NOISE SUPPRESSION FOR AUTOMOTIVE APPLICATIONS BASED ON DIRECTIONAL INFORMATION

Martin Fuchs, Tim Haulick and Gerhard Schmidt*

Temic SDS, Research
Söflinger Str. 100, 89077 Ulm, Germany
gerhard.schmidt@temic-sds.com

ABSTRACT

In noise suppression systems for automotive applications the use of adaptive beamformers has proven to be of great potential. Nevertheless, in diffuse noise fields the amount of noise attenuation is rather limited and depends on the number of microphones. In order to enhance the signal-to-noise ratio further, additional classical noise suppression schemes, like spectral subtraction, are often applied. Unfortunately, these schemes tend to introduce either speech distortions or leave a large amount of residual noise. In this paper we describe a method of extracting additional spatial information from a conventional beamformer in generalized sidelobe structure. This spatial information can be utilized, e.g., to control parameters like overestimation or spectral floor of classical noise suppression schemes in a frequency selective manner or to compute a simple attenuation factor for suppressing nonstationary noise. An outlook is given on further usage of the spatial information in other algorithmic parts of a hands-free telephone or a speech recognition system.

1. INTRODUCTION

In hands-free or voice recognitions systems utilized in cars the speech signals are superimposed by background noise. Depending on the speed of the car the average signal-to-noise ratio can even be lower than 0 dB. In order to enhance the signal quality and to increase the recognition rate beamforming and noise suppression are often applied. However, in case of a diffuse noise field (which models the noise fields measured in driving cars quite well) and medium or high frequencies the signal-to-noise ratio can be enhanced only by 3 dB per doubling of the number of microphones. In order to reduce the background noise further, additional noise suppression schemes are often applied. Several approaches have been presented to combine beamforming and noise suppression [1]. Additionally, it is possible to decide whether the speaker is active or not without knowing the background noise level only by comparing two short-term powers within the generalized sidelobe structure of a beamformer.

In the following we will describe briefly the structure of a Griffith-Jim beamformer and explain how the (frequency selective) activity information can be obtained. In a very basic experiment we show how this detection result can be utilized to adjust the coefficients of a standard spectral attenuation scheme. Even if the achievable speech quality is rather low this basic scheme shows that the method works even at signal-to-noise ratios below 0 dB at the microphone. Afterwards the method is extended in order to compute a scalar attenuation factor which can be used to suppress even nonstationary noise (e.g. speech of the back seat passengers or of the co-driver). To avoid *modulation effects* of the residual background noise level comfort noise injection is also applied. An outlook is given on further applications in which this spatial speech activity detection can be utilized.

* Now with Ilmenau University of Technology, Communications Research Laboratory, P.O. Box 100565, D-98683 Ilmenau, Germany

2. BASIC IDEA

A very common way to implement an adaptive beamformer is the structure according to Griffith and Jim [3]. To reduce the computational complexity subband processing is applied. For this reason, overlapping blocks of the microphone signals are transformed periodically into the subband domain via DFT-modulated polyphase filter banks. The basic structure of the system is depicted in Fig. 1. We will denote the resulting vectors as

$$\boldsymbol{y}_i(n) = [y_{i,0}(n),\, y_{i,1}(n),\, \ldots,\, y_{i,M-1}(n)]^{\mathrm{T}}\,. \tag{1}$$

The elements $y_{i,\mu}(n)$ correspond to the subband μ of the i^{th} microphone signal at time index n. In a first stage the delays of the subband signals are compensated for a predefined direction (usually the one of the driver). In automotive applications the locations of the microphone array and the average seat positions are known a-priori, which means that no estimation of the source direction is required. The delay compensation can be achieved within the subband domain by multiplication with appropriate exponential terms:

$$\tilde{\boldsymbol{y}}_i(n) = \boldsymbol{y}_i(n) \odot \boldsymbol{\phi}_i \tag{2}$$

The sign $\odot$ denotes elementwise multiplication of two vectors. If the required delay for the i^{th} microphone is denoted by τ_i and the sampling rate by f_{s} the vectors $\boldsymbol{\phi}_i$ are defined as:

$$\boldsymbol{\phi}_i = \left[1,\, e^{-j\frac{2\pi}{M}\tau_i f_{\text{s}}},\, e^{-j\frac{2\pi}{M}\tau_i f_{\text{s}}2},\, \ldots,\, e^{-j\frac{2\pi}{M}\tau_i f_{\text{s}}(M-1)}\right]^{\mathrm{T}}. \tag{3}$$

