A COMPACT MICROPHONE ARRAY SYSTEM WITH SPATIAL POST-FILTERING FOR AUTOMOTIVE APPLICATIONS

Markus Buck, Tobias Wolff, Tim Haulick, Gerhard Schmidt

Harman/Becker Automotive Systems
Acoustic Speech Enhancement – Research
89077 Ulm, Germany

ABSTRACT

Compact microphone arrays allow for directional filtering with a minimum of installation space. They are therefore particularly suitable for automotive applications. Typically, compact arrays are realized as differential arrays or filter-and-sum beamformers which both show limited performance in terms of directivity. In this contribution we present a novel system for directional filtering for compact arrays. This system consists of two closely spaced microphones and incorporates an adaptive beamformer as well as a spatial post-filter which is designed to suppress non-stationary noise.

Index Terms— Speech enhancement, array signal processing, MAP estimation

1. INTRODUCTION

Compact microphone arrays are characterized by a microphone distance which is much smaller than the smallest acoustic wavelength that has to be processed. Since only little installation space is required, compact microphone arrays are interesting for several fields of application such as hearing aids or mobile devices. But also for automobiles the compact size is advantageous when looking for a possible position for integrating the microphones.

Most often compact microphone arrays are operated as differential arrays achieving a directivity of up to 6 dB for the two-channel case [1]. However, there is the drawback that the main direction points into the direction of the array axis (end-fire direction). Alternatively, a *minimum variance distortionless response* (MVDR) beamformer can be applied. For end-fire steering it has been shown that this method is equivalent to a differential array for low frequencies but superior for higher frequencies in terms of directivity [2]. Steering directions other than end-fire are possible, though the directivity reduces strongly in this case.

Further noise reduction can be achieved by post-filtering. Of course single-channel methods such as spectral subtraction or Wiener filtering can be applied to the spatially processed output signal. These methods, however, do not enhance the directionality since no spatial information is used within the post-filter. Multi-channel approaches exploiting the cross correlation between the microphone signals [3] are not very effective for compact arrays since the small microphone distance yields a high signal coherence even in the case of diffuse noise. Another post-filter approach [4] is based on controlling the adaptation of the noise estimate depending on the spatial information. However, the noise cannot be tracked during speech activity. In [5] it has been proposed to estimate the noise power density spectrum for a differential array by steering a null into the direction of the desired signal source. Thereby a noise power estimate can be obtained even during speech activity. This method has been generalized in [6] in the sense that the noise estimate is generated by means of a blocking matrix. An instantaneous signal-to-noise ratio is then obtained by statistical optimization.

In this paper we present a complete system that offers a high directionality with two closely spaced microphones. The system incorporates the post-filtering method of [6] and is therefore particularly suitable to suppress non-stationary noise. In Sec. 2 we give an overview of the system. The algorithmic parts are described in Sec. 3 where the most important aspects of each stage are highlighted. In the evaluation in Sec. 4 we demonstrate that the proposed system outperforms an MVDR beamformer with a Wiener post-filter.

2. SYSTEM OVERVIEW

The basic structure of the proposed system is depicted in Fig. 1. In order to reduce the computational complexity subband processing is applied. The two microphone signals are transformed into the subband domain via analysis filterbanks. Each subband signal is processed independently. After processing, the subband signals are transformed back to the time domain by a synthesis filterbank.

