Rajan *et al. EURASIP Journal on Advances in Signal Processing* (2016) 2016:35
DOI 10.1186/s13634-016-0332-4

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Signal processing techniques for seat belt microphone arrays

Vasudev Kandade Rajan[1*], Mohamed Krini[2†], Klaus Rodemer[2] and Gerhard Schmidt[1†]

Abstract

Microphones integrated on a seat belt are an interesting alternative to conventional sensor positions used for hands-free telephony or speech dialog systems in automobile environments. In the setup presented in this contribution, the seat belt consists of three microphones which usually lay around the shoulder and chest of a sitting passenger. The main benefit of belt microphones is the small distance from the talker's mouth to the sensor. As a consequence, an improved signal quality in terms of a better signal-to-noise ratio (SNR) compared to other sensor positions, e.g., at the rear view mirror, the steering wheel, or the center console, can be achieved. However, the belt microphone arrangement varies considerably due to movements of the passenger and depends on the size of the passenger. Furthermore, additional noise sources arise for seat belt microphones: they can easily be touched, e.g., by clothes, or might be in the path of an air-stream from the automotive ventilation system. This contribution presents several robust signal enhancement algorithms designed for belt microphones in multi-seat scenarios. The belt microphone with the highest SNR (usually closest to the speaker's mouth) is selected for speech signal enhancement. Further improvements can be achieved if all belt microphone signals are combined to a single output signal. The proposed signal enhancement system for belt microphones includes a robust echo cancelation scheme, three different microphone combining approaches, a sophisticated noise estimation scheme to track stationary as well as non-stationary noise, and a speech mixer to combine the signals from each seat belt to a single channel output in a multi-seat scenario.

Keywords: Seat belt microphones, Microphone arrays, Speech enhancement, Moving microphones

1 Introduction

If speech-based services such as hands-free telephony [1, 2], in-car communication [3, 4], or voice control [5] should be used in cars, microphones that convert the acoustic signals into electric counterparts are required. In order to capture the speech signals of the passengers in an optimal way, the question on the placement of the microphones is natural. Here, several competing interests arise. Engineers who are responsible for optimizing the performance of voice control systems might favor a small distance between the microphone and the mouth of the speaker. This might lead to solutions that are not preferred by designers and final customers (being the second and third group in that process).

Currently, several positions have been found as a compromise among the three groups: automotive microphones are placed in the roof of the car (e.g., BMW), in the rear-view mirror (e.g., Daimler), in the overhead console (e.g., Audi), or on the steering wheel (e.g., Porsche) to mention just a few positions. When selecting those places, usually the expected average noise and speech levels are taken into account.

While the speech level is mainly a function of the distance between the talker's mouths and the microphones, the noise level depends on a lot of factors such as the noise distribution in the passenger compartment. In addition, it is evaluated if the position allows for simple wiring and if the microphone can be used for more than one passenger. Some systems also like to exploit spatial filtering such as beamforming [6] or diversity-based approaches [7]. Then, it must be ensured that more than one microphone can be mounted.

Recently, a new interesting microphone position and type is available (see Fig. 1): microphones that are

*Correspondence: vakr@tf.uni-kiel.de
†Equal contributors
[1]Digital Signal Processing and System Theory, University of Kiel, Kaiserstrasse 2, Kiel, Germany
Full list of author information is available at the end of the article

Fig. 1 Belt microphone (with permission from [8])

integrated into the seat belt [8]. From here on, such microphones will shortly be called as *belt microphones*. Details about belt microphones will be presented in the next chapter.

2 Belt microphones

In the following, we will view the belt microphones as an array consisting of three sensors. All of them are omnidirectional microphones, spaced 160 mm apart, fixed on one seat belt. Each microphone is approximately 10 mm in diameter, and all wiring needed for voltage supply and signal transport is weaved into the seat belts so that it appears invisible. The microphones are able to receive signals between 100 and 8500 Hz at a maximum sound pressure level of 115 dB. Figure 2 shows an example of a seat belt microphone system installed in a vehicle.

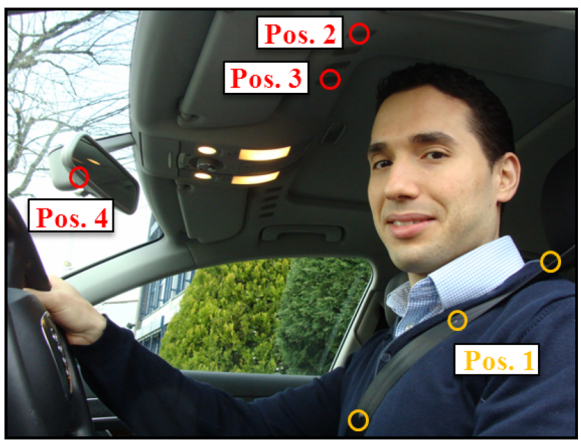

Fig. 2 Belt microphones (*Pos. 1*) and microphones positioned at the roof and at the mirror (*Pos. 2–4*)

The region of placement of these microphones is roughly between the shoulder and the center of the upper body of a sitting passenger. The exact position can vary considerably depending on the size of the passenger and also the seat position. However, due to the arrangement of the three microphones, at least one is usually close to the speaker's mouth. It is important to note that the entire geometry is likely to change due to movements of the passenger. The array can be of linear type with all microphones in one line, but could also be spread on a convex curve.

In Fig. 3, a belt microphone system is compared with three hands-free microphones placed at different positions (see Fig. 2) in terms of the average signal-to-noise ratio (SNR) for driving speeds between 120 and 160 km/h.

The distances from different microphone positions to the mouth are 20–27 cm (Pos. 1), 28 cm (Pos. 2/3), and 58 cm (Pos. 4). All microphones are calibrated to have the same speech power at standstill. This comparison shows that at higher frequencies, the behavior of all microphones is almost similar, whereas at low and medium frequencies, the belt microphone outperforms conventional hands-free microphones. An improvement of up to 6–10 dB in SNR can be achieved.

Even if belt microphones have a strong potential to improve the speech acquisition in automotive environments, we face several challenges with this microphone type. We will highlight the three major ones in the following paragraphs:

- **Continuously changing echo paths**
  An undesired characteristic of belt microphones is their changing position. Every time the driver or passenger moves his/her body, the positions of the microphones are changed. This is a recurring phenomenon during the course of normal driving. It is not an easy task for adaptive signal processing schemes such as echo cancelation filters to cope up with this movement. Every time the position is changed, the "true" frequency response is different from the estimated one which results in echo bursts. The sudden appearance of echoes can be quite unpleasant for the remote communication partner and can occur several times during a conversation. This serious restriction must be handled with robust and reliable detection and suppression schemes. It motivates the (re-) investigation of so-called room change detectors and shadow filters approaches.
- **Array processing**
  Another challenge resulting from the varying microphone positions is to process them as an array. One promising algorithm to this problem is the so-called adaptive microphone selection [9]. The

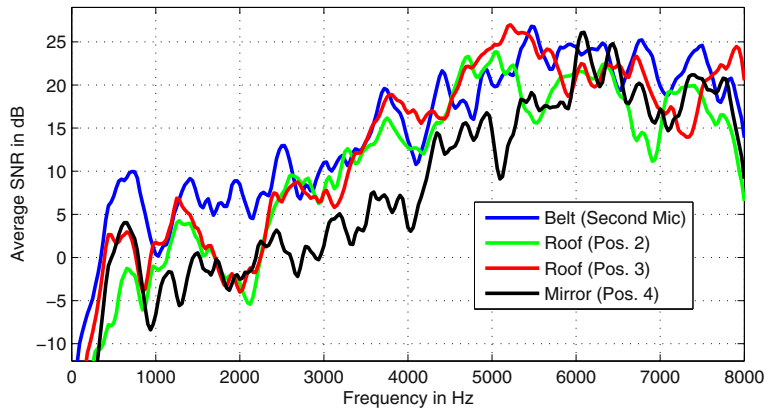

Fig. 3 SNR measured at different microphone positions

algorithm applies a sensor switching based on the corresponding SNR.
However, for optimal usage of the array structure, all microphone signals should be processed and utilized simultaneously. This motivates the investigation of robust adaptive beamforming schemes.