A first signal enhancement can be achieved by summing the delay compensated microphone signals:

$$\boldsymbol{y}_{\text{ds}}(n) = \frac{1}{N}\sum_{i=0}^{N-1} \tilde{\boldsymbol{y}}_i(n)\,. \tag{4}$$

Further enhancement can be obtained by subtracting an adaptively generated noise estimate $\hat{\boldsymbol{n}}(n)$ from the delay-and-sum signal $\boldsymbol{y}_{\text{ds}}(n)$:

$$\boldsymbol{e}(n) = \boldsymbol{y}_{\text{ds}}(n) - \hat{\boldsymbol{n}}(n) \tag{5}$$

In order to apply unconstrained adaptive algorithms, so-called blocking matrices are utilized. In the simplest case of only two microphones ($N = 2$) one can simply use the difference of the delay compensated signals:

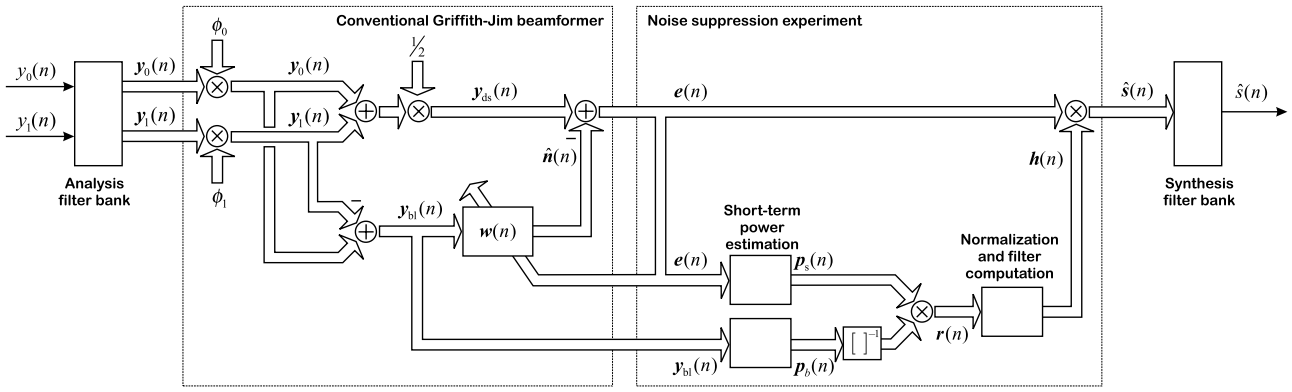

Fig. 1. Basic structure of the noise suppression system for a two-microphone setup ($N = 2$).

$$\boldsymbol{y}_{\mathrm{bl}}(n)|_{N=2} = \tilde{\boldsymbol{y}}_0(n) - \tilde{\boldsymbol{y}}_1(n)\,. \tag{6}$$

By filtering this input signal with an adaptive filter $\boldsymbol{w}(n)$ the noise estimate is generated:

$$\hat{\boldsymbol{n}}(n)|_{N=2} = \boldsymbol{w}^{\mathrm{H}}(n)\,\boldsymbol{y}_{\mathrm{bl}}(n)\,. \tag{7}$$

The filter $\boldsymbol{w}(n)$ can be adapted using one of the standard algorithms, e.g. NLMS or affine projection. If more than two microphones are used one adaptive filter is applied for each blocking signal. In order to reduce the effects of multipath propagation within the interior of the car or to deal with microphone tolerances the magnitudes of the filters are usually limited and updated only during noise only periods. Due to this limitation the directivity of the beamformer at low frequencies is rather poor. In Fig. 2 one of the microphone signals (top) and the output of the beamformer (center diagram) are depicted for a setup consisting of 4 microphones. The lowest diagram shows the result of a noise suppression experiment which will be described in Sec. 2.2.