Since a MVDR beamformer is applied, a fixed steering direction has to be specified. A time delay compensation stage synchronizes the corresponding direct path component in both microphone channels by compensating for their relative time delays. A stand alone unit for time delay compensation has the advantage that the subsequent parts of the beam-

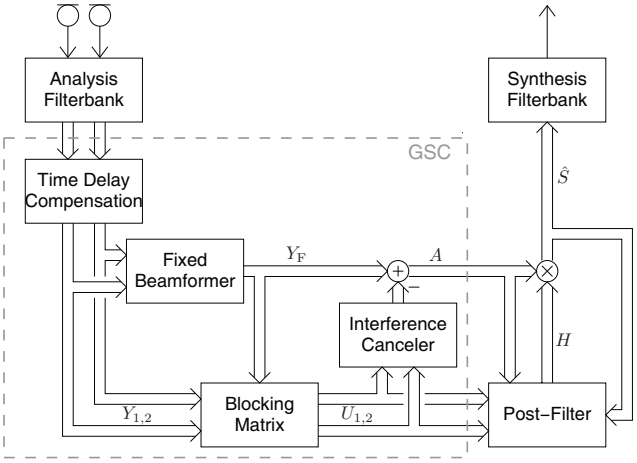

Fig. 1. Overview of the algorithmic parts.

former can be operated efficiently as a broadside beamformer. The adaptive beamformer is realized as a *generalized sidelobe canceler* (GSC) with an adaptive blocking matrix similar to [7]. The non-adaptive path of the GSC is represented by a fixed beamformer. The adaptive part of the system consists of the blocking matrix and a multi-channel interference canceler.

3. ALGORITHMIC DETAILS

3.1. Subband processing

The analog signals provided by the microphones are digitized by A/D converters with a sampling rate of f_{S}. The resulting time domain broadband signals $x_1(n)$ and $x_2(n)$ are transformed to the subband signals $X_1(e^{j\Omega_\mu}, k)$ and $X_2(e^{j\Omega_\mu}, k)$ by an analysis filterbanks. k is the time index of the subsampled subband signals, μ denotes the frequency index, and Ω_μ are the frequency supporting points.

The analysis filterbank is realized as a poly-phase filterbank. A Hann window of length N_{FFT} is used for weighting the signal segments. The weighted signal segments are transformed by a discrete Fourier transform of length N_{FFT} into the frequency domain where a subsampling by the rate R is applied.

After the core signal processing the subband signals $\hat{S}(e^{j\Omega_\mu}, k)$ are transformed back to the time domain by a poly-phase synthesis filterbank. An inverse Fourier transform is applied to the vector of subband signals at time index k. The resulting time frames are weighted again with a Hann window of length N_{FFT} and added with a frameshift of R.

3.2. Adaptive beamforming

For the present application it is advantageous to implement the adaptive beamformer in a GSC structure as depicted in Fig. 1 since the output signals of the blocking matrix are further exploited within the post-filter.

3.2.1. Time delay compensation

Due to the small distance between the microphones the relative time delay τ of the direct path signal components is quite small. To compensate for this relative time delay the subband signals are multiplied with complex-valued scalar values which shift the phases but do not change the amplitudes:

$$Y_1(e^{j\Omega_\mu}, k) = X_1(e^{j\Omega_\mu}, k) \cdot e^{j\Omega_\mu f_{\mathrm{S}} \frac{\tau}{2}} \ , \quad (1)$$

$$Y_2(e^{j\Omega_\mu}, k) = X_2(e^{j\Omega_\mu}, k) \cdot e^{-j\Omega_\mu f_{\mathrm{S}} \frac{\tau}{2}} \ . \quad (2)$$

The system is not limited to end-fire steering ($\alpha = 0^\circ$ or $\alpha = 180^\circ$). With $\tau = \frac{d}{c} \cos\alpha$ the beamformer can be steered to arbitrary angles of incidence α. d denotes the distance between the microphones and c the speed of sound.

3.2.2. Fixed beamformer

The fixed beamformer simply averages the time aligned microphone signals: $Y_{\mathrm{F}}(e^{j\Omega_\mu}, k) = \frac{1}{2}(Y_1(e^{j\Omega_\mu}, k) + Y_2(e^{j\Omega_\mu}, k))$. There is not much benefit in terms of directionality since the microphone distance d is quite small compared to the wavelengths of the sound signal. However, uncorrelated noise like self-noise of the sensors is reduced by approximately 3 dB.

