MULTI-CHANNEL SPEECH ENHANCEMENT FOR CAR APPLICATIONS

Huajun Yu, Tim Fingscheidt

Institute for Communications Technology, TU Braunschweig, D–38106 Braunschweig
E-Mail: {huajun.yu, t.fingscheidt}@tu-bs.de

Abstract: Beamforming for car applications has gained much attention in the past. With upcoming wideband speech telephony appropriate solutions operating at a sampling frequency of 16 kHz are required. While a number of proposals aims at quite idealistic conditions, our approach intentionally employs low-cost microphones, and it is optimized and tested with real multi-channel signals acquired using these sensors. Moreover, we assume the microphone array to be integrated into the head-unit of the car. Although from a signal-to-noise ratio perspective this is not an ideal location, it is yet very attractive, since no further wiring is necessary and radio navigation systems manufacturers can offer compact and optimized solutions. To achieve the required level of directivity (and therefore noise reduction) in car noise, we exploit the *a priori* noise field coherence of diffuse noise. An adaptive smoothing approach for post-filter estimation along with a new combination of the beamformer and the post-filter is proposed well suited for the low-cost microphones. Meanwhile, an intrusive instrumental evaluation methodology will be introduced. We will show that a significant level of noise attenuation can be achieved, while simultaneously the quality of the speech component will be improved compared to the state of the art.

1 Introduction

Nowadays, a hands-free equipment comprising speech enhancement is mandatory to allow safe (and convenient) telecommunications in a car. With upcoming wideband speech telephony (sampling frequency $f_s = 16$ kHz) a bandwidth of $50 \cdots 7000$ Hz has to be supported. For this field of application multi-channel microphone array techniques have drawn lots of interest. Compared to single-channel speech enhancement, microphone array based beamforming algorithms exploit not only spectral information but also spatial information. An example of a typical beamformer approach is the minimum variance distortionless response (MVDR) beamformer [1], which includes the delay-and-sum beamformer, and the superdirective beamformer. However, in practice, especially in car environment, applying a beamformer alone achieves insufficient noise attenuation. This is due to the characteristics of the car noise energy dominating mostly in the low frequency region, in which the beamformer has a very low directivity factor. Concerning this technical weakness of beamformer several approaches have been proposed to cope with it. One of these approaches is the multi-channel Wiener filter. Simmer et al. [1] have shown that the multi-channel Wiener filter (i.e., the optimal broadband minimum mean square error (MMSE) estimator), can be decomposed into a single-channel Wiener post-filter working on the output of the MVDR beamformer. A popular post-filter technique, which employs the multi-channel Wiener filter structure, was first reported by Zelinski [2]. However, Zelinski assumes the noise of different microphone signals to be uncorrelated which leads to an ideal incoherent noise field. Unfortunately, this is not correct in many applications including the car noise environment, in which the microphones are usually closely spaced. This exhibits a high correlation for noises in the low frequency region. It has been reported in [3] that a diffuse noise

field is an appropriate noise field model in a car environment. By adopting this *a priori* noise field model, McCowan et al. have extended the Zelinski post-filter to a noise coherence based post-filter structure [4] showing improved noise attenuation performance.
In this paper we present a beamformer and post-filter solution for a microphone array with 4 microphones *integrated* into the head-unit of a car. It turns out to be cost-effective not only due to the use of cheap microphones, but particularly because extensive wiring to some typical hands-free microphone positions such as rear mirror or light module can be omitted. However, the challenge of the acoustically sub-optimal head-unit position is to identify appropriately working beamformer and post-filter algorithms. Instead of using the combination of a superdirective beamformer with the McCowan post-filter as in [4], we will combine the very robust yet hardly effective delay-and-sum beamformer with the McCowan post-filter. Furthermore, the McCowan post-filter will be appropriately modified by using an adaptive smoothing factor for the auto- and cross-power spectral densities (psd) estimation, which is essential for the post-filter estimation [5]. In order to instrumentally evaluate the performance of the different post-filters for noise attenuation, an intrusive instrumental evaluation methodology will be introduced and applied. This framework gives us the possibility for evaluating the noise attenuation performance and the quality of the isolated speech component by applying the signal-to-noise ratio improvement (ΔSNR) and the mean opinion score (MOS), respectively.
The paper is organized as follows: In Section 2 we will recapitulate the relevant beamformer algorithms along with the formulation of the baseline Zelinski post-filter and the McCowan post-filter. Our new modified approach for the McCowan post-filter estimation will be presented in Section 3. Furthermore, the new structure of combining the delay-and-sum beamformer with the McCowan post-filter estimated by the modified approach will be shown. An outline of the intrusive instrumental evaluation methodology, the simulation setup, and a discussion of the experimental results will be given in Section 4.