- **Additional noise sources**

  An additional noise source arises for seat belt microphones because they are placed directly at the passenger's body. They might be touched accidentally by hand or rubbed by clothes such as ties, zippers of a jacket. Ventilation systems can severely degrade the signal quality when the air-stream is directed towards the passenger. Therefore, more sophisticated noise estimation schemes that are also able to track non-stationary noise sources are of great interest.

3 Structure of the contribution and notation

The authors have organized this article around the signal enhancement scheme designed for belt microphones in a multi-seat scenario as shown in Fig 4. All signal processing solutions involving various tasks like echo cancelation, speaker localization, signal equalization and delay alignment, microphone combination, noise estimation, residual echo and noise suppression, and speech mixer will be described in the following sections. Section 4.1 introduces a robust echo cancelation scheme to solve the major challenge of continuously changing echo paths with belt microphones. A reliable and robust localization of the moving signal source (passenger) will be presented in Section 4.2. The equalization and alignment methods for effectively combining the belt microphone signals for each seat will be illustrated in Section 4.3. Three different methods for combining belt microphones on each seat will be presented and their performance will be compared to each other in Section 4.4. A sophisticated low computational complexity noise estimation method which is able to track stationary as well as non-stationary noise will be discussed in Section 4.5. This is followed by Section 4.6, in which the residual echo and noise suppression scheme for attenuating different kinds of interferences at the output of the combiner will be considered. Details of a speech mixer that combines the different belt microphones from various seats to a single output will be illustrated in Section 4.7. Finally, the contribution concludes with a summary and an outlook in Sections 5 and 6, respectively.

All presented algorithms designed for belt microphones operate in the short-term frequency domain [10]. Thus, the entire structure is embedded into *analysis* and *synthesis filter banks*. For some applications, e.g., in-car communication (ICC) systems, special restrictions such as a low delay have to be fulfilled by the filter banks [11, 12].

In the following, we will use a sample rate $f_s = 16$ kHz, a frame shift of $r = 128$ samples, and a fast Fourier transform (FFT) order of $N_{\text{FFT}} = 512$ where the samples are weighted with a Hann window. As an example, the output of the analysis filter bank of the belt microphone signal contains the corresponding short-term spectra of the $M = 3$ microphone signals $Y_l^{(p)}(\mu, k)$, where the index $l \in \{0, \ldots, M-1\}$ is the microphone index, $(p) \in \{0, \ldots, P-1\}$ indicates the seat index, $\mu \in \{0, 1, \ldots, N_{\text{FFT}} - 1\}$ is the subband, and k is the frame index. Since the input signals are assumed to be real, it is sufficient to store and process only the first $N_{\text{Sbb}} = N_{\text{FFT}}/2 + 1$ frequency supporting points. For better readability, the seat index (p) and the microphone index l are dropped in most of the following sections. However, when we discuss beamforming and how the

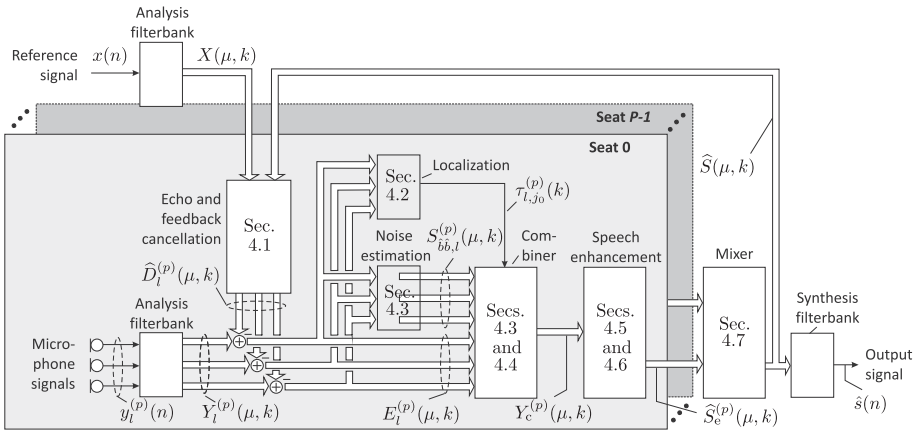

Fig. 4 Overview of the proposed signal enhancement system for processing belt microphones. The numbers in the frames refer to the related sections of this article

enhanced spectra of the individual seats can be combined, the indices will reappear.

4 Signal processing techniques for belt microphones

In the following subsections, the signal enhancement scheme designed for belt microphones in a multi-seat scenario will be described in detail.

4.1 Belt microphones used in echo path estimation

Belt microphones, like other automotive microphones, are often used for communication with a remote person. The setup is such that the remote person's voice is played back locally inside the automobile cabin which results in the well-understood problem of acoustic echoes [13]. The signals recorded by the belt microphones consist of these undesired echo components (besides the desired speech signals). As described in the previous section, the major challenge in using belt microphones for such a scenario is the continuously changing echo path. Along with the need for a robustly controlled adaptive filter, a method to handle the sudden changes in echo path is necessary. The task of the echo canceler is to produce a signal which is an estimate of the true echo signal.

4.1.1 Design of the echo canceler

The echo canceler is designed to operate in the subband domain as shown in Fig. 4. This subband domain operation offers an advantage of keeping the computational load low. In a multi-channel scenario such as here with at least three microphones per seat belt along with multiple seats, the total number of cancelation filters required are dependent on the number of loudspeakers present in the cabin as well. The number of cancelation filters required is given by number of loudspeakers $\times$ $(P \times M)$, where the loudspeakers are referred to as the reference channels and $(P \times M)$ is the total number of microphones for all seats. The echo path from each loudspeaker to a microphone is modeled as a finite impulse response (FIR) filter. The principle behind echo cancelation is first to estimate the total echoes at the microphone which is then subtracted from the microphone signal. For the sake of simplicity, a single channel setup is considered here although the method of cancelation remains the same for each filter. The belt microphone signal and its spectrum is represented by $y(n)$ and $Y(\mu,k)$, respectively. $Y(\mu,k)$ consists of the echo $D(\mu,k)$, the local speech component $S(\mu,k)$, and the background noise $B(\mu,k)$, respectively. Thus, the short-term spectrum of each belt microphone is given by

$$Y(\mu,k) = D(\mu,k) + S(\mu,k) + B(\mu,k). \tag{1}$$

The estimated echo spectrum by the echo canceler is represented by $\widehat{D}(\mu,k)$. This estimated spectrum is subtracted from the belt microphone spectrum to get the error spectrum given by

$$E(\mu,k) = Y(\mu,k) - \widehat{D}(\mu,k), \tag{2}$$

where the estimated echo is obtained by the convolution of the reference spectrum with the estimated multi-path transmission from the loudspeaker to the microphone

$$\widehat{D}(\mu,k) = \widehat{\boldsymbol{H}}^{\mathrm{H}}(\mu,k)\,\boldsymbol{X}(\mu,k), \tag{3}$$

with

$$\widehat{\boldsymbol{H}}(\mu,k) = \left[\widehat{H}(\mu,k,0),\ldots,\widehat{H}(\mu,k,L-1)\right]^{\mathrm{T}}, \tag{4}$$

$$\boldsymbol{X}(\mu,k) = \left[X(\mu,k),\ldots X(\mu,k-L+1)\right]^{\mathrm{T}}, \tag{5}$$

where L is the length of the FIR filter. The echo path is estimated in terms of its frequency response between the loudspeaker and the seat belt microphones represented by $\widehat{\boldsymbol{H}}(\mu,k)$. The coefficients of the FIR filters are updated

using the normalized least-mean square (NLMS) update rule [14]

$$\widehat{\boldsymbol{H}}(\mu,k+1)=\widehat{\boldsymbol{H}}(\mu,k)+\nu(\mu,k)\frac{\boldsymbol{X}(\mu,k)\,E^{*}(\mu,k)}{\|\boldsymbol{X}(\mu,k)\|^{2}}, \quad (6)$$

where $\nu(\mu,k)$ is the adaptive step-size control parameter. The step-size parameter in the NLMS equation controls the filter update and ranges between 0 and 1 [14]. This parameter is a critical aspect of modern day echo cancelers. It enables the system to

- Update the filters based on the reference spectral power distribution
- Inherit double-talk detection capabilities
- Protect the filters in high noise scenarios