3.2.3. Adaptive blocking matrix

The blocking matrix has the objective to generate noise reference signals $U_i(e^{j\Omega_\mu}, k)$ which are free of desired speech components. In the simplest case the blocking matrix can be realized by pairwise subtracting the input signals from one another. However, by applying adaptive filters the speech components can be suppressed more effectively and in a more robust manner. An adaptive blocking matrix is beneficial for several aspects: if the acoustic environment is reverberant the reverberation can be modeled by the filters. If the speaker is not exactly located in the desired direction, the filters can compensate for the mismatch. Deviations in the microphone transfer functions can also be aligned by this structure.

In this contribution an adaptive blocking matrix based on the *normalized least-mean square* (NLMS) algorithm similar to [7] is applied:

$$U_i(e^{j\Omega_\mu}, k) = Y_i(e^{j\Omega_\mu}, k) - Y_{\mathrm{F}}(e^{j\Omega_\mu}, k)\, B_i^*(e^{j\Omega_\mu}, k) \ , \quad (3)$$

$$B_i(e^{j\Omega_\mu}, k+1) = B_i(e^{j\Omega_\mu}, k) + \beta_{\mathrm{BM}}(k) \frac{Y_{\mathrm{F}}(e^{j\Omega_\mu}, k)\, U_i^*(e^{j\Omega_\mu}, k)}{\overline{|Y_{\mathrm{F}}(e^{j\Omega_\mu}, k)|^2}} \ . \quad (4)$$

The asterisk stands for conjugate complex, and the overline denotes a temporally smoothed quantity. The adaptive filters $B_i(e^{j\Omega_\mu}, k)$ are only updated during speech activity. No constraint is applied to the filter coefficients.

3.2.4. Multi-channel interference canceler

The multi-channel interference canceler aims to compensate for residual noise components in the output signal of the fixed beamformer by subtracting filtered versions of the signals $U_i(e^{j\Omega_\mu}, k)$. The filters $W_{i,l}(e^{j\Omega_\mu}, k)$ are adapted in speech pauses with the NLMS adaptation rule:

$$A(e^{j\Omega_\mu}, k) = Y_{\mathrm{F}}(e^{j\Omega_\mu}, k) - \sum_{i=1}^{2} \sum_{l=0}^{L-1} U_i(e^{j\Omega_\mu}, k)\, W_{i,l}^*(e^{j\Omega_\mu}, k) \;, \tag{5}$$

$$W_{i,l}(e^{j\Omega_\mu}, k+1) = W_{i,l}(e^{j\Omega_\mu}, k) + \beta_{\mathrm{IC}}(k) \frac{U_i(e^{j\Omega_\mu}, k-l)\, A^*(e^{j\Omega_\mu}, k)}{\sum_{n=1}^{2} \sum_{p=1}^{L-1} |U_n(e^{j\Omega_\mu}, k-p)|^2} \;. \tag{6}$$

The effectiveness of the multi-channel interference canceler depends strongly on the correlation between the residual noise component within $Y_{\mathrm{F}}(e^{j\Omega_\mu}, k)$ and the noise reference signals $U_i(e^{j\Omega_\mu}, k)$. In case of a low correlation, hardly any improvement can be expected from the multi-channel interference canceler.

3.3. Spatial post-filter

For additional noise reduction the post-filter method proposed in [6] is applied to the beamformer output signal. Particularly non-stationary noise components should be suppressed more effectively. The post-filter takes effect by dynamic spectral weighting

$$\widehat{S}(e^{j\Omega_\mu}, k) = A(e^{j\Omega_\mu}, k) \cdot H(e^{j\Omega_\mu}, k) \;. \tag{7}$$

The filter coefficients $H(e^{j\Omega_\mu}, k)$ are determined in the style of the Wiener filter:

$$H(e^{j\Omega_\mu}, k) = \max\left\{1 - \hat{\gamma}^{-1}(e^{j\Omega_\mu}, k), H_{\min}\right\} \;. \tag{8}$$