2 Beamforming and Post-filtering

Let us regard a microphone array with M channels. After applying the short-time Fourier transform of length K, the vector of microphone signals can then be formulated with frame index ℓ and frequency bin k as $\mathbf{Y}'(\ell,k) = \mathbf{S}'(\ell,k) + \mathbf{N}'(\ell,k)$ with noisy signal $\mathbf{Y}'(\ell,k) = (Y_1'(\ell,k)\; Y_2'(\ell,k) \cdots Y_M'(\ell,k))^T$, additive noise $\mathbf{N}'(\ell,k) = (N_1'(\ell,k)\; N_2'(\ell,k)\; \cdots\; N_M'(\ell,k))^T$, and speech $\mathbf{S}'(\ell,k) = S(\ell,k) \cdot \mathbf{D}(k)$. The term $S(\ell,k)$ denotes the desired source signal and $(\cdot)^T$ is the vector transpose. $\mathbf{D}(k)$ is the propagation vector modeling the delays of each channel for the desired source signal based on a reference microphone depending on the microphone array geometry

$$\mathbf{D}(k) = \left(\exp \frac{-j2\pi k\tau_1}{c} \;\cdots\; \exp \frac{-j2\pi k\tau_M}{c} \right)^T, \tag{1}$$

with c being the speed of the sound.
Following this signal model the multi-channel signals $\mathbf{Y}'(\ell,k)$ will be filtered by the well studied MVDR beamformer (see, e.g., [1])

$$\mathbf{W}_{\mathrm{MVDR}}(\ell,k) = \frac{\mathbf{\Phi}_{NN}^{-1}(\ell,k)\mathbf{D}(k)}{\mathbf{D}^H(k)\mathbf{\Phi}_{NN}^{-1}(\ell,k)\mathbf{D}(k)}, \tag{2}$$

with $\mathbf{W}_{\mathrm{MVDR}}(\ell,k)$ being the filter coefficients vector, $\mathbf{\Phi}_{NN}(\ell,k)$ being the $M \times M$ normalized cross-power spectral density matrix of the noise, and $(\cdot)^H$ denoting the Hermitian operator, respectively.

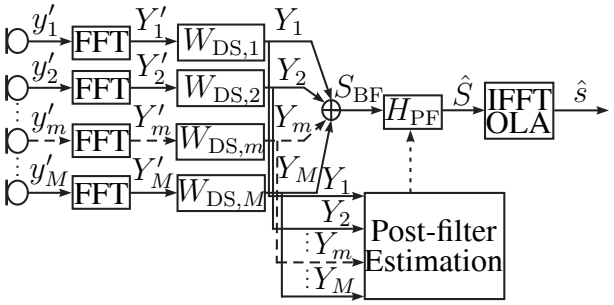

Figure 1 - Block diagram of a delay-and-sum (DS) beamformer with a post-filter.