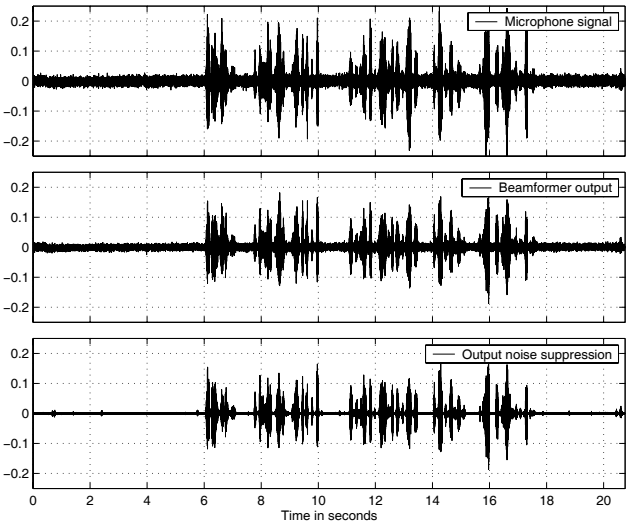

Fig. 2. Input and output signals of the beamformer and the basic noise suppression scheme.

2.1. Spatial Information

To detect speech activity from a predefined source direction short-term powers in each subband μ of the beamformer output and of the blocking signal are estimated by first-order IIR smoothing of the squared magnitudes [4]:

$$p_{\mathrm{s},\mu}(n) = \beta\, p_{\mathrm{s},\mu}(n-1) + (1-\beta)\,|e_\mu(n)|^2, \tag{8}$$

$$p_{\mathrm{b},\mu}(n) = \beta\, p_{\mathrm{b},\mu}(n-1) + (1-\beta)\,|y_{\mathrm{bl},\mu}(n)|^2. \tag{9}$$

The choice of the time constant β depends on the subsampling rate and the sampling frequency. Usually β is chosen from the interval $0.4 < \beta < 0.8$. The ratio $r_\mu(n)$ of both short-term powers is computed individually for each subband:

$$r_\mu(n) = \frac{p_{\mathrm{s},\mu}(n)}{p_{\mathrm{b},\mu}(n)}. \tag{10}$$

The resulting power ratio can have a highly varying range of values over the frequency (mainly caused by the characteristics of the blocking beamformer). To derive a common range, $r_\mu(n)$ has to be normalized by a moving average $\bar{r}_i(n)$. The normalization value is obtained by slowly increasing or decreasing the preceding value according to

$$\bar{r}_\mu(n) = \begin{cases} \bar{r}_\mu(n-1)\,\Delta_{\mathrm{inc}}, & \text{if } r_\mu(n) > \bar{r}_\mu(n-1), \\ \bar{r}_\mu(n-1)\,\Delta_{\mathrm{dec}}, & \text{else.} \end{cases} \tag{11}$$

The decreasing adjustment constant Δ_{dec} is chosen slightly smaller than 1 and the increasing constant Δ_{inc} is slightly larger than 1. Finally, the normalized power ratios can be computed as

$$\tilde{r}_\mu(n) = \frac{r_\mu(n)}{\bar{r}_\mu(n)}\,. \tag{12}$$

2.2. Very Basic Noise Suppression

Now we will introduce a very basic noise suppression scheme as a simple example of how the directional information – embedded within the normalized power ratios $\tilde{r}_\mu(n)$ – can be applied. The subband output signals of the beamformer $e_\mu(n)$ are further enhanced by applying attenuation factors $h_\mu(n)$:

$$\hat{s}_\mu(n) = e_\mu(n)\,h_\mu(n)\,. \tag{13}$$

The resulting output signals $\hat{s}_\mu(n)$ are supposed to deliver a denoised estimate of the subband speech signal $s_\mu(n)$. A very straight-forward way of adjusting the filter values can be constructed based on the directional information: When a threshold decision yields that the signal power within subband μ does not belong to the desired beamformer direction (e.g. co-driver is speaking or diffuse noise is present), the filter values are decreased by a factor $\tilde{\Delta}_{\mathrm{dec}}$. In the other case they are increased by $\tilde{\Delta}_{\mathrm{inc}}$:

$$h_\mu(n) = \begin{cases} \min\left\{h_\mu(n-1)\,\tilde{\Delta}_{\text{inc}},\ 1\right\}, & \text{if } \tilde{r}_\mu(n) > K_{\text{r}}, \\ \max\left\{h_\mu(n-1)\,\tilde{\Delta}_{\text{dec}},\ \frac{1}{4}\right\}, & \text{else.} \end{cases} \quad (14)$$