A pseudo-optimal step-size control for the NLMS algorithm is derived in [15]. The result is given by

$$\nu_{\text{opt}}(\mu,k)=\frac{\mathrm{E}\left\{|E_{\text{u}}(\mu,k)|^{2}\right\}}{\mathrm{E}\left\{|E(\mu,k)|^{2}\right\}}, \quad (7)$$

where $E_{\text{u}}(\mu,k)$ is the undisturbed error spectrum

$$E_{\text{u}}(\mu,k)=E(\mu,k)-S(\mu,k)-B(\mu,k) \quad (8)$$

and E{...} is the expectation operator. For the application of belt microphone, the pseudo-optimal step-size is approximated by

$$\nu(\mu,k)=\frac{\overline{X}^{2}(\mu,k)\,\beta^{2}_{\text{coupl}}(\mu,k)}{\overline{E}^{2}(\mu,k)}, \quad (9)$$

where $\beta_{\text{coupl}}(\mu,k)$ is referred to as the coupling between the reference spectrum and the error spectrum. The magnitude spectra of the reference signal $X(\mu,k)$ and the error signal $E(\mu,k)$ are smoothed by first-order infinite impulse response (IIR) filtering:

$$\overline{E}(\mu,k) = \beta_0\,|E(\mu,k)|+(1-\beta_0)\overline{E}(\mu,k-1), \quad (10)$$

$$\overline{X}(\mu,k) = \beta_0\left|X\left(\mu,k-\delta_{\text{delay}}(k)\right)\right| +(1-\beta_0)\,\overline{X}(\mu,k-1). \quad (11)$$

In Eq. (11), the variable $\delta_{\text{delay}}(k)$ captures the delay of the impulse response between the loudspeaker and the belt microphone. This variable helps in choosing the value of the input along time that has the largest contribution to the echo. This delay is computed by averaging the largest value per subband in the estimated frequency response matrix. The coupling factors have two roles to play:

- To ensure the tracking of the ratio of the squared magnitudes of the reference and the error signal
- To indicate the instantaneous coupling between the two quantities

It is desired that the filters converge to the true frequency response as fast as possible. Given this, the coupling factors are computed and adjusted based on multiplicative constants. These time constants are responsible for the speed and accuracy trade-off of the tracking according to

$$\beta_{\text{coupl}}(\mu,k)=\begin{cases}\tilde{\beta}_{\text{coupl}}(\mu,k-1)\,\Delta_{\beta,\text{inc}},\\ \text{if } \overline{X}(\mu,k)\,\beta_{\text{coupl}}(\mu,k-1)<\overline{E}(\mu,k),\\ \tilde{\beta}_{\text{coupl}}(\mu,k-1)\,\Delta_{\beta,\text{dec}},\\ \text{else},\end{cases} \quad (12)$$

where $\Delta_{\beta,\text{inc}}$ and $\Delta_{\beta,\text{dec}}$ are the increment and decrement time constants. The definition of $\tilde{\beta}_{\text{coupl}}(\mu,k)$ will be given in the next section. After sufficient excitation time, the filters converge to the desired frequency response.

4.1.2 Accommodating the altering position of belt microphones

The constantly altering position of the belt microphones results in an incorrectly estimated frequency response $\widehat{\boldsymbol{H}}(\mu,k)$ as compared to the true frequency response. The coupling factors absorb the change to a certain extent. For example the minor movement of the microphone caused due to breathing, slight movement of body, etc. is accounted by the multiplicative constants. Nevertheless, depending on the chosen time constants ($\Delta_{\beta\ \text{inc}}$, $\Delta_{\beta\ \text{dec}}$ in Eq. (12)), echo leakage can occur during the time of re-adaptation. In cases where the shift is more significant, it can freeze the system. This particular problem has been addressed by several authors earlier [16–18]. The problem of changes in the true echo paths causes the estimated echo path to be different from the current echo path. The altering position of the belt microphones is modeled here as a change in the echo path. A change in the system distance can be caused by other factors like a adjusting the volume of the playback of the reference signal, or delay change of the entire system could also lead to an increase in the system distance. Such behaviors are called "echo path change" or "room change" in the literature [19, 20]. Several algorithms are suggested for detecting echo path changes mostly in combination with double-talk detection methods [21, 22]. The approach proposed in [22] is based on correlation techniques in the time-domain, whereas in [21], a subband-based solution with two filters is presented. One filter is responsible for the single-talk echo cancelation and the second for the double-talk and echo path change detection.

4.1.3 Coupling trigger to handle room changes

After room changes, the re-adaptation of the estimated filter coefficients seems to be an appropriate action to converge to the new frequency response. The solution

presented here is integrated into the adaptive step-size control approach presented in the previous section. The coupling factors, which adjust the filter update according to Eq. (9), are triggered. The factors responsible for the computation of the step-size are:

- The short-term spectrum of undisturbed error
- The short-term spectrum of the (measurable) disturbed error

The average short-term magnitude of the undisturbed error spectrum $E_{\mathrm{u}}(\mu,k)$ is computed by multiplying the smoothed magnitude of the reference spectrum with the coupling factors. This can be seen as the estimated ratio of the squared magnitudes of the reference signal and the error signal.

When the filters reach a certain convergence, the resulting step-size will be small due to converged values of the coupling factor and the error signal. At this stage, a change in the computed step-size can only be achieved by a change in the coupling factors. This will cause the computed step-size to be large (close to one) since the high value of the coupling factor will ensure that the numerator is greater than the denominator (see Eq. (9). The consistently large step-size will ensure that the filters converge to the new frequency response. The coupling factors need a certain time to reach the new convergence levels. During this time, as during normal operation of the system, the residual echo power is computed according to

$$\mathrm{E}\left\{|E_{\mathrm{u}}(\mu,k)|^2\right\} \approx \overline{X}^2(\mu,k)\,\beta^2_{\mathrm{coupl}}(\mu,k). \tag{13}$$

The computed residual echo power is suppressed by the postfilter as described in Section 4.6. By triggering the coupling factors, echo artifacts caused, e.g., by movements of the belts, can be handled now. It is now clear that a forced change in the coupling factor can handle the room change which gives the filters and the NLMS algorithm time to adapt to the new frequency response without *freezing* the system. The question now is to detect such events. During remote-side single talk, the short-term power of the error spectrum $E(\mu,k)$ will suddenly increase when the belt microphones alter their positions. This occurs as the estimated filter coefficients are incorrect and the convolution with the reference signal does not lead to the amount of echo that is actually present in the belt microphone signal.

An ideal solution for the recovering from the room change is to have a parallel set of filters that contain the coefficients of the new frequency response. Since this is hard to achieve, filters that indicate in this direction are useful. Such schemes can be realized with a second set of filters, in parallel to the main filters, referred to as *shadow filters* [23, 24]. The shadow filters are updated with the same NLMS update rule as for the main filters, but the step-size parameter is always set to 1 whenever there is activity in the reference signal. This ensures that the shadow filters converge very quickly to the true frequency response but with the problem of very quick divergence.

To reduce the overall computational load of the system, shadow filters are not placed in parallel to every subband but only a few chosen subbands. The index μ_{sh} refers to the shadow filter subbands which is $\mu_{\mathrm{sh}} \subset \mu$. Since the shadow filters are always updated with step-size 1, they adapt much faster than the main filters. During the time of change in the position of the belt microphones, the power of the error signal produced by the shadow filters is much lower than the power of the error of the main filters. It is exactly with this error power difference that a room change can be detected:

$$\tilde{\beta}_{\mathrm{coupl}}(\mu,k) = \begin{cases} \frac{\overline{E}(\mu,k)}{\overline{X}(\mu,k)}, \\ \quad \text{if } \overline{E}_{\mathrm{main}}(k) > T_{\mathrm{change}}\,\overline{E}_{\mathrm{sh}}(k), \\ \beta_{\mathrm{coupl}}(\mu,k), \\ \quad \text{else}, \end{cases} \tag{14}$$

with

$$\overline{E}_{\mathrm{sh}}(k) = \sum_{\mu_{\mathrm{sh}}=0}^{N_{\mathrm{Sbb,sh}}-1} |E_{\mathrm{sh}}(\mu_{\mathrm{sh}},k)|^2, \tag{15}$$

$$\overline{E}_{\mathrm{main}}(k) = \sum_{\mu_{\mathrm{sh}}=0}^{N_{\mathrm{Sbb,sh}}-1} |E(\mu_{\mathrm{sh}},k)|^2. \tag{16}$$

$N_{\mathrm{Sbb,sh}}$ is the total number of subbands for which shadow filters have been placed. $E_{\mathrm{sh}}(\mu_{\mathrm{sh}},k)$ is the error signal obtained from the shadow filters similar to the main filters. During normal updates, the shadow filters will diverge because of a lack of optimum control. During this time, the error power of the main filters is lower than the error power of the shadow filters and is detected according to

$$\overline{E}_{\mathrm{sh}}(k) > T_{\mathrm{div}}\,\overline{E}_{\mathrm{main}}(k), \tag{17}$$

where T_{div} is the threshold at which the divergence of the shadow filters is detected. When this occurs, the coefficients from the main filter are copied to the shadow filters.