The single-channel beamformer output is then given by

$$S_{\mathrm{BF}}(\ell,k) = \mathbf{W}^H_{\mathrm{MVDR}}(\ell,k) \cdot \mathbf{Y}'(\ell,k). \tag{3}$$

Due to the weak directivity of the MVDR beamformer in the low frequency region, the performance of the MVDR beamformer is limited for a car environment, where its noise energy dominates in the low frequency region. Hence, a multi-channel Wiener filter, which can be decomposed into an MVDR beamformer followed by a single-channel Wiener filter, is often utilized to improve the limited performance in terms of the noise attenuation [1]. The multi-channel Wiener filter can be formulated as

$$\mathbf{W}_{\mathrm{opt}}(\ell,k) = \mathbf{W}_{\mathrm{MVDR}}(\ell,k) \cdot H_{\mathrm{PF}}(\ell,k), \tag{4}$$

where the Wiener post-filter (PF) is defined as

$$H_{\mathrm{PF}}(\ell,k) = \frac{\phi_{SS}(\ell,k)}{\phi_{SS}(\ell,k) + \phi_{NN}(\ell,k)}, \tag{5}$$

with $\phi_{SS}(\ell,k)$ and $\phi_{NN}(\ell,k)$ being the clean speech signal and noise auto-power spectral densities after beamforming, respectively.
With (3), and (4), the output of the post-filter in the frequency domain is given by

$$\hat{S}(\ell,k) = H_{\mathrm{PF}}(\ell,k) \cdot \mathbf{W}^H_{\mathrm{MVDR}}(\ell,k) \cdot \mathbf{Y}'(\ell,k). \tag{6}$$

The structure of a post-filter as proposed by Zelinski [2] is shown in Fig. 1. Zelinski has used the delay-and-sum (DS) beamformer, which is actually a special case of the MVDR beamformer with $\mathbf{W}_{\mathrm{DS}}(k) = \frac{1}{M}\mathbf{D}(k)$ under the assumption of an homogeneous incoherent noise field. The Zelinski post-filter is given by:

$$H_{\mathrm{ZE}}(\ell,k) = \frac{\frac{2}{M(M-1)} \sum\limits_{i=1}^{M-1} \sum\limits_{j=i}^{M} \mathrm{Re}\left\{\hat{\phi}_{Y_iY_j}(\ell,k)\right\}}{\frac{1}{M} \sum\limits_{i=1}^{M} \hat{\phi}_{Y_iY_i}(\ell,k)}, \tag{7}$$

with $\mathrm{Re}\{\cdot\}$ being the real operator, used to force $\hat{\phi}_{SS}(\ell,k)$ in the numerator to be real-valued. The auto-power spectral densities $\hat{\phi}_{Y_iY_i}(\ell,k)$ and cross-power spectral densities $\hat{\phi}_{Y_iY_j}(\ell,k)$ are estimated recursively as

$$\begin{aligned}
\hat{\phi}_{Y_iY_i}(\ell,k) &= \alpha\hat{\phi}_{Y_iY_i}(\ell-1,k) + (1-\alpha)Y_i^*(\ell,k)Y_i(\ell,k) \in \mathbb{R}, \\
\hat{\phi}_{Y_iY_j}(\ell,k) &= \alpha\hat{\phi}_{Y_iY_j}(\ell-1,k) + (1-\alpha)Y_i^*(\ell,k)Y_j(\ell,k) \in \mathbb{C},
\end{aligned} \tag{8}$$

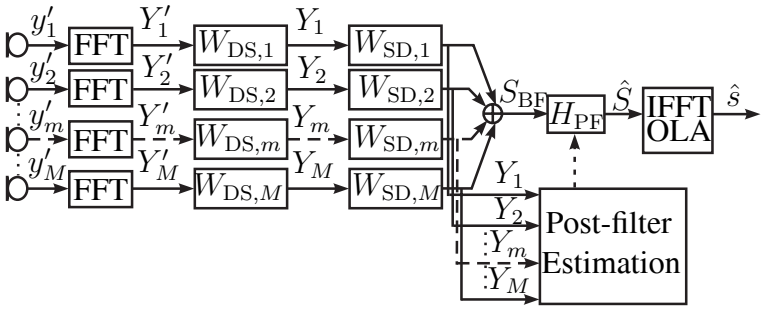

Figure 2 - Block diagram of a filter-and-sum (superdirective, SD) beamformer with a post-filter.

