the best. Therefor distributed microphones can be aligned to each of these speakers and can be mounted in their vicinity. One application of this setup is to optimize speech for speech recognition. It is desirable to cancel the influence of possible interfering speakers in the microphone signals [1, 2]. For this task an adaptive filter is used to cancel the interfering signal from the target one. In contrast to the very similar and well known echo cancellation problem a noise component also occurs on the filter input signal. In this paper an optimal step size is proposed that considers this noise component for controlling the adaptation of a normalized least mean square (NLMS) algorithm [3] in the short-time frequency domain. A Signal-to-Interference-Ratio (SIR) based adaptation control similar to [1] is also investigated. The performance of both approaches is evaluated.

System Overview

The proposed system can include multiple input and multiple output channels but here a two-channel structure with only one target channel is considered. An overview of this simplified arrangement can be seen in Fig. 1. The target microphone signal $X_2(\lambda,k)$ is the sum of a desired part $X_{S2}(\lambda,k)$ from speaker P_2, an interfering component $X_{I2}(\lambda,k)$ from the interfering speaker P_1 and a noise component $X_{N2}(\lambda,k)$ caused by the noise signal N_2:

$$X_2(\lambda,k) = X_{S2}(\lambda,k) + X_{I2}(\lambda,k) + X_{N2}(\lambda,k)\,. \quad (1)$$

The discrete frequency index in the subband domain is denoted by λ and k is the discrete frame index. As the adaptation is stopped during double talk periods we assume single talk ($X_{S2}(\lambda,k)=0$) of the interfering speaker P_1 for the step size derivation. The aim of the processing is to cancel $X_{I2}(\lambda,k)$ by filtering $X_1(\lambda,k) = X_{S1}(\lambda,k)+X_{N1}(\lambda,k)$ and subtracting the filtered signal from the target one. The output of the system results in:

$$Z_2(\lambda,k) = X_2(\lambda,k) - \widehat{\mathbf{H}}^{\mathrm{H}}(\lambda,k)\mathbf{X}_1(\lambda,k)\,, \quad (2)$$

where the upper H represents the hermitian operator and $\widehat{\ }$ denotes an estimation. The vector $\widehat{\mathbf{H}}(\lambda,k) = [\widehat{H}^{(0)}(\lambda,k),...,\widehat{H}^{(L-1)}(\lambda,k)]^{\mathrm{T}}$ contains the coefficients of the adaptive filter with length L. The filter input vector is denoted by $\mathbf{X}_1(\lambda,k) = [X_1(\lambda,k),...,X_1(\lambda,k-L+1)]^{\mathrm{T}}$.

Optimal Step Size

An optimal step size for the NLMS algorithm for the echo cancellation problem is given in [4]. In the following the same approach is used in a more generic way to derive an adaptive step size including the noise components on the signals. The mismatch between the actual transfer function $\mathbf{H}(\lambda,k)$ and the estimated one is called system mismatch vector [4] and is defined as:

$$\Delta\mathbf{H}(\lambda,k) = \mathbf{H}(\lambda,k) - \widehat{\mathbf{H}}(\lambda,k)\,. \quad (3)$$

Assuming $\mathbf{H}(\lambda,k)$ is time-invariant an update rule for $\Delta\mathbf{H}(\lambda,k)$ can be formulated by

$$\Delta\mathbf{H}(\lambda,k+1) = \Delta\mathbf{H}(\lambda,k) - \alpha(\lambda,k)\frac{Z_2^*(\lambda,k)\mathbf{X}_1(\lambda,k)}{||\mathbf{X}_1(\lambda,k)||^2}\,, \quad (4)$$

where $\alpha(\lambda,k)$ denotes the step size. To achieve an optimal step size $\alpha_{\mathrm{opt}}(\lambda,k)$ a cost function $J(\alpha)$ is minimized that is the L_2 norm of the system mismatch vector:

$$J(\alpha) = \mathrm{E}\left\{\|\Delta\mathbf{H}(\lambda,k+1)\|^2\right\}\,. \quad (5)$$