An example result for this simple noise suppression is shown in the lowest diagram of Fig. 2. The attenuation was limited to 0.25 in each subband in order to reduce artifacts during speech periods. This very basic method is able to reduce the background noise considerably, even if the quality of the resulting signal is slightly lower than with other well-established and tested frequency selective controlled noise suppression schemes (see e.g. [5, 7]). Furthermore, it is also possible to suppress non-stationary background noise (e.g. speech of the co-driver) as long as the noise source is not within the steering direction of the beamformer.

3. EXTENSIONS

The promising results of the basic experiment described in the previous section motivate a search for further applications. Furthermore, it is possible to enhance the spatial activity detection by performing the update of the normalization only during speech pauses:

$$\bar{r}_\mu(n) = \begin{cases} \bar{r}_\mu(n-1)\,\Delta(n), & \text{during speech pauses,} \\ \bar{r}_\mu(n-1) & \text{otherwise,} \end{cases} \quad (15)$$

with

$$\Delta(n) = \begin{cases} \Delta_{\text{inc}}, & r_\mu(n) > \bar{r}_\mu(n-1), \\ \Delta_{\text{dec}}, & \text{else.} \end{cases} \quad (16)$$

In this case the resulting normalized power ratios will be larger than 0 dB in case of speech activity from the steering direction, around 0 dB in noise-only periods and below 0 dB during speech or noise activity from any other than the steering direction. Fig. 3 shows an example for such a normalized power ratio for a section of speech where the co-driver and the driver speak alternately. The recording was performed within a car driving at a speed of 100 km/h. The values of $\tilde{r}_\mu(n)$ have been averaged over all subbands for better visibility before plotting. It is obvious that a distinction between speech activity of the driver, speech from the co-driver, and noise-only periods can be made from the resulting values of $\tilde{r}_\mu(n)$. Various possibilities for exploiting the directional information introduced above can be thought of. One could utilize the subband ratios $\tilde{r}_\mu(n)$ to adjust parameters – such as overestimation or spectral floor – of classical spectral substraction based noise suppression filters. Another application is the usage of so-called comfort noise injection, which will be described in the next section.

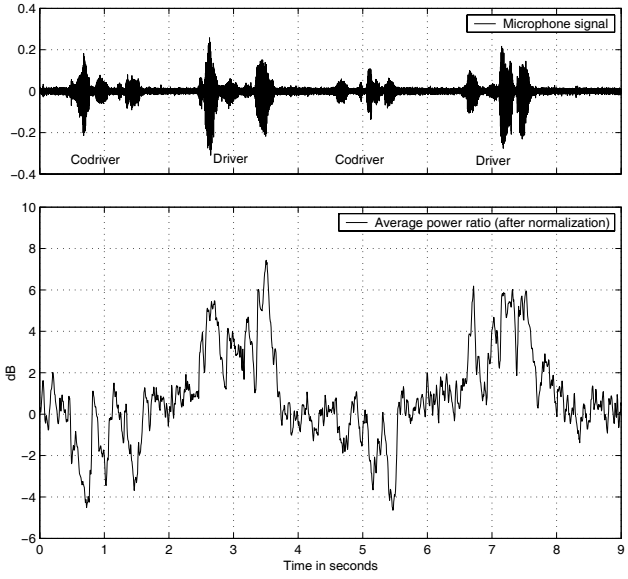

Fig. 3. Frequency-averaged normalized power ratio.