where α is a fixed smoothing factor.
However, according to [3] the car noise field can be well modeled as a diffuse noise field. For a diffuse noise field the coherence function between two microphones is given as $\Gamma_{ij}(k) = \mathrm{sinc}\left(\frac{2\pi k d_{ij}}{c}\right)$ with d_{ij} being the distance between two microphones i and j [6] . Accordingly, the $M \times M$ noise coherence matrix for the diffuse noise field is $\mathbf{\Gamma}_{NN,\mathrm{dif}}(k) = \Big(\Gamma_{ij}(k)\Big)$. With this *a priori* coherence matrix, McCowan et al. have extended the Zelinski post-filter [4].
Fig. 2 shows the structure of the McCowan post-filter with the superdirective (SD) beamformer, which is a special case of the MVDR beamformer using the noise coherence matrix for the diffuse noise field [7]:

$$H_{\mathrm{MC}}(\ell,k) = \frac{\frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i}^{M} \hat{\phi}_{SS}^{(ij)}(\ell,k)}{\frac{1}{M} \sum_{i=1}^{M} \hat{\phi}_{Y_i Y_i}(\ell,k)}, \tag{9}$$

$$\hat{\phi}_{SS}^{(ij)}(\ell,k) = \frac{\mathrm{Re}\left\{\hat{\phi}_{Y_i Y_j}(\ell,k)\right\} - \mathrm{Re}\left\{\Gamma_{ij}(k)\right\} \beta_{ij}(\ell,k)}{1 - \mathrm{Re}\left\{\Gamma_{ij}(k)\right\}},$$

with $\beta_{ij}(\ell,k) = \frac{1}{2}\left[\hat{\phi}_{Y_i Y_i}(\ell,k) + \hat{\phi}_{Y_j Y_j}(\ell,k)\right]$ and $\Gamma_{ij}(k)$ being the coherence function between two microphones.

3 Modified Post-filter Estimation

It has been shown in [4] that $H_{\mathrm{MC}}(\ell,k)$ achieves improved noise attenuation performance in a diffuse noise field. However, speech distortion and musical tones can still be perceived using McCowan post-filter. According to Guerin et al. [5], the smoothing factor α can be adaptively estimated as

$$\alpha(\ell,k) = \alpha_1 - \alpha_2 \cdot \frac{\mathrm{SNR}(\ell,k)}{1 + \mathrm{SNR}(\ell,k)}, \tag{10}$$

with SNR being the signal-to-noise ratio at the beamformer output. According to (10) $\alpha(\ell,k)$ will be limited to the interval $[\alpha_1 - \alpha_2, \alpha_1]$. For a low SNR, $\alpha(\ell,k)$ will reach its upper limit α_1, leading to a smooth estimation of the auto- and cross-power spectral densities. This will limit the occurrence of musical tones [5]. In cases where SNR is high, $\alpha(\ell,k)$ will reach its minimum $\alpha_1 - \alpha_2$, leading to a good estimation for fast speech variation. Since the SNR does not change so much frame by frame, the SNR term in (10) can be approximated by

$$\frac{\mathrm{SNR}(\ell,k)}{1 + \mathrm{SNR}(\ell,k)} \cong H_{\mathrm{MC}}(\ell-1,k), \tag{11}$$