The spectral floor $H_{\min}$ determines the maximum attenuation. $\hat{\gamma}(e^{j\Omega_\mu}, k)$ is an estimate of the a posteriori *signal-to-noise ratio* (SNR) which is defined as $\gamma(e^{j\Omega_\mu}, k) = \frac{|A(e^{j\Omega_\mu}, k)|^2}{|A_{\mathrm{n}}(e^{j\Omega_\mu}, k)|^2}$. The beamformer output signal $A(e^{j\Omega_\mu}, k)$ consists of a speech component $A_{\mathrm{s}}(e^{j\Omega_\mu}, k)$ and an additive noise component $A_{\mathrm{n}}(e^{j\Omega_\mu}, k)$: $A(e^{j\Omega_\mu}, k) = A_{\mathrm{s}}(e^{j\Omega_\mu}, k) + A_{\mathrm{n}}(e^{j\Omega_\mu}, k)$. It has to be emphasized that $\hat{\gamma}(e^{j\Omega_\mu}, k)$ is an instantaneous estimate of the a posteriori SNR. While most other post-filter approaches freeze their noise estimates during speech activity the present post-filter is able to track the noise estimate and therefore the SNR at any time.

A preliminary estimate of $\gamma(e^{j\Omega_\mu}, k)$ is obtained by taking the magnitude-squared output signals of the blocking matrix $|U_i(e^{j\Omega_\mu}, k)|^2$ as estimates for the noise power and $|A(e^{j\Omega_\mu}, k)|^2$ as estimate for the signal power, respectively:

$$\tilde{\gamma}(e^{j\Omega_\mu}, k) = \frac{|A(e^{j\Omega_\mu}, k)|^2}{W_{\mathrm{eq}}(e^{j\Omega_\mu}, k) \sum_{i=1}^{2} |U_i(e^{j\Omega_\mu}, k)|^2} \;. \tag{9}$$

The equalization weights $W_{\mathrm{eq}}(e^{j\Omega_\mu}, k)$ are adjusted in speech pauses in order to get $E\{\tilde{\gamma}(e^{j\Omega_\mu}, k|_{\text{speech pause}})\} \approx 1$ for each subband. $E\{\cdot\}$ denotes the expectation operator. While the multi-channel interference canceler uses the complex values of $U_i(e^{j\Omega_\mu}, k)$ to compensate for the noise, the post-filter only makes use of their magnitudes but not of their phases.

The preliminary estimate $\tilde{\gamma}(e^{j\Omega_\mu}, k)$ can be considered as a realization of a random variable. A more adequate estimate of the a posteriori SNR is obtained by statistical optimization in the sense of the maximum a posteriori (MAP) probability. This statistical optimization is done in the logarithmic domain. With $\Gamma(e^{j\Omega_\mu}, k) = 10\log_{10}\gamma(e^{j\Omega_\mu}, k)$ and $\tilde{\Gamma}(e^{j\Omega_\mu}, k) = 10\log_{10}\tilde{\gamma}(e^{j\Omega_\mu}, k)$ the preliminary estimate can be expressed as the sum of the nominal value $\Gamma(e^{j\Omega_\mu}, k)$ and a random error $\Delta(e^{j\Omega_\mu}, k)$:

$$\tilde{\Gamma}(e^{j\Omega_\mu}, k) = \Gamma(e^{j\Omega_\mu}, k) + \Delta(e^{j\Omega_\mu}, k) \;. \tag{10}$$

In the following the frequency and time variable are omitted for a better readability. The additive error Δ can be modelled as a realization of a normal distributed random variable with variance λ_Δ and zero mean. Thus, the conditional probability density function for the observed SNR given the undisturbed SNR is:

$$f_{\tilde{\Gamma}}(\tilde{\Gamma} \mid \Gamma) = \frac{1}{\sqrt{2\pi\lambda_\Delta}} \cdot \exp\left(-\frac{(\tilde{\Gamma}-\Gamma)^2}{2\lambda_\Delta}\right) \;. \tag{11}$$