3.1. Comfort Noise Injection

In most classical noise suppression schemes an estimate of the background noise spectral density is computed. This is commonly done with the help of minimum statistics based methods [6] or by following the input spectral density in speech pauses and holding it during speech activity periods. These methods rely on the assumption that the noise spectral density is short-term stationary. Non-stationary distortions such as speech of the codriver or of the backseat passengers should be suppressed by beamforming. However, in automotive applications this works only to a certain extent. Because of multipath propagation within the car interior a reduction of only 10 to 20 dB is achievable. The residual noise is normally not suppressed by standard noise reduction schemes, resulting in tonal distortions. In order to suppress such undesired residual speech signals and also to remove instationary noise components such as the sound produced by the indicator or by the windshield wiper a so-called comfort noise injection can be utilized. If a current frame is classified as containing nonstationary distortions it is replaced by artificially generated noise:

$$\hat{s}_\mu(n) = \begin{cases} \hat{b}_\mu(n)\,h_{\mu,\min}, & \text{in case of nonstationary noise,} \\ e_\mu(n)\,h_\mu(n), & \text{else.} \end{cases} \quad (17)$$

The artificial noise signals $b_\mu(n)$ are generated such that they have the same power spectral density than the stationary background noise. The multiplication with the maximal attenuation $h_{\mu,\min}$ leads to a stationary sounding residual noise. To classify a frame as a nonstationary noise frame the normalized power ratios $\tilde{r}_\mu(n)$ are averaged over a certain frequency range, denoted by Ψ:

$$\tilde{r}(n) = \frac{1}{|\Psi|} \sum_{\mu \in \Psi} \tilde{r}_\mu(n). \quad (18)$$

This range contains medium frequencies between 500 Hz and 2000 Hz in order to detect voiced periods and high frequencies between 4000 Hz and 5000 Hz to detect unvoiced sounds such as fricatives. The quantity $|\Psi|$ is denoting the number of elements within the set Ψ. Averaging over this frequency range results in more reliable classification compared to a frequency-selective classification. A frame is classified as a noisy one whenever the current averaged ratio as well as the proceeding 100 ratios $\tilde{r}(n)$, ..., $\tilde{r}(n-100)$ are below a threshold $\tilde{r}_0 = 3$ dB. Taking the last 100 values into account can easily be realized by applying a counting mechanism. To show the performance of this simple method a simulation example is presented in Fig. 4. The upper diagram depicts a time-frequency analysis of the output signal of a beamformer with $N = 4$ microphones. The microphone signals were recorded in a car driving at a speed of 90 km/h. During the first 10 seconds the driver speaks, afterwards the codriver and than again the driver. Due to the multipath propagation within the car the beamformer is not able to reduce the signal of the codriver entirely. When applying the comfort noise injection the codriver's residual speech is removed completely (see lower diagram of Fig. 4). Due to the counting mechanism during the classification process nearly no artifacts are introduced during speech activity of the driver.

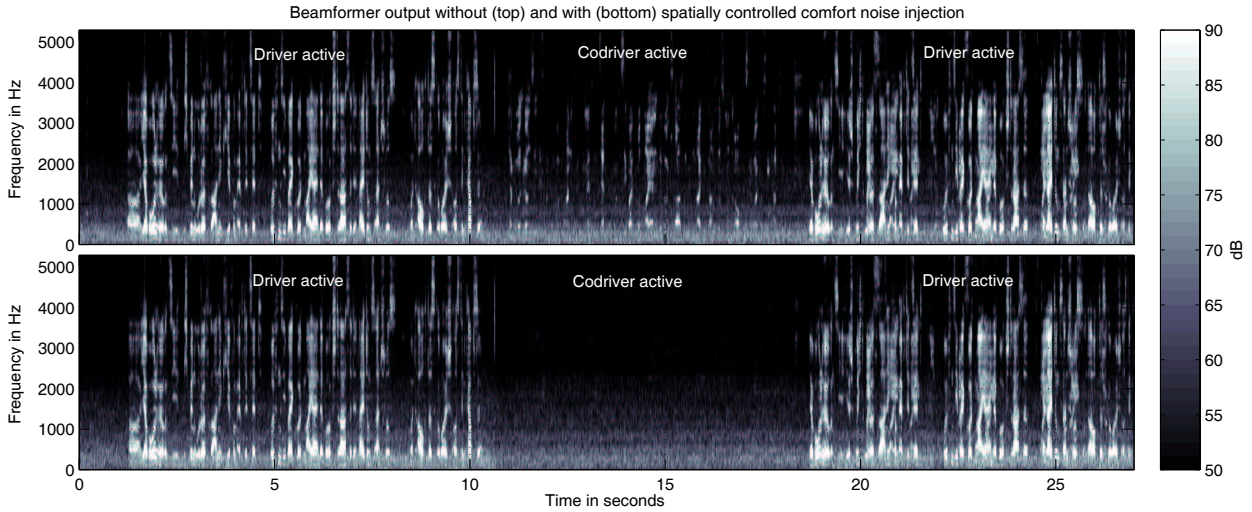

Fig. 4. Time-frequency analysis of a beamformer output signal with and without comfort noise injection.