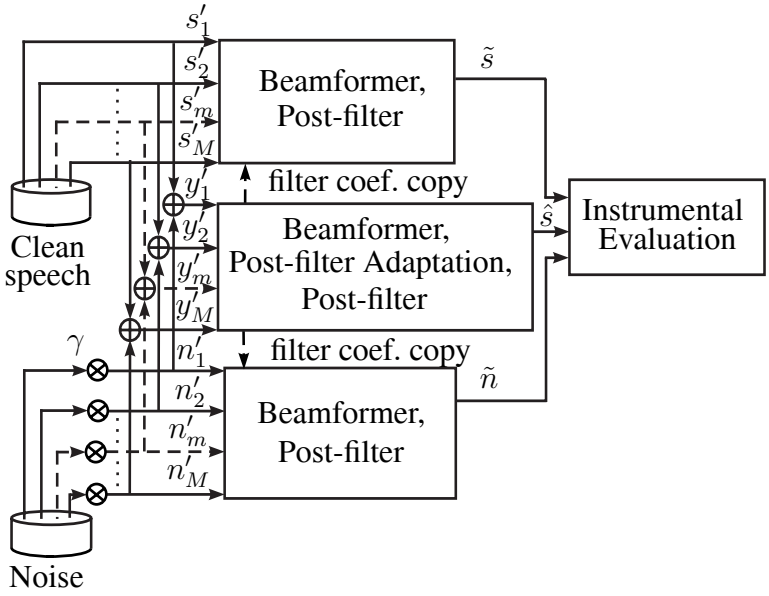

Figure 3 - Block diagram of the instrumental evaluation setup.

which leads to $\alpha(\ell, k) = \alpha_1 - \alpha_2 \cdot H_{\mathrm{MC}}(\ell - 1, k)$.
Care has to be taken in choosing the parameters α_1 and α_2, in order to avoid reverberation-like effects. Unlike the values given in [5], we found $(\alpha_1, \alpha_2) = (0.8, 0.5)$ to be optimal.
Another problem arises when a *superdirective* beamformer is used in a practical situation, since it is very sensitive to the self-noise amplification of the uncorrelated microphone noises and the precision of the microphone positions. Therefore, a *constrained* superdirective beamformer has to be used which degrades the beamformer performance [7]. In contrast to that the delay-and-sum beamformer behaves very robust in practice and is easier to implement. It has been reported in [4] that the beamformer alone achieves only a very small contribution to the noise attenuation compared to the post-filter. Hence, in our work along with the McCowan post-filter and the adaptive power spectral density estimation with appropriate smoothing factor limits, the more robust delay-and-sum beamformer will be utilized. This choice is even more recommended, as we employ low-cost microphones.

4 Instrumental and Subjective Evaluation

4.1 Methodology and Simulation Setup

In this paper an intrusive instrumental evaluation methodology is used to compare the wideband speech enhancement performance of different MVDR beamformers with single-channel Wiener post-filters. As shown in Fig. 3, the clean speech signal $s'_i(n)$ and noise signal $n'_i(n)$ with $i = 1, \cdots, M$ are the inputs to the speech enhancement system consisting of beamformer and post-filter. The post-filter coefficients will be computed based on the synthetically generated noisy signal $y'_i(n) = s'_i(n) + n'_i(n)$, where $n'_i(n)$ has been achieved by scaling multi-channel noise with a factor γ, yielding pre-defined values of the input signal-to-noise ratio $\mathrm{SNR}_{\mathrm{in}}$ according to the active speech level of ITU-T Recommendation P.56 [8].
The enhanced speech signal can be expressed by its components as $\hat{s}(n) = \tilde{s}(n) + \tilde{n}(n)$. By separate processing of the speech components $s'_i(n)$ and of the noise components $n'_i(n)$, we get the speech component of the output signal $\tilde{s}(n)$, and the attenuated (and distorted) noise component $\tilde{n}(n)$, respectively. The output signal-to-noise ratio $\mathrm{SNR}_{\mathrm{out}}$ can then be calculated. In this paper, the signal-to-noise ratio improvement ΔSNR, which is the difference between the $\mathrm{SNR}_{\mathrm{out}}$ and $\mathrm{SNR}_{\mathrm{in}}$, and the mean opinion score (MOS) are used to evaluate the noise attenuation performance and the quality of the speech. The wideband $\mathrm{MOS}_{\tilde{s}}$ with reference signal

Figure 4 - Head-unit-integrated microphone array with 30 channels for research in a middle-class car.