The a priori probability density function for the undisturbed SNR $f_\Gamma(\Gamma)$ can also be modeled by a normal distribution [6]. The mean value μ_Γ and the variance λ_Γ of this distribution, both, depend on the a priori SNR $\xi = \frac{E\{|A_{\mathrm{s}}|^2\}}{E\{|A_{\mathrm{n}}|^2\}}$:

$$\mu_\Gamma(\xi) = 10\log_{10}(\xi+1) \;, \tag{12}$$

$$\lambda_\Gamma(\xi) = \frac{\lambda_\Phi\, \xi}{0.5+\xi} \;. \tag{13}$$

with $\lambda_\Phi = 62.05$. Thus, the probability density function of the undisturbed SNR is

$$f_\Gamma(\Gamma) = \frac{1}{\sqrt{2\pi\lambda_\Gamma(\xi)}} \cdot \exp\left(-\frac{(\Gamma-\mu_\Gamma(\xi))^2}{2\lambda_\Gamma(\xi)}\right) \;. \tag{14}$$

Determining the SNR value which maximizes the a posteriori probability $f_\Gamma(\Gamma \mid \tilde{\Gamma}) = f_{\tilde{\Gamma}}(\tilde{\Gamma} \mid \Gamma) \cdot f_\Gamma(\Gamma)$ results in

$$\hat{\Gamma} = \frac{\lambda_\Phi\, \xi\, \tilde{\Gamma} + (\xi+0.5)\, \lambda_\Delta\, 10\log_{10}(\xi+1)}{\lambda_\Phi\, \xi + (\xi+0.5)\, \lambda_\Delta} \;. \tag{15}$$

Finally, the linearly scaled estimate of the instantaneous a posteriori SNR $\hat{\gamma}(e^{j\Omega_\mu}, k)$ results from $10\log_{10}\hat{\gamma}(e^{j\Omega_\mu}, k) = \hat{\Gamma}(e^{j\Omega_\mu}, k)$.

The a priori SNR $\xi(e^{j\Omega_\mu}, k)$ is estimated by the *decision directed approach* described in [6] and the variance of the error $\lambda_\Delta(e^{j\Omega_\mu}, k)$ is estimated in speech pauses by recursive smoothing.

4. EVALUATION

The proposed system has been evaluated in a Mercedes S-class where the array has been positioned at the overhead console. The microphone distance has been $d = 1\,\text{cm}$ and the speaker has been located at an angle $\alpha \approx 60^\circ$ relative to the end-fire direction. For the evaluation the microphone signals have been recorded during real driving situations. Noisy speech as well as pure noise signals have been recorded. In addition impulse reponses from the speaker's position to the microphones have been measured. The signals have been processed with $f_\text{S} = 11\,025\,\text{Hz}$, $N_\text{FFT} = 256$, and $H_\text{min} = 0.2$. The stepsizes of the adaptive filters are controlled based on the method proposed in [7]. The stepsizes are limited to $\max_k\{\beta_\text{BM}(k)\} = 0.01$ and $\max_k\{\beta_\text{IC}(k)\} = 0.05$. For the speech recognition tests clean speech signals from a Lombard database have been convolved with the measured impulse responses and mixed with the pure noise recordings corresponding to the specific noise conditions. The Harman/Becker continuous speech recognizer has been used for the tests. The test data base consisted of about 4000 german digits spoken in 600 utterances by 50 speakers. The speech recognizer has not been retrained for this special kind of signal processing.

