Implementation of a new Method for Noise Suppression in Automotive Environments

Tom Maschmann, Marco Gimm, Vasudev Kandade Rajan, Gerhard Schmidt
Christian-Albrechts-Universität zu Kiel, E-mail: [stu108805,mgi,vakr,gus]@tf.uni-kiel.de

Abstract

Talking inside a car can be difficult, due to a high amount of background noise while driving. In-car-communication (ICC) systems can be used to solve this problem. However, ICC-systems, as well as hands-free telephone systems or voice-controlled car applications only record a distorted signal, consisting of the users voice and background noise (engine, wind, tire noise, etc.), which degrades the speech intelligibility and quality. This contribution proposes a new method for noise suppression in automotive environments. Therefore, several approaches are combined using different speech- and speaker-dependent features. Furthermore, a conventional Wiener filter is extended by an adaptive spectral floor and overestimation of the background noise. The objective evaluation discloses a noise attenuation of approximately 10 dB which leads to an improvement of speech quality and intelligibility. A subjective test confirms this improvement.

Introduction

Using hands-free telephone systems, in-car-communication systems or voice-controlled car applications, can make the driving experience more comfortable and safer. All these systems are using stationary microphones, usually situated in the cars dashboard or seat belts. While driving, several noise sources, such as air condition, opened windows, tire or engine noise are recorded in addition to the speech signal. Both, the noise sources and the speech have their main power at low frequencies. This degrades the speech intelligibility and quality, especially in cases of low *signal-to-noise ratio* (SNR). Thus, reducing the noise is a challenging task for engineers [3].

The algorithm proposed in this paper combines several approaches. Speech- and speaker dependent features, such as the *pitch* frequency are used to extend the applied Wiener filter. Here, the so far constant spectral floor and overestimation of the noise are adaptively configured. During speech activity, these parameters shall decrease to less influence the speech signal. Different scenarios are used for the evaluation. Therefore, real car noise is mixed with clean speech signals in different SNRs. The evaluation is divided into an objective and a subjective part. Finally, conclusions and an outlook are drawn.

Fundamental Principles

First, fundamental principles and the mathematical theory of the used methods are explained.

Revisiting the Wiener Filter

For the following consider the recorded distorted signal $y(n)$ modeled as $y(n) = s(n) + n(n)$, where $s(n)$ is the clean speech signal and $n(n)$ the additive noise signal.

A simple and very common method to reduce additive car noise is provided by the *Wiener filter*. Its transfer function $H(e^{j\Omega})$ is chosen to be optimal in the sense of minimum mean square error (MMSE) $\overline{e^2}$, where the error signal $e(n)$ is the difference of the expected output $s(n)$ and the filter output $\hat{s}(n)$ (see Fig. 1) [2].

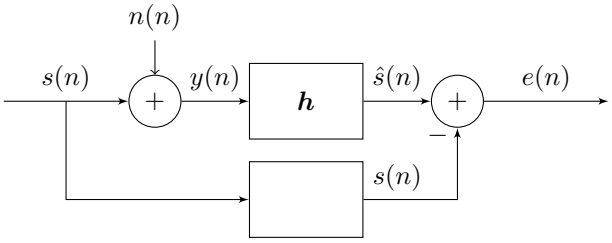

Figure 1: Wiener filter [2]

The short term Fourier transform (STFT) of the input signal $Y(\mu, k)$ is used, where μ is the subband index and k is the frame index.

Including the noise estimation $\hat{N}(\mu, k)$, the Wiener coefficients $\hat{H}_{opt}(\mu, k)$ are obtained as follows:

$$\hat{H}_{\text{opt}}(\mu, k) = \max\left\{\gamma_{\text{floor}},\, 1 - \beta_{\text{overest}} \frac{\left|\hat{N}(\mu, k)\right|^2}{\left|Y(\mu, k)\right|^2}\right\} \quad (1)$$

γ_{floor} and β_{overest} are constant parameters to reduce estimation errors. The spectral speech estimation $\hat{S}(\mu, k)$ is then given by $\hat{S}(\mu, k) = \hat{H}_{opt}(\mu, k)Y(\mu, k)$.