4.1.4 Results

The echo canceler presented before has been tested in various scenarios. The test setup consisted of one reference signal originating from the phone which is played back via the loudspeakers in the car with engines turned on. This signal is picked up as an echo by the belt microphones. The echo canceler filter length was set to about 100 ms. The entire echo canceler was tuned for performance in

terms of smoothing constants, multiplicative constants of the coupling factors, etc.

The first scenario for testing the echo canceler is a regular speech activity from both the remote side and the local side in the following order: a remote-side single talk, a local-side single-talk, and a double-talk. The test determines the amount of echo suppression by the echo canceler alone without the postfilter applied to suppress the residual echo. The plots in Fig. 5 compare the microphone signal spectrogram of the first belt microphone versus the corresponding error signal. The time spans of each activity situation are shown encircled along with the echo canceled regions. A cancelation of between −20 and −30 dB is achieved during the remote-side single-talk situation. During the local-side-only speech-activity, the speech is completely retained as seen clearly in the spectrograms. During the double-talk situation, one of the important factors is that the echo cancelation filters do not diverge. Also, when postfilter is applied, the conversation must be as transparent as possible. Transparent conversation here means the retention of the local speech components against the suppression of the echo components. An overview of the results of some important tests performed under the ITU tests [25] is shown through a quality pie in Fig. 6. All tests except the distortion RCV and distortion SND belong to the echo canceler. The green color indicates that the ITU tests have passed, while the yellow area indicates that the test has failed. The red area indicates the area to be covered in order to pass the test. The most important tests are the double talk-tests (preceding with DT-) which have all passed with the highest class of class 1. More details about the test can be found in the ITU recommendation document [25].

The second test scenario focuses on the room-change detection. For this test, the microphone signal was simulated in such a way that after about 9.5 s time, the impulse response is changed to another impulse response, both belonging to typical belt microphone positions. Figure 7 shows the microphone signal and the point at which the room change was applied. The second plot shows the error signal in which the echo reappears after the room change. The third plot is the trigger signal which indicates that the room change condition is met according to Eq. (14). Finally, the reaction of the coupling factor $\beta_{\text{coupl}}(\mu, k)$ is seen which is reset to about 0 dB after each trigger. The coupling factor is plotted for subband $\mu = 29$, but the trigger is applied to all subbands. The tunable threshold parameter T_{change} for this scenario was set at 25 dB. This parameter will determine the reaction time of the trigger to the room change. During the time after the room change and before the coupling trigger, there will be echo blips which would be suppressed by the postfilter through the residual echo. The power of the echo blips is dependent mainly on the distance between the adapted frequency response and the changed frequency response. During subjective tests, it has been seen that these blips mostly go unnoticed by the remote listener. During initial adaptation and re-adaptation after a room change, there will be many re-triggers because the filters are still converging to the changed frequency response. This is also seen in the third and the fourth plots where after the first trigger, there are four follow-up triggers which again reset the coupling factors. This results in a slightly higher residual echo power. This can be improved by averaging the trigger indicator over time, holding the room-change detection for a while after the first trigger.

4.2 Localization

Based on the array geometry of the three microphones on each seat belt, the microphone signals can be combined with a beamformer as described in Section 4.4. Because

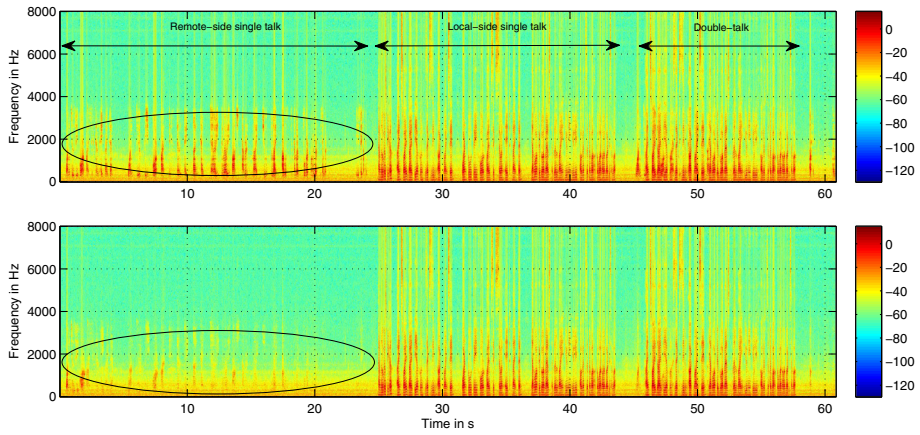

Fig. 5 Spectrogram comparison of the first microphone of the belt (*top*) with the echo canceled error signal (*bottom*)

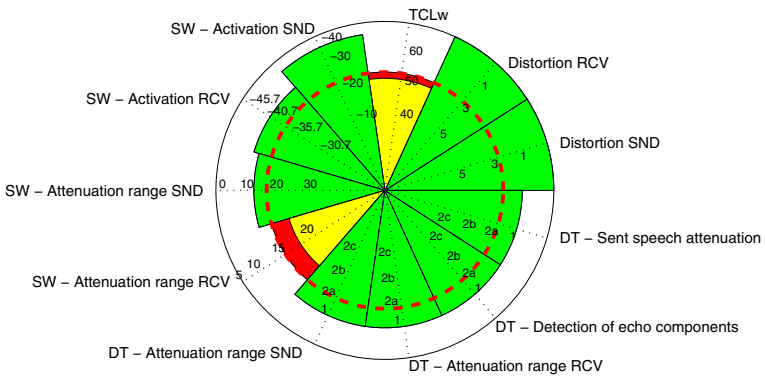

Fig. 6 Results overview for ITU tests

of the highly varying geometry of the array and changing position of the signal source, it is nearly impossible to generate any a priori knowledge about the direction of the beamformer. Therefore, a reliable and robust localization of the signal source has to be applied. To estimate the delays $\tau_{l,j}(k)$ in samples between the microphones and thereby the direction of the beamformer, a *generalized cross correlation* (GCC) function is utilized as presented in [26]. Here, a pairwise instantaneous *cross power spectral density* (CPSD) $\tilde{S}_{e_l,e_j}(\mu,k) = E_l(\mu,k)\,E_j^*(\mu,k)$ between the error spectra is determined for $l \neq j$. The CPSD is smoothed to avoid the jumps and variations seen on the instantaneous spectra through a first-order IIR filter given by[1]

$$\overline{S}_{e_l,e_j}(\mu,k) = \alpha_{\mathrm{E}}\,\tilde{S}_{e_l,e_j}(\mu,k) + (1-\alpha_{\mathrm{E}})\,\overline{S}_{e_l,e_j}(\mu,k-1), \tag{18}$$

where the smoothing constant α_{E} is chosen to be around 0.8 (240 dB/s). The delay is computed by the argument that maximizes the inverse Fourier transform of that quantity. Before transforming it, a weighting can be applied. We utilized here the so-called phase transformation (PHAT) [27] leading to the following normalized cross correlation:

$$s_{e_l,e_j}(\kappa,k) = \frac{1}{N_{\mathrm{FFT}}} \sum_{\mu=0}^{N_{\mathrm{FFT}}-1} \frac{\overline{S}_{e_l,e_j}(\mu,k)}{\left|\overline{S}_{e_l,e_j}(\mu,k)\right|}\, e^{j\frac{2\pi\mu}{N_{\mathrm{FFT}}}\kappa}. \tag{19}$$

