Head-Unit Integrated Microphone Array for Handsfree Wideband Telephony

Huajun Yu, Tim Fingscheidt
TU Braunschweig, Institut für Nachrichtentechnik (IfN), Schleinitzstr. 22, 38106 Braunschweig, Germany
Email: {yu, fingscheidt}@ifn.ing.tu-bs.de

Introduction

Recently, beamforming-based multi-channel speech enhancement techniques have drawn lots of interest for handsfree telephony in cars. With upcoming wideband telephony appropriate solutions operating at a sampling frequency of 16 kHz are required. A typical beamformer algorithm is the minimum variance distortionless response (MVDR) beamformer [1], which includes both the delay-and-sum beamformer and the superdirective beamformer. However, using just a beamformer does not provide sufficient signal-to-noise ratio (SNR) gains.

Several approaches have been proposed to cope with this problem. Zelinski has proposed in [2] to combine the delay-and-sum beamformer with a post-filter estimated under the assumption of an ideal uncorrelated noise field. However, car noise can be better modeled as a diffuse noise field [3]. With this a priori knowledge, McCowan et al. have proposed to use a superdirective beamformer with the post-filter estimated under the assumption of a diffuse noise field [4]. However, the superdirective beamformer is very sensitive to the microphone characteristics and the self-noise amplification of the uncorrelated microphone noises [1]. Since we use randomly selected low-cost microphones, a constrained superdirective beamformer with degraded performance has to be used. Another challenge is the integration of the microphone array into the head-unit. It turns out to be very cost-effective by omitting extensive wiring to some typical positions such as rear mirror or light module, while the position is acoustically sub-optimal. To deal with this problem we present a modified approach for McCowan's post-filter estimation by employing an adaptive smoothing factor for the auto- and cross-power spectral densities (psd) estimation. Furthermore, the newly modified post-filter will be combined with a very robust delay-and-sum beamformer being able to deal with the low-cost microphones.

In the sequel the baseline post-filter approach and the new modified post-filter will be described. The experimental setup and results will then be presented. Finally a conclusion is drawn.

Baseline Post-filter

Applying the short-time Fourier transform of length K, the vector of microphone array signals can then be formulated as $\mathbf{Y}'(\ell,k) = S(\ell,k)\cdot\mathbf{D}(k)+\mathbf{N}'(\ell,k)$ with frame index ℓ, frequency bin k, $\mathbf{N}'(\ell,k)$ being the additive noise and $\mathbf{D}(k)$ being the propagation vector for the delays of the desired single-channel source signal $S(\ell,k)$ for each microphone based on a reference microphone.

Due to the weak directivity of the MVDR beamformer in the low frequency region, a multi-channel Wiener filter can be used to improve the noise reduction performance for car noise with its dominating noise energy in the low frequency region. The enhanced speech signal is given as $\hat{S}(\ell,k) = H_{\mathrm{PF}}(\ell,k)\cdot\mathbf{W}^H_{\mathrm{MVDR}}(\ell,k)\cdot\mathbf{Y}'(\ell,k)$, with $\mathbf{W}_{\mathrm{MVDR}}(\ell,k)$ being the beamformer coefficients vector, $H_{\mathrm{PF}}(\ell,k)$ being the single-channel post-filter and $(\cdot)^H$ denoting the Hermitian operator, respectively. The delay-aligned signals $\mathbf{Y}(\ell,k) = \mathrm{diag}\{\mathbf{D}^*(k)\}\cdot\mathbf{Y}'(\ell,k)$ serve as inputs for the post-filter estimation with $\mathrm{diag}\{\cdot\}$ being the diag operator and $(\cdot)^*$ denoting the complex conjugate. McCowan et al. [4] have proposed to use a constrained superdirective beamformer with $H_{\mathrm{PF}}(\ell,k)$ being estimated as

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

$$H_{\mathrm{PF}}(\ell,k) = \frac{\frac{2}{M(M-1)}\cdot\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\hat{\phi}^{(ij)}_{SS}(\ell,k)}{\frac{1}{M}\sum_{i=1}^{M}\hat{\phi}_{Y_iY_i}(\ell,k)}, \quad (2)$$

where $\beta_{ij}(\ell,k) = \frac{1}{2}\left[\hat{\phi}_{Y_iY_i}(\ell,k)+\hat{\phi}_{Y_jY_j}(\ell,k)\right]$, $\mathrm{Re}\{\cdot\}$ is used to force $\hat{\phi}^{(ij)}_{SS}(\ell,k)$ to be real-valued, $\Gamma_{ij}(k)$ is the real-valued coherence function of two microphones for the diffuse noise field, $\hat{\phi}_{Y_iY_i}$ and $\hat{\phi}_{Y_iY_j}$ are the auto- and cross-psd estimated from the delay-aligned microphone signals $Y_i(\ell,k)$ by a fixed smoothing factor α.