Smoothing

Both, the Wiener filter and the noise estimation are using a smoothed version of the distorted input signal. Smoothing can be applied along the time and/or the frequency axis and is used to reduce spectral outliers and huge bias. A first-order IIR filter is used for both smoothing possibilities, e.g along the time axis:

$$\overline{\left|X(\mu, k)\right|^2} = \beta_t \overline{\left|X(\mu, k-1)\right|^2} + (1 - \beta_t)\left|X(\mu, k)\right|^2, \quad (2)$$

where β_t indicates the smoothing parameter along the time axis and $X(\mu,k)$ is a general short term spectrum. Smoothing along the frequency axis can be applied in forward (f) and backward (b) direction:

$$\text{f: } \overline{\left|X(\mu,k)\right|^2} = \begin{cases} \left|X(\mu,k)\right|^2 & \text{, for } \mu = 0 \\ \beta_\text{f}\overline{\left|X(\mu-1,k)\right|^2} & \\ +(1-\beta_\text{f})|X(\mu,k)|^2\text{, else} & \end{cases}$$

$$\text{b: } \overline{\left|X(\mu,k)\right|^2} = \begin{cases} \left|X(\mu,k)\right|^2 & \text{, if } \mu = N-1 \\ \beta_\text{f}\overline{\left|X(\mu+1,k)\right|^2} & \\ +(1-\beta_\text{f})\left|X(\mu,k)\right|^2\text{, else.} & \end{cases} \tag{3}$$

Noise Estimation

As already mentioned, a suitbale noise estimation is required. Here, the *robust extended noise estimation* (RENE), proposed and evaluated in [4], is used. It is divided into two main steps:

1. Calculation of a probability weighting $\omega(\mu,k)$ for a speech pause.
2. Use of the probability weighting in a weighted sum of slow and fast estimators.

First, a slow changing noise estimator $\hat{N}_{det}(\mu,k-1)$ is compared to the smoothed input spectrum $\overline{|Y(\mu,k)|}$ [4].

If the smoothed input signal $\overline{|Y(\mu,k)|}$ is larger than the former noise estimation $\hat{N}_\text{det}(\mu,k-1)$, a rising noise level will be assumed and a multiplicative correction $\Delta_\text{inc,n}$ will be chosen to update the estimation. If the input signal is smaller than the estimated noise, indicating a falling noise level, Δ_{dec} will be chosen.

However, two exceptions extend this fundamental noise estimation. First, if an SNR threshold is exceeded, speech activity will be assumed and the small weighting $\Delta_\text{inc,s}$ will be selected. Secondly, if the smoothed input signal $\overline{|Y(\mu,k)|}$ is larger than the former noise estimation $\hat{N}_\text{det}(\mu,k-1)$ for a while, limited by a maximum period t_max, a constantly rising noise level will be assumed. The noise estimation shall follow this rise quickly, so a strong weighting $\Delta_\text{inc,f}$ is used.

$$\Delta(\mu,k) = \begin{cases} \Delta_\text{inc,f}\text{ , if } \hat{N}_\text{det}(\mu,k-1) < \overline{|Y(\mu,k)|} & \\ \qquad \cap \text{counter}_\text{inc}(\mu,k) > t_\text{max} & \\ \Delta_\text{inc,n}\text{ , if } \hat{N}_\text{det}(\mu,k-1) < \overline{|Y(\mu,k)|} & \\ \qquad < \text{SNR}_\text{tresh}\cdot\hat{N}_\text{det}(\mu,k-1) & \\ \Delta_\text{inc,s}\text{ , if SNR}_\text{tresh}\cdot\hat{N}_\text{det}(\mu,k-1) & \\ \qquad < \overline{|Y(\mu,k)|} & \\ \Delta_\text{dec}\text{ , else.} & \end{cases} \tag{4}$$

The noise estimation at frame index k is then modeled as $\hat{N}_\text{det}(\mu,k) = \hat{N}_\text{det}(\mu,k-1)\tilde{\Delta}(\mu,k)$, where $\tilde{\Delta}(\mu,k)$ is the weighting $\Delta(\mu,k)$ adjusted by a long-term trend [4].