The time-domain transformed samples are indexed by κ. Optionally, a lower FFT size can be used to reduce the computational complexity by dropping bins above, e.g., 3500 Hz. The robustness can be improved by setting the lowest bins to zero before applying the transform. The maximum distance between the belt microphones is 320 mm, and hence, the delay is limited to this distance given by $\tau_{\max}$. From the time-domain transformed frame,

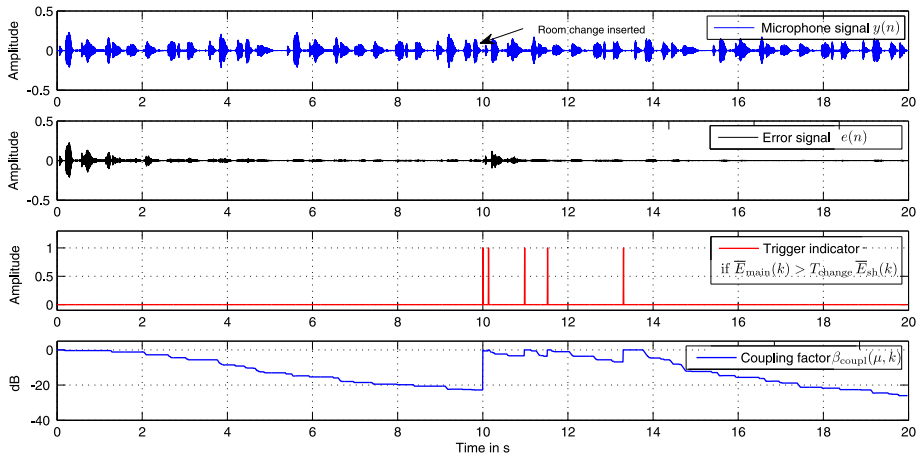

Fig. 7 Plots showing triggering of the coupling factor (shown for subband around 1 kHz) due to a room change detected by the trigger indicator. The room change was inserted after about 9.5 s as shown in the microphone signal

the delay $\tau_{l,j}(k)$ is computed by finding the argument κ that maximizes the cross correlation function $s_{e_l,e_j}(\kappa,k)$:

$$\tau_{l,j}(k) = \underset{-\tau_{\max}<\kappa<\tau_{\max}}{\operatorname{argmax}} \left\{ s_{e_l,e_j}(\kappa,k) \right\}. \tag{20}$$

4.3 Signal equalization and delay alignment

The nature of the belt microphones is such that they usually pickup slightly varied ambient noise even if they are in the same environment. To achieve good combination performance, it is important to correct this by equalizing the noise for all the microphones. This is achieved by using a simple multiplicative constant based on the noise PSD estimation for each microphone given by

$$\widehat{B}_l(\mu,k) = \begin{cases} \delta_{\text{inc}}\, \widehat{B}_l(\mu,k-1), \\ \qquad \text{if } \overline{E}_l(\mu,k) > \widehat{B}_l(\mu,k-1), \\ \delta_{\text{dec}}\, \widehat{B}_l(\mu,k-1), \\ \qquad \text{else}, \end{cases} \tag{21}$$

where $\widehat{B}_l(\mu,k)$ is the estimated magnitude spectrum of background noise for each microphone, δ_{inc} is the incremental constant, and δ_{dec} is the decremental constant, with $0 \ll \delta_{\text{dec}} < 1 < \delta_{\text{inc}}$. $\overline{E}_l(\mu,k)$ is a smoothed version of the magnitude of the spectrum $E_l(\mu,k)$ as opposed to a complex smoothed spectra obtained similarly as shown in Eq. (18). A slowly varying equalization factor $K_l(\mu,k)$ per microphone is computed and tracked based on the average background noise $\widehat{B}_{\text{avg}}(\mu,k) = 1/M \sum_{l=0}^{M-1} \widehat{B}_l(\mu,k)$ of all the three microphones given by

$$K_l(\mu,k) = \begin{cases} \delta_{\text{gain-inc}}\, K_l(\mu,k-1), \\ \qquad \text{if } \widehat{B}_{\text{avg}}(\mu,k) > \widehat{B}_l(\mu,k), \\ \delta_{\text{gain-dec}}\, K_l(\mu,k-1), \\ \qquad \text{else}. \end{cases} \tag{22}$$

The equalization factor, which is bounded by a maximum and a minimum value for safety reasons, is applied to the error spectra along with the estimated delay to obtain the pre-processed spectra on which the beamforming technique is applied. This is performed by

$$\widetilde{E}_l(\mu,k) = K_l(\mu,k)\, E_l(\mu,k)\, e^{-j\frac{2\pi\mu}{N_{\text{FFT}}}\tau_{l,j_0}(k)}. \tag{23}$$

Usually, the center microphone is used as a (delay) reference, meaning that we use $j_0 = 1$ in Eq. (23).

4.4 Combining belt microphones

The microphone combination computes one signal for each passenger from a subset of all microphones. From the arrangement of $M = 3$ microphones positioned on a seat belt, that microphone can be selected which has the best overall signal quality in terms of high SNR. Further improvements can be achieved if all microphone signals are combined to a single output signal. In the following subsections, three different combining methods are presented and compared to each other in terms of SNR and signal-to-interference ratio (SIR).

4.4.1 Max-SNR approach

A straightforward and robust method is to use only the microphone with the best signal quality that is measured based on instantaneous SNR:

$$\Gamma_l(\mu,k) = \frac{|E_l(\mu,k)|^2}{\left|\widehat{B}_l(\mu,k)\right|^2}. \tag{24}$$

This measure is smoothed over time and also decreased if the microphone signal suffers frequently from degradations by instationary distortions. The smoothed SNR $\overline{\Gamma}_l(\mu,k)$ is only updated during local speech activity of pth passenger and while no activity on the reference channel and from the neighboring speaker's is detected. The signal combination is done in two stages:

1. Select the microphone $i \in \{0, \ldots, M-1\}$ with the best quality measure:

$$\overline{\Gamma}_i(k) = \frac{1}{N_{\text{FFT}}} \sum_{\mu=0}^{N_{\text{FFT}}-1} \overline{\Gamma}_i(\mu,k). \tag{25}$$

   In order to avoid frequent switching in case the measures are close together, a hysteresis is introduced.
2. If instationary distortions have been detected in the microphone that is currently selected, the disturbed frequency bins are replaced by those of signal with the next best quality. In case that all signals are distorted, comfort noise is injected.

The estimates of the time delays between microphones can also be exploited and combined with the smoothed SNR for enhanced microphone selection. Details can be found, e.g., in [28].

4.4.2 SNR-based weighting

For combining the microphone signals to one output signal, a modified filter and sum beamformer is used with an SNR-based signal weighting. In the literature, e.g., in [7, 29], SNR-based beamforming is widely presented. The SNR-based weighting beamformer is a modified filter-and-sum beamformer which means that each input to the beamformer will be filtered and the sum of all the filtered inputs forms the output of the beamformer. In the context of this paper, the inputs to the beamformer are the three equalized and delay-aligned belt microphone spectra. This is shown in Eq. (26)

$$Y_{\text{fb}}(\mu,k) = \sum_{l=0}^{M-1} G_l(\mu,k)\, \widetilde{E}_l(\mu,k). \tag{26}$$

The filter weights $G_l(\mu,k)$ are a function of the normalized SNR computed per subband for the respective microphone. The SNR per subband $\Gamma_l(\mu,k)$ is computed according to Eq. (24). The normalized SNR is computed by dividing the subband SNR by the sum of all the subband SNRs of three microphones given by

$$\widetilde{\Gamma}_l(\mu,k) = \frac{\Gamma_l(\mu,k)}{\sum_{j=0}^{M-1} \Gamma_j(\mu,k)}. \tag{27}$$

Since the short-term SNR of the individual microphone signals is highly varying, the filter function is computed as a smoothed version of the normalized SNRs. In addition, the filter function should be updated only during speech activity. The smoothing is again performed by an IIR filter with the smoothing constant that is switched between a constant and 0 to ensure that the previous values are kept during non-speech frames. This is captured in Eqs. (28) and (29):

$$G_l(\mu,k) = [1-\alpha_l(k)]\; G_l(\mu,k-1) + \alpha_l(k)\,\widetilde{\Gamma}_l(\mu,k), \tag{28}$$

where

$$\alpha_l(k) = \begin{cases} \alpha_{\mathrm{SNR}}, & \text{if } \frac{1}{N_{\mathrm{FFT}}} \sum_{\mu=0}^{N_{\mathrm{FFT}}-1} \overline{\Gamma}_l(\mu,k) > T_{\mathrm{SNR}}, \\ 0, & \text{else}, \end{cases} \tag{29}$$

where α_{SNR} is the smoothing constant and T_{SNR} is the SNR threshold parameter for voice activity detection.

