MULTICHANNEL SPEECH ENHANCEMENT IN CARS: EXPLICIT VS. IMPLICIT ADAPTATION CONTROL

Guillaume Lathoud

IDIAP Research Institute,
Martigny, Switzerland
Email: lathoud@idiap.ch

Julien Bourgeois, Jürgen Freudenberger

DaimlerChrysler Research and Technology,
Ulm, Germany
Email: julien.bourgeois@daimlerchrysler.com
juergen.freudenberger@daimlerchrysler.com

Speech-based command interfaces are becoming more and more common in cars. Applications include automatic dialog systems for hands-free phone calls as well as more advanced features such as navigation systems. However, interferences, such as speech from the codriver, can significantly hamper the performance of the speech recognition component, which is crucial for those applications. This issue can be addressed with *adaptive* interference cancellation techniques such as the Generalized Sidelobe Canceller (GSC). In order to cancel the interference (codriver) while not cancelling the target (driver), adaptation must happen *only* when the interference is active and dominant.

This paper proposes a novel approach for pre-estimation of target and interference energies, along with its application to "explicit" adaptation control: a hard decision whether the filter(s) should be updated or not. It is compared with an "implicit" adaptation control method that does not require such pre-estimation. Two physical setups S1 and S2 are considered, with respectively 2 directional microphones, or 4 directional microphones in the rear-view mirror, as shown in Fig. 1. The spacing between microphones is 17 cm in S1, and 5 cm in S2. Section 1 presents an adaptation algorithm for each setup. Section 2 describes the two adaptation control methods. Experiments on real in-car recordings are presented in Section 3. Section 4 concludes. A detailed version of this paper can be found in [6], and will include additional experimental results in the future.

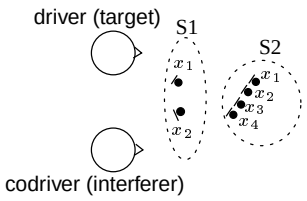

Fig. 1: Physical setup.

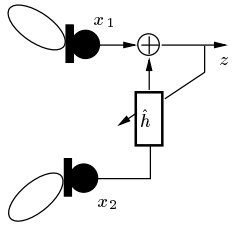

Fig. 2: Noise Canceller.

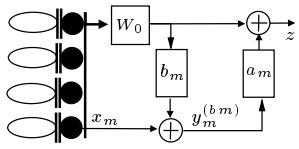

Fig. 3: GSC.

1. ADAPTIVE INTERFERENCE CANCELLATION ALGORITHMS

Setup S1 provides an input Signal-to-Interference Ratio (SIR) of about 6 dB at the driver's microphone signal $x_1(t)$. The other signal $x_2(t)$ is used as a reference, i.e. an estimate of the interference signal. In order to remove the interference from $x_1(t)$, the linear filter depicted by Fig. 2 is used. The filter $\hat{\mathbf{h}}$ is adapted to minimize the output power $\mathbf{E}\{z^2(t)\}$, using the NLMS algorithm [1] with step size μ, as in Eq. 1. To prevent target cancellation, the filter $\hat{\mathbf{h}}$ of length L must be adapted *only* when the interference is active and dominant.

$$\hat{\mathbf{h}}(t+1) = \hat{\mathbf{h}}(t) - \mu \mathbf{E}\{z(t)\mathbf{x}_2(t)\}/\|\mathbf{x}_2(t)\|^2 \quad \text{where} \quad \mathbf{x}_2(t) = (x_2(t), x_2(t-1), \ldots, x_2(t-L+1))^T \tag{1}$$

In setup S2, $M = 4$ directional microphones are in the rear-view mirror, all pointing at the target. It is therefore not possible to use any of them as an estimate of the interference signal. A suitable approach is the linearly constrained minimum variance beamforming [2] and its robust GSC implementation [3]. It consists of two filters b_m and a_m for each input signal $x_m(t)$, with $m = 1 \ldots M$, as depicted by Fig. 3. Each filter b_m (resp. a_m) is adapted to minimize the output power of $y_m^{(b_m)}(t)$ (resp. $z(t)$), as in Eq. 1. To prevent leakage problems, the b_m (resp. a_m) filters must be adapted *only* when the target (resp. interference) is active and dominant.

2. IMPLICIT AND EXPLICIT ADAPTATION CONTROL

For both algorithms described in Section 1, an adaptation control is required that slows down or stops the adaptation according to target and interference energies. Two methods are proposed: "implicit" and "explicit". The implicit method introduces a continuous, adaptive step-size $\mu(t)$, whereas the explicit method relies on a binary decision, whether to adapt or not.

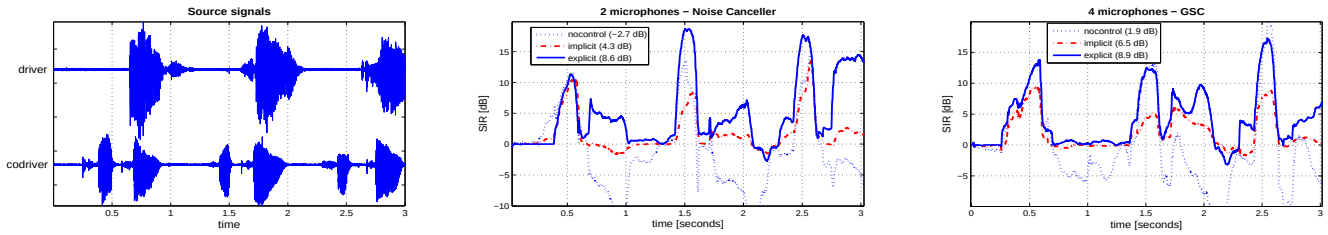

Fig. 4: SIR improvement (100 ms moving average, first 3 seconds shown). Average over 10 seconds is indicated in the legend (brackets).