The weighting $\omega(\mu,k)$ is modeled as a probability for a speech pause:

$$\omega(\mu,k) = \min\left\{\frac{\left|\hat{N}_\text{det}(\mu,k)\right|^2}{\overline{\left|Y(\mu,k)\right|^2}}, 1\right\}. \tag{5}$$

Finally, the noise estimation $\hat{N}(\mu,k)$ is a weighted sum of the smoothed input spectrum $\overline{|Y(\mu,k)|}$ and the slow estimator $\hat{N}_\text{det}(\mu,k)$ [4]:

$$\hat{N}(\mu,k) = \big(1-\omega(\mu,k)\big)\cdot\hat{N}_\text{det}(\mu,k) + \omega(\mu,k)\cdot\overline{\left|Y(\mu,k)\right|} \tag{6}$$

During speech activity, $\omega(\mu,k)$ converges to 0 and the slow estimator is used. In contrast, during speech pauses, it is assumed that the input signal simply consists of noise. Thus $\omega(\mu,k)$ becomes larger.

Pitch Frequency Estimation

During *voiced* speech, the *pitch frequency* and its harmonics are forming peaks in the spectral envelope [2]. Interferred by background noise, the pitch frequency is not necessarily clearly visible. It can be found using the *harmonic product spectrum*. Here the logarithmic form is used [1]:

$$P(\mu,k) = \sum_{m=1}^{K}\log_{10}\left(\left|X(m\mu,k)\right|^2\right). \tag{7}$$

Each sample of $P(\mu,k)$ depends of the corresponding samples of the spectrum $X(\mu,k)$ and its K harmonics. Due to the peaks in the spectral envelope of voiced speech, the pitch and its harmonics form the largest sum in the product spectrum.

Proposal

In a next step, the Wiener filter is extended. The information of the speaker's *pitch frequency* is used to avoid an unwanted signal attenuation in subbands, where speech is present. Thus, the *pitch frequency* is estimated and a multiplicative weight β_pitch is calculated using the *mel domain*. Therefore, a mel-matrix with 20 triangular filters is used.

In addition, the so far constant parameters γ_floor and β_overest are adaptively configured. During speech activity, the spectral floor and overestimation of noise shall be smaller, because the speech signal superimposes the noise level in loudness. Thus, the noisy signal can pass the filter less modified to avoid unwanted attenuation. The input spectrum $Y(\mu,k)$ is smoothed over frequency and/or over time. The time-smoothed version is only used for the pitch-estimation.

System Implementation

In the following, the implementation of the proposed new method is discussed.

Usage of the Noise Estimation

The used noise estimation RENE is optimized in two parts to improve the performance, especially in higher

subbands. The condition for the usage of the small parameter $\Delta_{inc,s}$ is now tested in the mel domain. If speech is present in a melband M_i, $\Delta_{inc,s}$ will be used for the corresponding subbands. In addition, the probability weighting $\omega(\mu,k)$ changes. Therefore, an *input-to-noise ratio* (INR) $\mathrm{INR}(\mu,k)$ of the frequency smoothed input spectrum $\overline{|Y(\mu,k)|}_f$ and $\hat{N}_{det}(\mu,k)$ is calculated. Furthermore, an upper and lower limits $\mathrm{INR}_{\mathrm{max/min}}$ are set. If more than $M_{\mathrm{mel1}} = 10$ subbands of $\mathrm{INR}(\mu,k)$ exceed the upper limit, speech will be assumed and $G(\mu,k)$ is modeled as $G(\mu,k) = (\mathrm{INR}(\mu,k))^2$. If less than $M_{\mathrm{mel2}} = 3$ subbands exceed the limit, $G(\mu,k)$ is modeled as $G(\mu,k) = \frac{\mathrm{INR}(\mu,k)}{2}$. The values of $\mathrm{INR}(\mu,k)$ are limited to the upper and lower limit. The weighting $\omega(\mu,k)$ is then calculated as follows:

$$\omega(\mu,k) = \frac{\left(\frac{1}{G(\mu,k)} - \frac{1}{\mathrm{INR}_{\mathrm{max}}}\right)}{\left(\frac{1}{\mathrm{INR}_{\mathrm{min}}} - \frac{1}{\mathrm{INR}_{\mathrm{max}}}\right)}. \tag{8}$$

A range from 0 to 1 is provided by using the limits. By squaring or halving $\mathrm{INR}(\mu,k)$, $\omega(\mu,k)$ converges faster to either 0 or 1.