Besides undesired speech signals it is also possible to suppress other non-stationary but directional noise components. The sounds produced by the indicator or by the windshield wiper can not be suppressed by standard noise suppression characteristics such as spectral subtraction or minimum mean-square error amplitude estimation according to Ephraim and Malah [2]. Thus, segmentation procedures – preprocessing units of speech recognizers – often misinterpret the indicator noise as speech segments and do not terminate the connection to the recognizer after the true speech period is finished. By utilizing directionally controlled comfort noise injection according to this section these problems can be avoided.

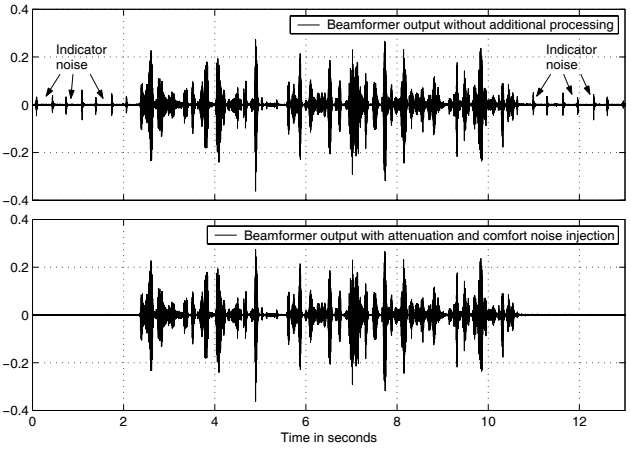

Fig. 5. Output signals of a beamformer and a succeeding noise suppression unit without (top) and with (bottom) directionally controlled comfort noise injection.

In Fig. 5 the output signals of two 4-sensor beamformers with succeeding noise suppression units – one with and one without comfort noise injection – are presented. The input signals were recorded in a car driving at 30 km/h with activated indicator. When the comfort noise injection is activated the noise peaks caused by the indicator disappear during speech pauses of the driver. It should be noted that the suppression is only possibly during speech pauses of the driver (due to the directional control). However, especially for segmentation purposes this could be a large improvement.

4. CONCLUSIONS AND OUTLOOK

In this contribution a method for extracting spatial information about the current sound field was described. The spatial information in terms of a normalized power ratio can be obtained either in a frequency selective or in a broadband manner. The latter is more reliable. One possibility (comfort noise injection) to utilize this spatial information within a single-channel noise suppression scheme has been described in detail. The very promising results motivate to investigate further application areas. The broadband ratio can be used, e.g., for controlling echo cancellation filters (stop the adaptation during speech activity of the driver).

5. REFERENCES

[1] M. Dörbecker, S. Ernst: Combination of Two-Channel Spectral Subtraction and Adpative Wiener Post-Filtering for Noise Reductions and Dereverberation, Proc. EUSIPCO '96, pp. 995-998, 1996.

[2] Y. Ephraim, D. Malah: *Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator,* IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109-1121, December 1984.

[3] L. J. Griffith, C. W. Jim: An Alternative Approach to Linearly Constrained Adaptive Beamforming, IEEE Transactions on Antennas and Propagation, vol. AP-30, no. 1, pp. 24-34, January 1982.

[4] O. Hoshuyama, et al.: A Realtime Robust Adaptive Microphone Array, Proc. ICASSP '98, pp. 3605-3608, 1998.

[5] K. Linhard, T. Haulick: Spectral Noise Subtraction with Recursive Gain Curves, Proc. ICSLP '98, pp. 1479-1482, 1998.

[6] R. Martin: An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals, Proc. EUROSPEECH '93, pp. 1093-1096, 1993.

[7] H. Puder: Single Channel Noise Reduction Using Time-Frequency Dependent Voice Activity Detection, Proc. IWAENC '99, pp. 68-71, 1999.