$\mathrm{E}\{\cdot\}$ indicates the expectation operator. In case of convergence the minimization of $J(\alpha)$ results in a fast decrease of the system mismatch. Hence $\alpha_{\mathrm{opt}}(\lambda,k)$ is obtained by differentiating the cost function with respect to $\alpha(\lambda,k)$ and setting to zero. According to the signal model the output signal can be written as the sum of an undisturbed error $Z_{I2}(\lambda,k) = \Delta\mathbf{H}^{\mathrm{H}}(\lambda,k)\mathbf{X}_{S1}$ and a noise component $Z_{N2}(\lambda,k)$. Considering the expression

$$Z_2(\lambda,k) = Z_{I2}(\lambda,k) + Z_{N2}(\lambda,k) \quad (6)$$

it can be written for $\alpha_{\mathrm{opt}}(\lambda,k)$ assuming $X_{S1}(\lambda,k)$ and $X_{N1}(\lambda,k)$ as well as $Z_{I2}(\lambda,k)$ and $Z_{N2}(\lambda,k)$ are uncorrelated and further $||\mathbf{X}_1(\lambda,k)||^2 \approx ||\mathbf{X}_1(\lambda,k+1)||^2$:

$$\alpha_{\mathrm{opt}}(\lambda,k) = \frac{\mathrm{E}\left\{|Z_{I2}(\lambda,k)|^2\right\} + \Re\{C_1\} - \Re\{C_2\}}{\mathrm{E}\left\{|Z_2(\lambda,k)|^2\right\}}\,. \quad (7)$$

In contrast to the well known solution for the echo cancellation problem derived in [4] two additional terms in the nominator occur:

$$C_1 = \mathrm{E}\left\{X_{N2}^*(\lambda,k)\Delta\mathbf{H}^{\mathrm{H}}(\lambda,k)\mathbf{X}_{N1}(\lambda,k)\right\}\,, \quad (8)$$

$$C_2 = \mathrm{E}\left\{\widehat{\mathbf{H}}^{\mathrm{T}}(\lambda,k)\mathbf{X}_{N1}^*(\lambda,k)\Delta\mathbf{H}^{\mathrm{H}}(\lambda,k)\mathbf{X}_{N1}(\lambda,k)\right\}. \quad (9)$$

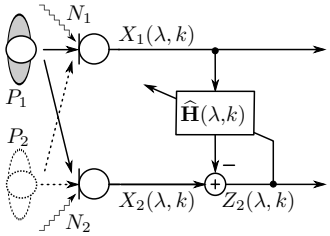

Figure 1: Setup of the interfering speaker cancellation.

These terms originate from the background noise on the reference channel. Because they are not available directly they have to be estimated. In order to play safe the real parts are approximated by negative absolute values:

$$\widehat{\alpha}_{\text{opt}}(\lambda,k) = \frac{\widehat{K}(\lambda,k)\mathrm{E}\left\{|\mathbf{X}_{S1}(\lambda,k)|^2\right\} - |C_1| - |C_2|}{\mathrm{E}\left\{|Z_2(\lambda,k)|^2\right\}}. \quad (10)$$

Obviously the first expression in the nominator was also replaced by the interfering signal multiplied by the coupling factor $\widehat{K}(\lambda,k)$. $\widehat{K}(\lambda,k)$ is estimated during speech periods between the filter input and the system output as proposed in [4]. With the inequality of Schwarz, the assumption of white noise and $\|\Delta\mathbf{H}(\lambda,k)\|^2 = K(\lambda,k)$ it follows for the absolute values of Eq. 8 and Eq. 9:

$$|C_1| \leq \sqrt{\widehat{K}(\lambda,k)\cdot L}\cdot|\sigma_{X_{1N}}(\lambda,k)\sigma^*_{X_{2N}}(\lambda,k)|\,, \quad (11)$$

$$|C_2| \leq \sqrt{\widehat{K}(\lambda,k)}\cdot L\cdot\sigma^2_{X_{1N}}(\lambda,k)\cdot\left\|\widehat{\mathbf{H}}(\lambda,k)\right\|\,. \quad (12)$$

The power of $X_{1N}(\lambda,k)$ is denoted by $\sigma^2_{X_{1N}}(\lambda,k)$. Using these upper bounds for $|C_1|$ and $|C_2|$ Eq. 10 is determined for controlling the adaptation. With the proposed method the step size is robustly kept to zero in noise periods and the adaptation is stopped.