4.4.3 Adaptive beamformer

In the following, we will describe details about all components that are necessary to perform a robust beamforming approach with belt microphones. Before proceeding with combining the three microphone spectra using an adaptive beamformer based on the *generalized sidelobe canceler* [30], they are pre-processed with two blocks, namely the *delay alignment* and *equalization*. In addition, a *localization* as shown in Fig. 8 is computed. The delay alignment is performed in order to compensate the elapsed time between the mouth of the talking passenger and the individual microphones. Localization and equalization are realized as described in Sections 4.2 and 4.3.

The SNR-weighted beamformer presented in the previous section referred to as the *first beamformer* (see Fig. 8) is used as a precursor to the adaptive blocking matrix and the interference canceler presented in the following section.

Adaptive blocking matrix: The adaptive blocking matrix (ABM) generates a noise reference for the interference canceler (IC) as shown in Fig. 8. A fixed blocking matrix [31], which subtracts adjacent equalized and time-aligned microphone subband signals, is not suitable for belt microphones due to the strong microphone SNR variations. The ABM subtracts adaptively filtered versions of the first beamformer output $Y_{\mathrm{fb}}(\mu,k)$ from each channel input $\widetilde{E}_l(\mu,k)$ and provides the noise reference signals $U_l(\mu,k)$ for the IC with $l \in \{0,\ldots,M-1\}$. The SNR differences between belt microphones and the mismatch of the steering direction can be compensated. Filters of the blocking matrix are adapted using the NLMS algorithm:

$$\boldsymbol{V}_l(\mu,k+1) = \boldsymbol{V}_l(\mu,k) + \beta_{\mathrm{bm}}(\mu,k)\,\frac{\boldsymbol{Y}_{\mathrm{fb}}(\mu,k)\,U_l^*(\mu,k)}{\|\boldsymbol{Y}_{\mathrm{fb}}(\mu,k)\|^2}, \tag{30}$$

where

$$\boldsymbol{V}_l(\mu,k) = [V_l(\mu,k,0),\ldots,V_l(\mu,k,N_{\mathrm{bm}}-1)]^{\mathrm{T}} \tag{31}$$

denote the subband filter coefficients and N_{bm} is the filter length. The vector

$$\boldsymbol{Y}_{\mathrm{fb}}(\mu,k) = [Y_{\mathrm{fb}}(\mu,k),\ldots,Y_{\mathrm{fb}}(\mu,k-N_{\mathrm{bm}}+1)]^{\mathrm{T}} \tag{32}$$

comprises the current and the last $N_{\mathrm{bm}}-1$ subband outputs of the first beamformer. The filters of the blocking matrix are adapted only if speech is picked up from the steering direction. For improved robustness, the filter coefficients can be limited (in terms of their magnitudes) by an upper and lower threshold [32, 33]. The step-size $\beta_{\mathrm{bm}}(\mu,k)$ is used to control the speed of the adaptation in every subband.

Interference canceler: The subband signals $U_l(\mu,k)$ are passed to the IC which adaptively removes the signal components that are correlated to the interference input signals from the beamformer output $Y_{\mathrm{fb}}(\mu,k)$. The adaptive filters

$$\boldsymbol{W}_l(\mu,k) = [W_l(\mu,k,0),\ldots,W_l(\mu,k,N_{\mathrm{ic}}-1)]^{\mathrm{T}} \tag{33}$$

of the IC are not adapted if speech is coming from the steering direction to avoid signal cancelation. N_{ic} denotes the filter length. For filter adaptation, again the NLMS algorithm is used:

$$\boldsymbol{W}_l(\mu,k+1) = \boldsymbol{W}_l(\mu,k) + \beta_{\mathrm{ic}}(\mu,k)\,\frac{\boldsymbol{U}_l(\mu,k)\,Y_{\mathrm{c}}^*(\mu,k)}{\sum_{l=0}^{M-1}\|\boldsymbol{U}_l(\mu,k)\|^2}. \tag{34}$$

The vector $\boldsymbol{U}_l(\mu,k) = [U_l(\mu,k),\ldots,U_l(\mu,k-N_{\mathrm{ic}}+1)]^{\mathrm{T}}$ comprises the last $N_{\mathrm{ic}}-1$ output signals of the ABM. The adaptive beamformer output is determined by

$$Y_{\mathrm{c}}(\mu,k) = Y_{\mathrm{fb}}(\mu,k) - \sum_{l=0}^{M-1} \boldsymbol{U}_l^{\mathrm{H}}(\mu,k)\,\boldsymbol{W}_l(\mu,k). \tag{35}$$

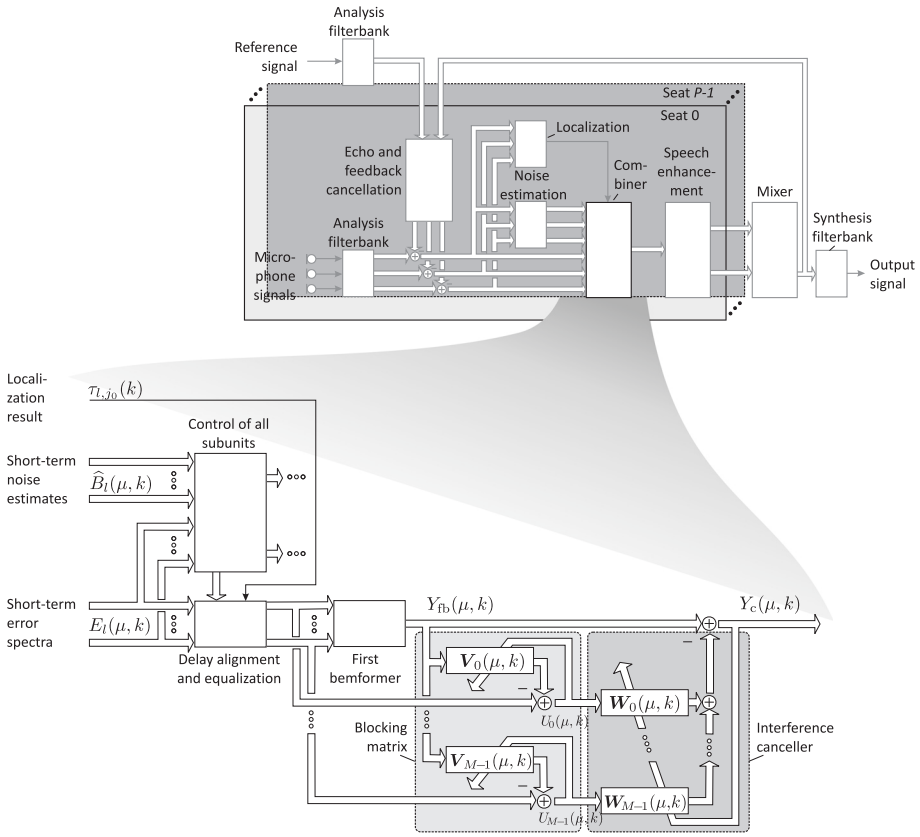

Fig. 8 Overview of the proposed adaptive beamformer for belt microphones and its integration in the entire processing structure

In order to increase the robustness of the beamformer, the norm of the adaptive filter coefficients can be limited [32, 33]. The control of the step-size $\beta_{\mathrm{ic}}(\mu,k)$ is described in the following section.