$s(n)$ being chosen as the best clean speech signal from $s_i'(n)$ is computed according to ITU-T Recommendation P.862.2 [9].
The wideband speech performance assessment is carried out in a Volkswagen upper middle-class car Passat, employing an experimental head-unit-integrated microphone array with 30 channels as shown in Fig. 4. The applied microphone array consists of $M = 4$ microphones with $3.6\ \mathrm{cm}$ distance located on the left. The microphones have been randomly selected from a delivery of type MCE-4500 by Monacor. Multiple recordings are made synchronously for each channel with the clean speech being spoken from the driver position and the background noises recorded separately. In our experiment, two background noise conditions are investigated: (1) car engine in an idle state, window closed, air condition being set at 50%; (2) car driven on an expressway with a speed of $50\ \mathrm{km/h}$, window closed, air condition being set at 50%. $\mathrm{SNR_{in}}$ is scaled to values of -5 dB, 0 dB, 5 dB, 10 dB, and 15 dB. $\Delta\mathrm{SNR}$ and $\mathrm{MOS}_{\tilde{s}}$ are calculated and averaged over the whole data set being recorded in each of the background noise conditions.
The sampling frequency $f_s = 16\ \mathrm{kHz}$ is used. The signal is windowed by a Hann window of length 512, followed by an FFT with length $K = 512$ and a frame shift of 50% samples.
Figs. 5-8 show the results for different beamformer and post-filter combinations: Delay-and-sum (DS) or constrained superdirective (SD) beamformer, Zelinski [2] (ZE) or McCowan [4] (MC) post-filter, finally combined with a fixed (FIX) or an adaptive (ADA) smoothing factor for psd estimation.

4.2 Experimental Results

Fig. 5 shows the $\Delta\mathrm{SNR}$ results for different approaches in the first background noise condition. As expected the DS-ZE-FIX and DS-ZE-ADA approaches hardly provide any signal-to-noise ratio improvement. Employing the *a priori* noise field coherence (MC), the SD-MC-FIX approach achieves a well improved $\Delta\mathrm{SNR}$ compared to the DS-ZE-$*$ approaches. By adopting the proposed adaptive smoothing factor for psd estimation (SD-MC-ADA) the $\Delta\mathrm{SNR}$ of SD-MC-FIX can be improved further up to 0.5 dB. However, as we have explained in Section 3, the superdirective beamformer is very sensitive to the self-noise amplification and precision of the microphone positions, so that its performance will be degraded in a real implementation. Hence, as expected, Fig. 5 reveals that our new approach of combining the very robust delay-and-sum beamformer with the McCowan post-filter and the adaptive smoothing factor for psd estimation (DS-MC-ADA) achieves a significant improvement in $\Delta\mathrm{SNR}$ compared to all other approaches.
Fig. 7 shows the $\mathrm{MOS}_{\tilde{s}}$ values of the above approaches. As expected, the delay-and-sum beamformer with fixed and adaptive smoothing factor for the Zelinski post-filter preserves speech ex-

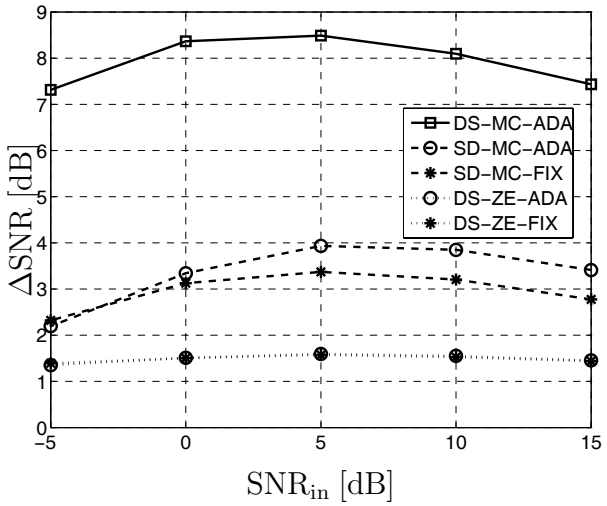

Figure 5 - ΔSNR for car engine in an idle state, window closed, and 50% level air conditioning.