Signal-to-Interference-Ratio

Another method to control the adaptive filter is to analyse the SIR between the two microphone signals. The estimation of the SIR is done whenever the signal power is above the background noise power. With the estimated power spectral densities $\widehat{S}_{X_1}(\lambda,k)$ and $\widehat{S}_{X_2}(\lambda,k)$ of the microphone signals the SIR in channel 2 is computed as:

$$\widehat{\mathrm{SIR}}_2(\lambda,k) = 10\log_{10}\left(\frac{\widehat{S}_{X_2}(\lambda,k)}{\widehat{S}_{X_1}(\lambda,k)}\right)\,. \quad (13)$$

Similar to [1] an adaptive threshold for speech activity detection is determined but here the detection is achieved by tracking the mean negative SIR for each channel. Additionally a threshold based on the signal-to-noise ratio is analysed. If no speech activity is detected the step size is set to zero and otherwise a fixed step size is used.

Evaluation

For evaluation in time domain the proposed optimal step size approach can be compared to the SIR based adaptation control. Therefore the obtained SIR improvement ($\mathrm{SIR}_{\text{imp}}$) and the resulting signal distortion (SD) are considered in a double talk scenario with one target channel. The test signals are created by Lombard speech signals filtered with measured car impulse responses at a microphone distance of 110 cm and mixed according to the introduced signal models. The background noise originates from a car driving at 130 km/h. For voice activity detection an additional broadband decision based on the SIR is used with both approaches. To determine the SIR improvement

$$\mathrm{SIR}_{\text{imp}}(t) = \frac{\mathrm{SIR}_{\text{out}}(t)}{\mathrm{SIR}_{\text{in}}(t)} \quad (14)$$

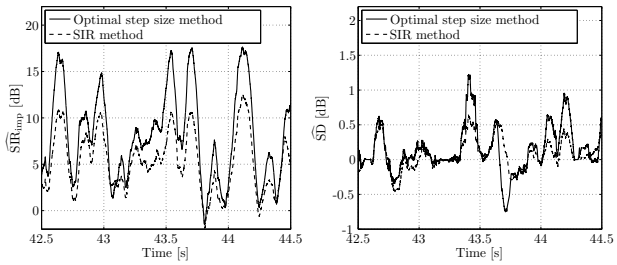

Figure 2: $\mathrm{SIR}_{\text{imp}}$ and SD in dB for an extract of the signal (2 seconds) smoothed by a 80 ms moving average filter.

the output and the input SIR have to be computed whenever speech is present regardless of the active speaker:

$$\mathrm{SIR}_{\text{in}}(t) = \frac{E\left\{x^2_{S2}(t)\right\}}{E\left\{x^2_{I2}(t)\right\}}\,,\ \mathrm{SIR}_{\text{out}}(t) = \frac{E\left\{z^2_{S2}(t)\right\}}{E\left\{z^2_{I2}(t)\right\}}\,. \quad (15)$$

The discrete time index is denoted by t. Further the distortion measure SD is only computed during activity of the target speaker P_2 and is defined as:

$$\mathrm{SD}(t) = \frac{E\left\{x^2_{S2}(t)\right\}}{E\left\{z^2_{S2}(t)\right\}}\,. \quad (16)$$

In Fig. 2 it can be seen that with the optimal step size a higher SIR improvement and a slightly increased undesired distortion is achieved compared to the SIR method. The larger improvement and distortion result from higher filter coefficients in some subbands where the closer microphone offers lower energy than the distant one. In contrast to the optimal step size method the SIR criterion avoids an adaptation in these subbands.

Conclusion

An adaptation control method based on the computation of an approximation of an optimal step size that considers the influence of background noise is proposed. It can be shown that with the proposed optimal step size a higher amount of interfering speaker cancellation is possible compared to an SIR based adaptation control.

References

[1] A. Lombard, W. Kellermann: Multichannel cross-talk cancellation in a call-center scenario using frequency-domain adaptive filtering, *Proc. Int. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, Washington, 2008.

[2] G. Lathoud, J. Bourgeois, J. Freudenberger: Sector-based detection for hands-free speech enhancement in cars, *EURASIP Journal on Applied Signal Processing*, 2006.

[3] B. Widrow, S. D. Stearns: *Adaptive Signal Processing*, Prentice-Hall Signal Processing Series, Upper Saddle River, NJ, USA, 1985.

[4] A. Mader, H. Puder, G. U. Schmidt: Step-size control for acoustic echo cancellation filter - an overview, *Signal Processing*, **80**, no. 9, 1697–1719, 2000.