Adaptation control: The step-sizes for the ABM and the IC are controlled based on the speech activity estimation from the steering direction. As a measure for the speech activity, a ratio of the smoothed short-term powers $S_{y_{\mathrm{fb}}y_{\mathrm{fb}}}(\mu,k)$ and $S_{uu}(\mu,k)$ of the first beamformer and of the ABM output, respectively, averaged over a certain frequency range, is used:

$$r_{\mathrm{SD}}(k) = \frac{\sum_{\mu=N_{\mathrm{u}}}^{N_{\mathrm{o}}} S_{y_{\mathrm{fb}}y_{\mathrm{fb}}}(\mu,k)}{\sum_{\mu=N_{\mathrm{u}}}^{N_{\mathrm{o}}} S_{uu}(\mu,k)}\,. \tag{36}$$

The short-term powers are smoothed through a first-order IIR filter according to:

$$S_{y_{\mathrm{fb}}y_{\mathrm{fb}}}(\mu,k) = (1-\alpha)\, S_{y_{\mathrm{fb}}y_{\mathrm{fb}}}(\mu,k-1) + \alpha\ |Y_{\mathrm{fb}}(\mu,k)|^2 \tag{37}$$

and

$$S_{uu}(\mu,k) = (1-\alpha)\, S_{uu}(\mu,k-1) + \alpha\, \beta(\mu,k) \sum_{l=0}^{M-1} |U_l(\mu,k)|^2\,. \tag{38}$$

The smoothing constant is chosen around $\alpha = 750$ dB/s, and the lower and upper frequencies used in Eq. (36) were set by $\Omega_{N_{\mathrm{u}}} = 1$ kHz and $\Omega_{N_{\mathrm{o}}} = 6$ kHz. $\beta(\mu,k)$ is controlled such that in periods of stationary background noise, the ratio $r_{\mathrm{SD}}(k)$ becomes one. Only high values of $r_{\mathrm{SD}}(k)$ indicate signal energy from the steering direction. Thus, the filters of the ABM are adjusted only when $r_{\mathrm{SD}}(k)$ exceeds a predetermined threshold $t_{\mathrm{bm}} = 0.3$ using:

$$\beta_{\mathrm{bm}}(\mu,k) = \begin{cases} \beta_{\mathrm{bm}}^{(\max)}, & \text{if } S_{b_{\mathrm{fb}}b_{\mathrm{fb}}}(\mu,k)\,K < \left|Y_{\mathrm{fb}}(\mu,k)\right|^2 \\ & \quad \wedge\, r_{\mathrm{sd}}(k) \geq t_{\mathrm{bm}}\,, \\ 0, & \text{else}\,, \end{cases} \tag{39}$$

where $S_{b_{\mathrm{fb}}b_{\mathrm{fb}}}(\mu,k)$ denotes the estimated PSD of the noise at the first beamformer output and K is set to 6 dB. The

adaptive filters of the IC are controlled using:

$$\beta_{\mathrm{ic}}(k) = \begin{cases} \beta_{\mathrm{ic}}^{(\max)}, & \text{if } r_{\mathrm{sd}}(k) < t_{\mathrm{ic}}, \\ 0, & \text{else}, \end{cases} \tag{40}$$

with $t_{\mathrm{ic}} = 0.2$. The maximum step-sizes can be set to $\beta_{\mathrm{ic}}^{(\max)} = 0.1$ and $\beta_{\mathrm{bm}}^{(\max)} = 0.2$.

4.4.4 Results

For evaluating the performance of the adaptive beamformer, real-world recordings have been made at a speed of 120 km/h using the microphones on the driver's seat belt. The belt position was set to a worst-case position for the beamformer. Thus, the distance of the three microphones of the belt was highly different from each other, and therefore, the single microphone SNRs are varying significantly.

A frequency-selective SNR and SIR analysis has been made to compare the non-adaptive and the adaptive beamformers with the best single microphone. For the analysis, the speech signal with a duration of about 120 s recorded in a car at 120 km/h has been used. No remote speech was considered for this test scenario. During the first 60 s, the passenger sitting beside the drive was speaking (interference speech). Afterward, the driver was active also for about 60 s (desired speech). The number of filter taps for the interference canceler and for the adaptive blocking matrix were chosen as $N_{\mathrm{ic}} = 3$ and $N_{\mathrm{bm}} = 3$. The maximum step-sizes for the adaptive filters were set as indicated in the last paragraph. For a simple analysis, two belt microphones with high SNRs have been used to reduce computational complexity. The results in terms of SNR can be seen in Fig. 9. The SNR performance of the first beamformer is slightly better than that of the single best microphones (on average about 2 dB). The overall SNR performance can be further increased by about 5 dB when using the proposed adaptive beamformer compared to the best belt microphone. The SIR analysis as shown in Fig. 10 indicates that on average, about 2 dB SIR improvement with the first beamformer and 6 dB with the adaptive beamformer can be achieved. The proposed adaptive beamformer is suitable for highly suboptimal array geometries and shows robust performance when the signal source position is changing fast. The SIR can be further enhanced if a spatial postfilter is applied as postprocessor for adaptive beamforming.

4.5 Belt microphones in speech enhancement: low complexity noise estimation

Another common problem faced in the automobile environment is the presence of highly varying background noise. With increasing number of microphones like in the case of belt microphones, a reliable and robust noise estimation scheme that requires a low computational complexity is essential. A straightforward solution to estimate the noise accurately is to track the segments of the beamformer output spectrum that do not contain speech. Naturally, the behavior of this spectrum is dependent on the nature of noise present in the given environment which can be classified as non-stationarity in many cases in automobiles. Generally for such environments, the noise spectrum can be described as non-flat with a low-pass characteristic dominated below 500 Hz. Apart from this low-pass characteristic, changes in speed, opening and closing of windows, passing cars, etc. cause the noise floor to vary with time. A close look at one frequency bin of the noise spectrum reveals the following properties:

1. Instantaneous power can vary a large extent from the mean power even during steady conditions.
2. A steady increase or a steady decrease of power is observed during certain situations (e.g., during acceleration).

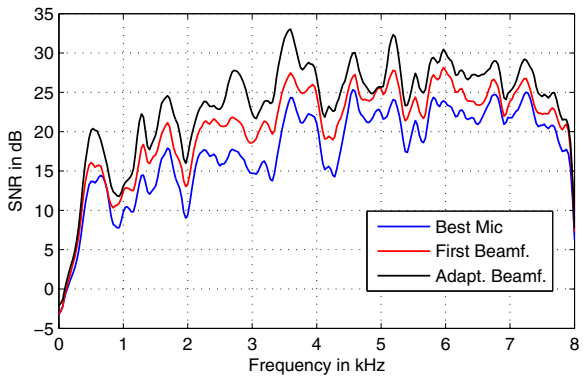

Fig. 9 SNR comparison between the best belt microphone and the beamformer outputs

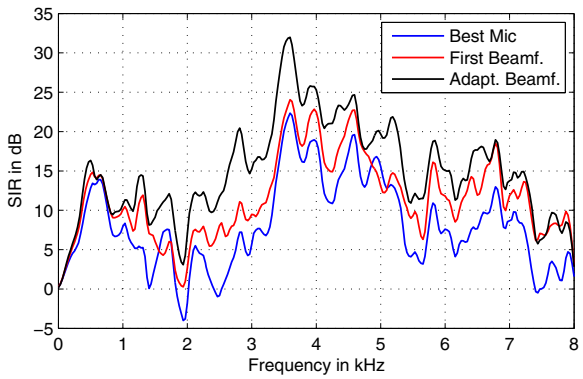

Fig. 10 SIR comparison between the best belt microphone and the beamformer outputs

The signal model considered here is an additive noise which is described by

$$Y_{\mathrm{c}}(\mu,k) = S(\mu,k) + B_{\mathrm{c}}(\mu,k), \tag{41}$$

where $S(\mu,k)$ is the local speech in the echo canceled and beamformed spectrum and $B_{\mathrm{c}}(\mu,k)$ is the local background noise. Consider a simple estimator used to track the change in magnitude per subband

$$\widehat{B}_{\mathrm{c}}(\mu,k) = \begin{cases} \widehat{B}_{\mathrm{c}}(\mu,k-1)\,\Delta_{\mathrm{inc}}, & \\ & \text{if } \overline{Y_{\mathrm{c}}}(\mu,k) > \widehat{B}_{\mathrm{c}}(\mu,k-1), \\ \widehat{B}_{\mathrm{c}}(\mu,k-1)\,\Delta_{\mathrm{dec}}, & \\ & \text{else}, \end{cases} \tag{42}$$

where $\widehat{B}_{c}(\mu,k)$ is the estimated background noise magnitude spectrum. This estimator follows a smoothed input $\overline{Y_{c}}(\mu,k)$ based on the previous noise estimate. The speed at which it tracks the noise floor is controlled by the increment constant Δ_{inc} and the decrement constant Δ_{dec}. The advantage of this algorithm is its low computational complexity. With careful tuning of increment and decrement constants combined with a highly smoothed input, an estimate of the background noise can be obtained. However, this estimator would fail for the following reasons