While the benefit of adaptive beamformers has been examined in many other publications, this evaluation focusses on the post-filter. The reference system is an adaptive beamformer (see Sec. 3.2) followed by a single-channel Wiener filter. The proposed system incorporating the spatial post-filter is compared to this. From the results in Fig. 2 it can be seen that in terms of speech recognition the proposed system shows similar performance as the reference system for situations with rather stationary noise (*120 km/h* and *160 km/h*). For non-stationary noise situations like an opened driver window (*window*) or an interfering front passenger (*double-talk*) the proposed system clearly outperforms the reference system. The spatial post-filter is thus superior to the single channel post-filter under these conditions.

Applying the *log-spectral distance measure* (LSD) [6] indicates that this improvement is achieved without degrading the speech distortion.

5. CONCLUSION

In this contribution a practical system consisting of an MVDR beamformer and a spatial post-filter has been presented. Due to the post-filter, which is designed to suppress non-stationary noises, the proposed system achieves high directionality with only two closely spaced microphones. The system is cheap to realize and supports arbitrary steering directions. Therefore it can be integrated with very little space requirements and can easily be adjusted for use in a specific application. These features make it particularly suitable for automotive applications. Furthermore this technology is also applicable for hearing aids as well as for mobile devices.

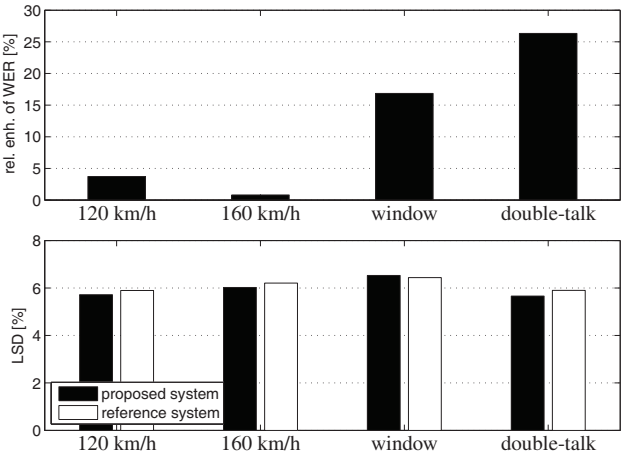

Fig. 2. Evaluation of the proposed system in different noise scenarios: 120 km/h, 160 km/h, 100 km/h with a window opened by 5 cm, and 120 km/h with an interfering speaker on the other front seat. Upper plot: relative improvement of the word error rate (WER). Lower plot: log spectral distance.

6. REFERENCES

[1] G. W. Elko: Superdirectional microphone arrays, in S. Gay, J. Benesty (Eds.): *Acoustic Signal Processing for Telecommunication*, Kluwer, Dordrecht, Netherlands, pp. 181-237, 2000.

[2] M. Buck, M. Rößler: First-order differential microphone arrays for automotive applications, *International Workshop on Acoustic Echo and Noise Control (IWAENC 01)*, Darmstadt, Germany, pp. 19-22, Sept. 2001.

[3] K. U. Simmer, J. Bitzer and C. Marro: Post-Filtering Techniques, in M. Brandstein, D. Ward (Eds.): *Microphone Arrays: signal processing techniques and applications*, Springer, New York, USA, pp. 39-60, 2001.

[4] I. Cohen, I. Gannot, B. Berdugo: An integrated real-time beamforming and postfiltering system for nonstationary noise enviroments, in *EURASIP Journal on Applied Signal Processing*, pp. 1064-1073, Nov. 2003.

[5] M. Ihle: Differential microphone arrays for spectral subtraction, *International Workshop on Acoustic Echo and Noise Control (IWAENC 03)*, Kyoto, Japan, Sept. 2003.

[6] T. Wolff, M. Buck: Spatial maximum a posteriori post-filtering for arbitrary beamforming, *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 08)*, Trento, Italy, pp. 53-56, May 2008.

[7] O. Hoshuyama, A. Sugiyama: Robust adaptive beamforming, in M. Brandstein, D. Ward (Eds.): *Microphone Arrays: signal processing techniques and applications*, Springer, New York, USA, pp. 87-106, 2001.