Adaptive Attenuation and Overestimation

The speech activity dependent variation of the spectral floor $\gamma_{\mathrm{total}}(\mu,k)$ and the overestimation $\beta_{\mathrm{total}}(\mu,k)$ is achieved by multiplying the constant parameters γ_{floor} and β_{overest} by a parameter $\gamma(\mu,k)$, $\beta(\mu,k)$ and a trend parameter $\gamma_{\mathrm{global}}(k)$, $\beta_{\mathrm{global}}(k)$, respectively. Therefore, an INR is calculated in the mel domain. If it is above a threshold, speech is assumed and the corresponding subbands in the frequency domain of $\gamma(\mu,k)$ and $\beta(\mu,k)$ will be set to 2 dB and -1 dB, respectively. Then, the trend parameters $\gamma_{\mathrm{global}}(k)$ and $\beta_{\mathrm{global}}(k)$, indicating a trend for speech activity, are modeled as an average over frequency of $\gamma(\mu,k)$ and $\beta(\mu,k)$. Afterwards, $\gamma_{\mathrm{global}}(k)$ and $\beta_{\mathrm{global}}(k)$ are averaged from the last five values. This leads to a less attenuation and overestimation during speech activity (see Fig. 2).

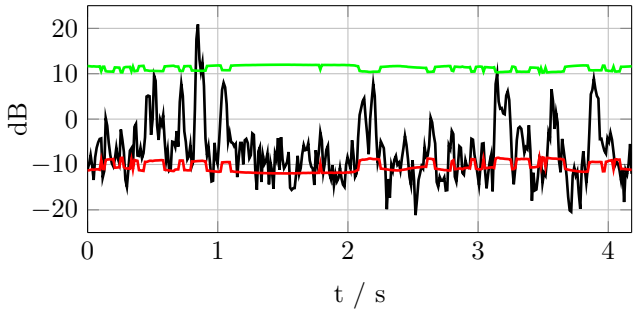

Figure 2: Adaptive attenuation *gamma*$\mathrm{total}(\mu,k)$ (red), adaptive overestimation $\beta_{\mathrm{total}}(\mu,k)$ (green) and the distorted signal (blue) at 1.1 kHz and SNR of 15 dB

The adaptive spectral floor $\gamma_{\mathrm{total}}(\mu,k)$ and overestimation $\beta_{\mathrm{total}}(\mu,k)$ are then given by:

$$\begin{aligned}\gamma_{\mathrm{total}}(\mu,k) &= \gamma_{\mathrm{floor}} \cdot \gamma(\mu,k) \cdot \overline{\gamma_{\mathrm{global}}(k)} \\ \beta_{\mathrm{total}}(\mu,k) &= \beta_{\mathrm{overest}} \cdot \beta(\mu,k) \cdot \overline{\beta_{\mathrm{global}}(k)}\end{aligned} \tag{9}$$

Pitch-dependent Weighting

A possible pitch frequency μ_{pitch} can be assumed in a range between 100 Hz - 450 Hz[1]. Here, the logarithmic harmonic product spectrum is calculated. The biggest calculated peak $\mu_{P_{\mathrm{max}}}$ of the harmonic product spectrum must fulfill two criteria to be set as μ_{pitch}:

1. The product spectrum at $\mu_{P_{\mathrm{max}}}$ must be greater than a constant threshold
2. The PSD[2] at $\mu_{P_{\mathrm{max}}}$ must be greater than a noise dependent threshold