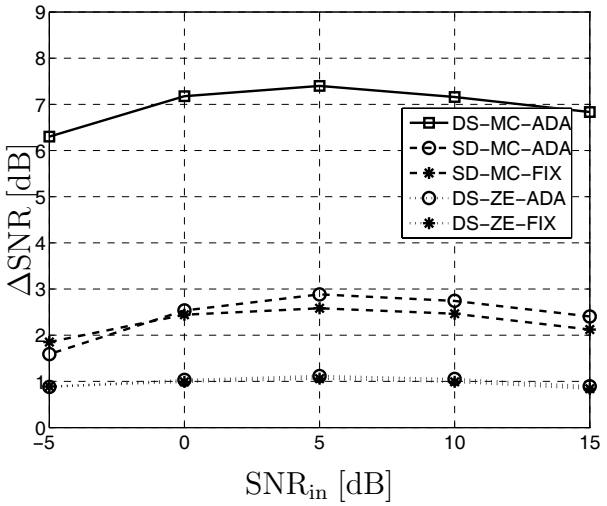

Figure 6 - ΔSNR for car driven with 50 km/h, window closed, and 50% level air conditioning.

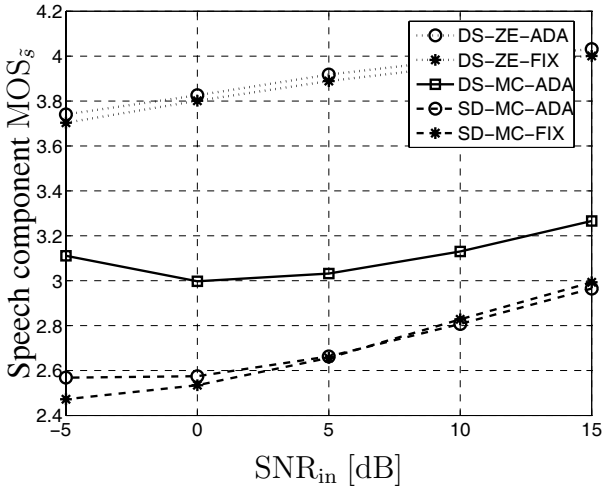

Figure 7 - Speech component $MOS_{\tilde{s}}$ for car engine in an idle state, window closed, and 50% level air conditioning.

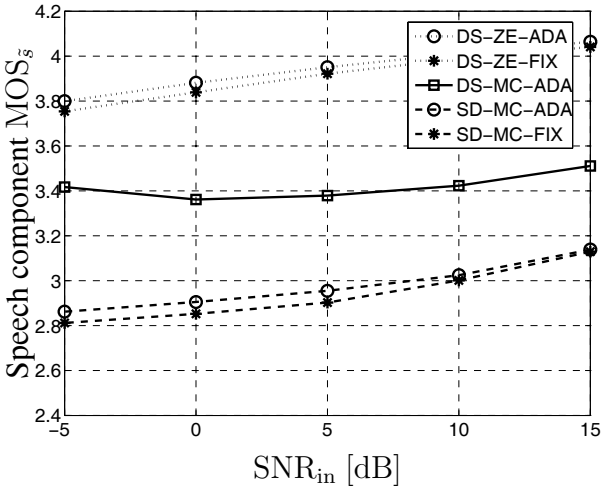

Figure 8 - Speech component $MOS_{\tilde{s}}$ for car driven with 50 km/h, window closed, and 50% level air conditioning.