Modified Post-filter Estimation

By using the a priori knowledge of the diffuse noise field for the estimation of McCowan's post-filter, the noise attenuation performance is improved in [4]. However, speech distortion and musical tones can be perceived using McCowan's post-filter. According to Guerin et al. [5], the smoothing factor α can be estimated as

$$\alpha(\ell,k) = \alpha_1 - \alpha_2\cdot\frac{\mathrm{SNR}(\ell,k)}{1+\mathrm{SNR}(\ell,k)}, \quad (3)$$

with $\mathrm{SNR}(\ell,k)$ being the signal-to-noise ratio at the beamformer output. For a low $\mathrm{SNR}(\ell,k)$, $\alpha(\ell,k)$ will reach its upper limit α_1, leading to a smooth estimation of the auto- and cross-power spectral densities. This limits the occurrence of musical tones [5]. When $\mathrm{SNR}(\ell,k)$ is high, $\alpha(\ell,k)$ will reach its minimum $\alpha_1-\alpha_2$, leading to a good estimation for fast speech variation. Since $\mathrm{SNR}(\ell,k)$ does not change so much frame by frame, the $\mathrm{SNR}(\ell,k)$ term in (3) can be approximated by

$$\frac{\mathrm{SNR}(\ell,k)}{1+\mathrm{SNR}(\ell,k)} \cong H_{\mathrm{PF}}(\ell-1,k), \quad (4)$$

which leads to $\alpha(\ell,k) = \alpha_1 - \alpha_2\cdot H_{\mathrm{PF}}(\ell-1,k)$.

Table 1: ΔSNR and speech component quality (PESQ-MOS) for a car in idle state (v= 0 km/h), window closed, air conditioning at 50% level

ΔSNR (dB)					
SNR_{in}	-5 dB	0 dB	5 dB	10 dB	15 dB
Zelinski	1.39	1.50	1.57	1.52	1.44
McCowan	2.32	3.13	3.37	3.21	2.78
New	7.31	8.37	8.49	8.10	7.43
PESQ-MOS					
SNR_{in}	-5 dB	0 dB	5 dB	10 dB	15 dB
Zelinski	3.70	3.80	3.89	3.96	4.00
McCowan	2.47	2.53	2.66	2.83	2.99
New	3.11	3.00	3.03	3.13	3.27

Table 2: ΔSNR and speech component quality (PESQ-MOS) for a car driven at v = 50 km/h, window closed, air conditioning at 50% level

ΔSNR (dB)					
SNR_{in}	-5 dB	0 dB	5 dB	10 dB	15 dB
Zelinski	0.89	0.99	1.06	0.99	0.84
McCowan	1.85	2.44	2.58	2.46	2.12
New	6.30	7.18	7.40	7.16	6.83
PESQ-MOS					
SNR_{in}	-5 dB	0 dB	5 dB	10 dB	15 dB
Zelinski	3.75	3.84	3.92	3.99	4.04
McCowan	2.81	2.85	2.90	3.00	3.13
New	3.42	3.36	3.38	3.42	3.51

Experimental Setup and Results

The applied microphone array in this paper consists of M = 4 microphones with 3.6 cm distance between each pair of microphones. It is located on the left side of the radio and navigation system display of an upper middle-class car, i.e., a Volkswagen Passat. The microphones have been randomly selected from a typical delivery of MCE-4500 microphones by Monacor. Multiple recordings are made separately for each channel with the clean speech signals and background noises in a synchronous manner. In our experiment two background noise conditions are investigated: (1) The car engine is running in idle state (i.e., 0 km/h), with window closed and air conditioning being set to 50% of the full level; (2) The car is driven on an expressway with a speed of 50 km/h, with window closed and air conditioning being set to 50% of the full level. An intrusive instrumental evaluation methodology based on separately and sychronously processed and recorded speech and noise signals will be used in this paper. The input signal-to-noise ratio SNR_{in} can then be set from -5 dB to 15 dB. With this methodology, it is possible to evaluate the noise reduction performance by signal-to-noise ratio improvement (ΔSNR), and the preservation of the speech component quality based on the perceptual evaluation of speech quality mean opinion score only based on the noise-free speech component (PESQ-MOS) [6]. The simulation is then performed with real acquired data for the approaches proposed by Zelinski [2], McCowan et al. [4], and the newly modified post-filter with an adaptive smoothing factor for psd estimation with a delay-and-sum beamformer.

In Table 1 and Table 2 we have shown the results for both two background noise conditions. Although the Zelinski approach has reached the best PESQ-MOS points, it provides yet hardly any noise attenuation. However, as expected, the new proposed approach with the modified post-filter estimation by using an adaptive smoothing factor has improved the PESQ-MOS at an average above 0.4 points against McCowan's post-filter, while achieving a significant improvement of several dB in ΔSNR against Zelinski's and McCowan's approach. Therefore, our new proposed modification of the post-filter estimation delivers the best results concerning both noise attenuation performance and preservation of the speech component quality. Finally, musical tones and reverberation-like effects are very much reduced by using the new approach.

Conclusion

We have briefly presented a wideband speech beamformer solution for a head-unit integrated microphone array with four low-cost microphones. Using an intrusive instrumental evaluation, we can show that the combination of the robust delay-and-sum beamformer with our newly modified post-filter estimation achieves a significantly better noise attenuation performance while still preserving the speech component quality to a large extent. Furthermore, musical tones and reverberation-like effects could be prevented with high fidelity.

References

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

[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] "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.