Implicit method: the system mismatch to the optimum $\mathbf{h}(t)$ is defined as $\mathbf{m}(t) \triangleq \mathbf{h}(t) - \hat{\mathbf{h}}(t)$ and the actual error signal $\epsilon(t) \triangleq \mathbf{m}(t)^T \mathbf{x}_2(t)$. It can be shown [4] that an optimal step-size is given by $\mu(t) = \mathbf{E}\{\epsilon^2(t)\}/\mathbf{E}\{z^2(t)\}$. For a white excitation signal we can estimate $\mathbf{E}\{\epsilon^2(t)\}$ as: $\mathbf{E}\{\epsilon^2(t)\} = \mathbf{E}\{x_2^2(t)\}\mathbf{E}\{\|\mathbf{m}(t)\|\}$.
As $\mathbf{E}\{\|\mathbf{m}(t)\|\}$ is unknown, we approximate it with a constant μ_0 close to the system mismatch expected when close to convergence. In setup S1 we can replace Eq. 1 with: $\hat{\mathbf{h}}(t+1) = \hat{\mathbf{h}}(t) - \mu_0 \mathbf{E}\{z(t)\mathbf{x}_2(t)\}/\|\mathbf{z}(t)\|^2$. This effectively reduces the step-size when the current target power estimate is large and conversely it adapts faster in absence of the target. This approach also applies to filters a_m in setup S2. For filters b_m in setup S2, this turns to $\mu = \mu_0/\|\mathbf{y}_m^{(bm)}\|^2$.

Explicit method: for both setups, a sector-based audio source detection method [5] was developed to directly estimate the input Signal-to-Interference Ratio (SIR) at $x_1(t)$, as detailed in [6]. It is a frequency-domain method, which determines, *for each frequency bin*, which of the target or interference is the most active. It does not use any post-processing, it only relies on geometrical knowledge. The estimated input SIR is then used to determine when the target (respectively the interference) is dominant, by setting two thresholds. This decision determines whether or not the fixed step-size adaptation should be applied.

3. EXPERIMENTS

For both setups, we measured the SIR improvement over microphone $x_1(t)$ on real 16 kHz recordings in a Mercedes S320 vehicle. A 5-second recording with artificial heads was used to tune all parameters in the methods described in Sections 1 and 2. Another 10-second recording with real human speakers was used to test the methods. It contains a significant degree of overlap between the two speakers (56% of speech frames). Filters of length $L = 256$ were used, with half overlapping frames of length 64 samples.
Estimation of input SIR: In log domain, a RMS error of 2.36 dB and 2.31 dB are obtained for setups S1 and S2, respectively.
Adaptation: The first 3 seconds are depicted in Fig. 4. It includes a baseline method that does not use any adaptation control. The baseline result highlights the target cancellation problem and confirms the necessity of adaptation control. Both implicit and explicit methods are robust against this problem, and the explicit method provides the best results. This analysis is valid for both setups, and is confirmed by average SIR improvement values over the entire recording (values in brackets).

4. CONCLUSION

Two adaptation control methods were proposed to cancel the codriver interference from the driver's speech signal. At no additional cost, the "implicit" adaptation method provides robustness against leakage, but slower convergence. The main contribution of this paper is a novel technique for input SIR estimation, along with its application to "explicit" adaptation control. It provides both robustness and better performance. Both implicit and explicit methods are suitable for real-time implementation. Future work includes tests on other noise cases, including background road noise and other passengers. A detailed version of this paper can be found in [6] and will be updated in the future with new experimental results.
The authors acknowledge the support of the European Union through the HOARSE project.

References

[1] B. Widrow and S.D. Stearns, *Adaptive Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1985.
[2] L. J. Griffiths and C. W. Jim, *An Alternative Approach to Linearly Constrained Adaptive Beamforming*, IEEE Trans. Antennas and Propagation, vol. AP-30, no. 1, pp. 27–34, January 1982.
[3] O. Hoshuyama and A. Sugiyama, *A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix using Constrained Adaptive Filters*, in Proc. Int. Conf. Acoustics, Speech, Signal Process., Atlanta, GA, May 1996.
[4] A. Mader, H. Puder and G. Schmidt: *Step-Size Control for Acousic Echo Cancellation Filters - An Overview*, Signal Processing, vol. 80, no. 9, pp. 1697–1719, September 2000.
[5] G. Lathoud and M. Magimai.-Doss, *A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers*, To appear in Proc. Int. Conf. Acoustics, Speech, Signal Process., Philadelphia, PA, March 2005.
[6] G. Lathoud, J. Bourgeois and J. Freudenberger, *Sector-Based Detection for Speech Enhancement in Cars*, IDIAP Research Report RR-04-67 (ftp://ftp.idiap.ch/pub/reports/2004/rr-04-67.pdf), December 2004.