traordinary well, however, without achieving perceivable noise attenuation. On the other hand, using the new approach an improvement of 0.2 to 0.4 $\mathrm{MOS}_{\tilde{s}}$ points can be achieved against the SD-MC-$*$ approaches, respectively. Even for $\mathrm{SNR}_{\mathrm{in}}$ with -5 dB and 0 dB, the $\mathrm{MOS}_{\tilde{s}}$ of the speech component still stays above 3.0 points. We found that the quality of the speech component is improved while even achieving a significant noise attenuation performance. Finally, the musical tones and reverberation-like effects are very much reduced by using the new approach. Figs. 6 and 8 show the results of $\Delta\mathrm{SNR}$ and $\mathrm{MOS}_{\tilde{s}}$ for the second background noise condition, respectively, similar conclusions can be drawn as in the first background noise condition. However, in all approaches noise attenuation performance has decreased to some extent. This is due to the problem of the mismatch between the moving car noise field and the diffuse noise field model. Yet the $\mathrm{MOS}_{\tilde{s}}$ values of all approaches have improved, which shows a trade-off of noise attenuation performance and the preservation of the speech component. Again, the proposed DS-MC-ADA approach achieves by far the best noise attenuation performance while still attaining a better quality of the speech component compared to the constrained superdirective approaches.

5 Conclusions

In this paper, we have presented a beamformer solution for a new head-unit-integrated microphone array with 4 low-cost microphones. Using an intrusive instrumental evaluation, we have shown that the combination of the robust delay-and-sum beamformer with the McCowan postfilter estimated using an adaptive smoothing factor has achieved a significant noise attenuation performance while still preserving the speech component to a large extent. Furthermore, musical tones and reverberation-like effects have been avoided with high fidelity.

6 Acknowledgments

We thank Georg Eisner and Gordon Seitz of Volkswagen AG, Wolfsburg, for a lot of helpful discussions. Many thanks to Simon Bork and Martin Herrenkind of IAV GmbH, Gifhorn, for array hardware construction and data recordings. The project was funded by Volkswagen AG, Wolfsburg, Germany.

References

[1] K.U. Simmer, J. Bitzer, and C. Marro, "Post-Filtering Techniques," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. 2001, Springer Verlag, pp. 40–60.

[2] R. Zelinski, "A Microphone Array with Adaptive Post-filtering for Noise Reduction in Reverberant Rooms," in *Proc. of ICASSP'88*, New York, NY, USA, Apr. 1988, pp. 2578–2581.

[3] J. Meyer and K.U. Simmer, "Multi-Channel Speech Enhancement in a Car Environment Using Wiener Filtering and Spectral Subtraction," in *Proc. of ICASSP'97*, Munich, Germany, Apr. 1997, pp. 1167–1170.

[4] I.A. McCowan and H. Bourlard, "Microphone Array Post-Filter based on Noise Field Coherence," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[5] A. Guerin, R. Le Bouquin-Jeannes, and G. Faucon, "A Two-Sensor Noise Reduction System: Applications for Hands-free Car Kit," *EURASIP Journal on Applied Signal Processing*, vol. 11, pp. 1125–1134, 2003.

[6] R.K. Cook, R.V. Waterhouse, R.D. Berendt, S. Edelman, and M.C. Thompson Jr., "Measurement of Correlation Coefficients in Reverberant Sound Fields," *Journal of the Acoustical Society of America*, vol. 27, no. 6, pp. 1072–1077, 1955.

[7] J. Bitzer and K.U. Simmer, "Superdirective Microphone Arrays," in *Microphone Arrays*, M. Brandstein and D. Ward, Eds. 2001, Springer Verlag, pp. 20–38.

[8] "ITU-T Recommendation P.56, Objective Measurement of Active Speech Level," ITU-T, Mar. 1993.

[9] "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU-T, Nov. 2005.