μ_{pitch} is then used to calculate a weight $\beta_{\mathrm{pitch}}(\mu,k)$. The overestimation of noise $B_{\mathrm{overest}}(\mu,k)$ in the Wiener filter shall be smaller at the pitch frequency and its harmonics. Therefore an INR in the mel domain is taken. If the INR is larger than a threshold, speech activity will be assumed and $\beta_{\mathrm{pitch}}(\mu,k)$ will be set to 0.3 at the pitch and its harmonics in the appropriate melband M_i:

$$\beta_{\mathrm{pitch}}(\mu,k) = \begin{cases} 0.3, \text{ if } \mathrm{INR}_{\mathrm{mel},i} > \mathrm{INR}_{\mathrm{thresh,pitch}} \\ \quad \cap \mu = k \cdot \mu_{\mathrm{pitch}}\ ,\ \mu \in M_i\ ,\ k \in \mathbb{N} \\ 1.2, \text{ else.} \end{cases} \tag{10}$$

The adjacent subbands $\mu_{\mathrm{pitch}-1}$ and $\mu_{\mathrm{pitch}+1}$ are also set to 0.3 to reduce possible estimation errors. $\beta_{\mathrm{pitch}}(\mu,k)$ is set to 1.2 between the pitch and its harmonics to even increase the overestimation in subbands, where no speech is present. The overestimation is then modeled as $\tilde{\beta}_{\mathrm{total}}(\mu,k) = \beta_{\mathrm{total}}(\mu,k) \cdot \beta_{\mathrm{pitch}}(\mu,k)$.

Finally, the Wiener Filter equation reads:

$$\hat{H}_{\mathrm{opt}}(\mu,k) = \max\left\{\gamma_{\mathrm{total}}(\mu,k),\ 1 - \tilde{\beta}_{\mathrm{total}}(\mu,k)\frac{|\hat{N}(\mu,k)|^2}{|Y(\mu,k)|^2}\right\} \tag{11}$$

Evaluation

Objective Evaluation

Different scenarios, e.g. constant speed, increasing or decreasing speed, at different SNRs are used for the objective evaluation. The criterion is the dB-distance. It is specified by the difference between real and estimated noise, unfiltered and filtered distorted signal and the difference of the new implementation compared to the former noise estimation [4] and a standard Wiener filter (see Eq. 1). The dB-distance for the noise, as an example, is given by:

$$D_{\mathrm{n,dB}}(\mu,k) = \Big|20\log_{10}\big|N(\mu,k)\big| - 20\log_{10}\big|\hat{N}(\mu,k)\big|\Big|. \tag{12}$$

For the evaluation, D_{dB} is averaged over time, over frequency and over both. The new implementation shows good results in all scenarios and SNRs. Fig.3 shows the dB-distance over time $\overline{D_{\mathrm{n,dB}}(k)}$ and frequency $\overline{D_{\mathrm{n,dB}}(\mu)}$, respectively, in a 100 km/h scenario and an SNR of 10

[1] The average pitch frequency of male speaker is 120 Hz, of female speaker 215 Hz and of children 400 Hz [3]

[2] *Power spectral density*

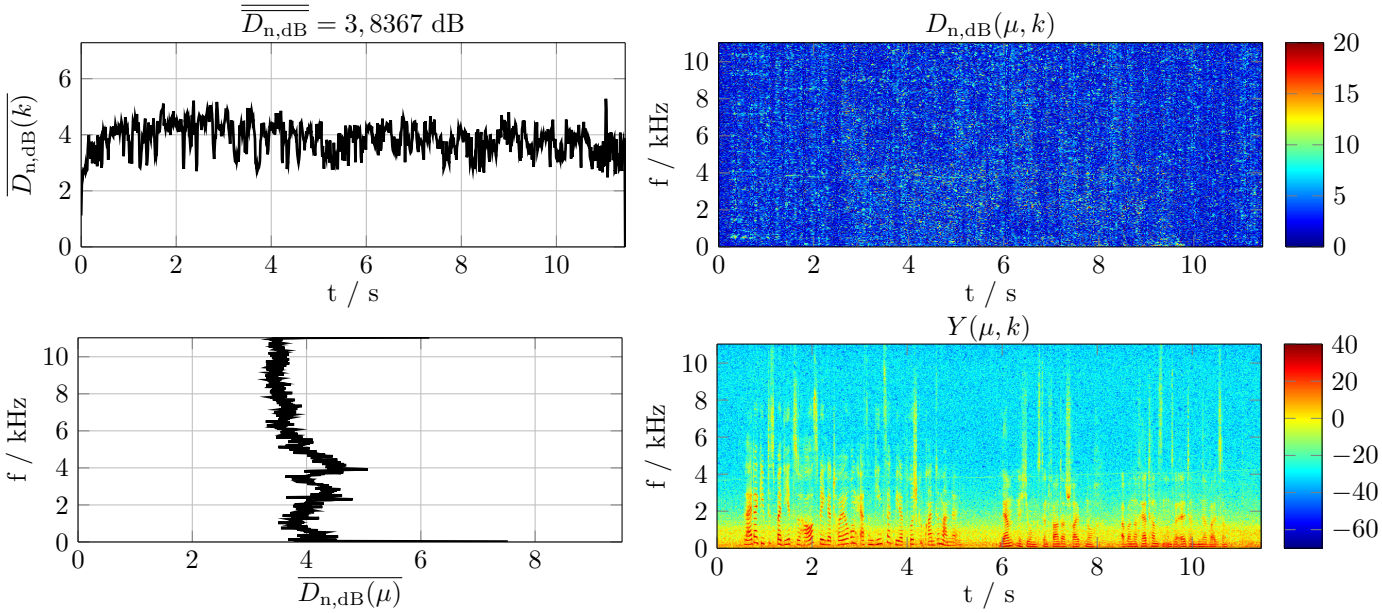

Figure 3: Estimated vs. real noise at 100km/h and a SNR of 10 dB

dB. The noise estimation error is 3.8 dB, averaged over time and frequency. A more detailed analysis can be found in [3].

Compared to the former noise estimation and a normal Wiener filter, the new implementation has better results during speech activity [3].

Subjective Evaluation

A *Comparison-Mean-Opinion-Score* test with 19 test persons is used for the subjectiv evaluation. The test is divided into two parts. First, the filtered and unfiltered signal are compared. Here, 7 possible answers can be choosen by the test person, from "A is much better than B" to "A is much worse than B".

In the second part, the new and the former method are compared through 2 possible answers, " A is better than B" or vice versa. The presented test sentences are given by the ITU[3] [5], divided into male and female speaker.

Comparing the unfiltered and filtered signal, the filtered signal is at least "slightly better", in 8 of 12 presented scenarios even "better". Also the comparison of the new and old method reveals a preference for the new implementation. In table 1, the mean value of the first test is depicted for 3 scenarios. Here, the value range is given from -3 (filtered signal is much worse) to $+3$ (filtered signal is much better).

Scenario	Mean value
Male speaker, 50km/h, SNR 10 dB	1.84
Male speaker, inc. speed, SNR 5 dB	1.73
Female speaker, inc, speed, SNR 10 dB	1.47

Table 1: Mean value of test persons depicted for 3 scenarios of 1st test

Conclusion and Outlook

The proposed method for noise suppression in automotive environments combines several approaches, e.g. the usage of the pitch-frequency or the speech dependent spectral floor and overestimation. In addition, the given noise estimation algorithm has been enhanced. The objective and subjective evaluation discloses an improvement of speech quality and intelligibility. Averaged over all scenarios and different SNR, the noise attenuation is approximately 10 dB. Thereby, the noise estimation error is relatively small. Further investigations could improve the systems parametrization and prove its behaviour in real-life situations.

[3] *International Telecommunication Union*

References

[1] L. R. Rabiner, R. W. Schafer: *Digital Processing of Speech Signals*, Prentice Hall, 1978.

[2] E. Hänsler, G. Schmidt: *Acoustic Echo and Noise Control*, Wiley, 2004.

[3] T. Maschmann: *Implementierung eines neuen Verfahrens zur Geräuschreduktion und -manipulation in Sprachsignalen*, Bachelor thesis, Kiel University, Faculty of Engineering, 2016. (in German)

[4] C. Baasch: *Verbesserung und Implementierung einer Geräuschschätzung in einem Echtzeitsystem für Anwendungen im Automobilbereich*, Bachelor thesis, Kiel University, Faculty of Engineering, 2012. (in German)

[5] International Telecommunication Union: *Appendix I to ITU-T Recommendation P.50*, 1998.