- Low time-constants will lag in tracking the noise power
- High time-constants will estimate speech as noise

4.5.1 Idea of the proposed noise estimator

Using the simple estimator in Eq. (42) as the basis, an improved noise estimation algorithm is proposed that tries to find a balance by keeping the computational complexity low and offering fast and accurate tracking. By recursive averaging of the estimated background noise in combination with the smoothed input spectrum, the noise estimate is obtained by

$$\begin{aligned} \widehat{B}_{f}(\mu,k) \;=\; & W_{\widehat{B}}(\mu,k)\;\overline{Y_{c}}(\mu,k) \\ & + \left(1 - W_{\widehat{B}}(\mu,k)\right)\widehat{B}_{\mathrm{pre}}(\mu,k), \end{aligned} \tag{43}$$

where the time-varying parameters $W_{\widehat{B}}(\mu,k)$ along with $\Delta_{\mathrm{final}}(\mu,k)$ (applied to estimate $\widehat{B}_{\mathrm{pre}}(\mu,k)$) control the estimation. $\widehat{B}_{\mathrm{pre}}(\mu,k)$ is a slow varying noise estimation used similar to the basic noise estimation signal. The principle behind the new estimator is to choose the most suitable multiplicative constant in a given specific situation through the $\Delta_{\mathrm{final}}(\mu,k)$ parameter. Common situations are the presence of speech, a consistent background noise, increasing background noise, decreasing background noise, etc. A measure called *trend* is computed which indicates if the long-term direction of the input signal is going up or down. Details are described in the following paragraphs. The incremental and decremental time-constants along with the trend are finally applied together in Eq. (51).

4.5.2 Smoothing the input spectrum

The tracking of the noise estimator is dependent on the smoothed input signal $\overline{Y_{c}}(\mu,k)$. The input spectrum is smoothed using a first-order IIR filter

$$\overline{Y_{c}}(\mu,k) = \gamma_{\mathrm{smth}}\;|Y_{\mathrm{c}}(\mu,k)| + (1-\gamma_{\mathrm{smth}})\;\overline{Y_{\mathrm{c}}}(\mu,k-1), \tag{44}$$

where γ_{smth} is the smoothing constant. The smoothing constant must be chosen in such a way that it retains fine variations of the input spectrum as well as eliminate the high variation of the instantaneous spectrum. A value of 300 dB/s is chosen here.[2] Optionally, additional frequency-domain smoothing can be applied.

4.5.3 Trend: long-term activity measurement

One of the difficulties for noise estimators in non-stationary environments is differentiating between a speech part and an actual change in the noise level. This problem can be partially overcome by measuring the duration for a power increase, i.e., the difference between the estimated background noise level and the instantaneous power. If the increase is due to a speech source, then the power difference will drop down after the utterance of a syllable, whereas the power difference continues to stay high for a longer duration. This can be utilized as an indication of an increased background noise. By using these power differences, a trend measure is computed by the proposed noise estimation algorithm. By observing the direction of the trend, the noise floor changes can be tracked by avoiding track speech-like parts of the spectrum. The decision about the current state of the frame is made by comparing if the estimated noise of the previous frame is smaller than the smoothed input spectrum of the current frame, and a set of values are obtained. A positive value indicates that the direction is going up, and a negative value indicates that the direction is going down

$$A_{\mathrm{curr}}(\mu,k) = \begin{cases} A_{\mathrm{up}}, & \text{if } \overline{Y_{\mathrm{c}}}(\mu,k) > \widehat{B}_{\mathrm{c}}(\mu,k-1), \\ A_{\mathrm{down}}, & \text{else}, \end{cases} \tag{45}$$

where $\widehat{B}_{\mathrm{c}}(\mu,k-1)$ is the estimated noise of the previous frame. The values $A_{\mathrm{up}} = 1$ and $A_{\mathrm{down}} = -4$ are chosen empirically. The trend is smoothed along both the time and the frequency axis. A zero-phase forward-backward filter is used for smoothing along the frequency axis. Smoothing along the frequency ensures that isolated peaks caused by non-speech-like activities are suppressed.

Smoothing is applied by using

$$\overline{A_{\text{tr}}}(\mu,k) = \gamma_{\text{tr-fq}}\, A_{\text{curr}}(\mu,k) + (1-\gamma_{\text{tr-fq}})\overline{A_{\text{tr}}}(\mu-1,k), \tag{46}$$

for $\mu = 1,\ldots,N_{\text{Sbb}}$ and similarly backward smoothing is applied. Both frequency smoothing constants $\gamma_{\text{tr-fq}}$ are chosen to be at about 35 dB/Hz. This is temporally smoothed to obtain the time-smoothed trend factor $\overline{\overline{A_{\text{tr}}}}(\mu,k)$ by an IIR filter

$$\overline{\overline{A_{\text{tr}}}}(\mu,k) = \gamma_{\text{tr-tm}}\, \overline{A_{\text{tr}}}(\mu,k) + (1-\gamma_{\text{tr-tm}})\, \overline{\overline{A_{\text{tr}}}}(\mu,k-1), \tag{47}$$

where $\gamma_{\text{tr-tm}}$ is the smoothing constant chosen to be at about 15 dB/s. The behavior of the double-smoothed trend factor $\overline{\overline{A_{\text{tr}}}}(\mu,k)$ can be summarized as follows. The trend factor is a long-term indicator of the power level of the input spectrum. During speech parts, the trend factor temporarily goes up but comes down quickly. When the true background noise increases, then the trend goes up and stays there until the noise estimate catches up. A similar behavior is seen for a decreasing background noise power. This trend measure is used to further *push* the noise estimate in the desired direction. The trend is compared to an upward threshold and a downward threshold. When either of these thresholds are reached, then the respective time-constant to be used later is chosen by

$$\Delta_{\text{tr}}(\mu,k) = \begin{cases} \Delta_{\text{tr-up}}, \\ \quad \text{if } \overline{\overline{A_{\text{tr}}}}(\mu,k) > T_{\text{tr-up}}, \\ \Delta_{\text{tr-down}}, \\ \quad \text{else if } \overline{\overline{A_{\text{tr}}}}(\mu,k) < T_{\text{tr-down}}, \\ 1, \\ \quad \text{else.} \end{cases} \tag{48}$$

The values of $\Delta_{\text{tr-up}}$ and $\Delta_{\text{tr-down}}$ are chosen to be at 20 and -20 dB/s. The trend multiplicative $\Delta_{\text{tr}}(\mu,k)$ is used later in Eq. (50) to obtain the final multiplicative constant.

4.5.4 Tracking constants based on activity detection

The tracking of the noise estimation has to be performed for two cases:

- When the smoothed input is greater than the estimated noise
- When the smoothed input is smaller than the estimated noise

Incrementing the noise estimate The short-term input spectrum can be greater than the estimated noise due to three reasons:

- When there is speech activity
- When the previous noise estimate has dipped too low and has to rise up
- When there is a continuous increase in the true background noise

The first case is handled by checking if the level of $\overline{Y_{\text{c}}}(\mu,k)$ is greater than a certain SNR threshold T_{snr}, in which case the chosen incremental constant Δ_{speech} has to be very slow because speech should not be tracked. For the second case, the incremental constant is set to Δ_{noise} which means that this is a case of normal rise and fall during tracking. For the case of a continuous increase in the true background noise, the estimate must catch up with it as fast as possible. For this, a counter $k_{\text{cnt}}(\mu,k)$ is utilized. The variable counts the duration for which the input spectrum has stayed above the estimated noise. If this counter reaches a threshold $K_{\text{inc-max}}$, then the $\Delta_{\text{inc-fast}}$ is chosen. The counter is incremented by 1 every time $\overline{Y_{\text{c}}}(\mu,k)$ is greater than $\widehat{B}_{\text{c}}(\mu,k-1)$ and reset to 0 otherwise. Equation (49) captures these conditions

$$\Delta_{\text{inc}}(\mu,k) = \begin{cases} \Delta_{\text{inc-fast}}, & \text{if } k_{\text{cnt}}(\mu,k) > K_{\text{inc-max}}, \\ \Delta_{\text{speech}}, & \text{else if } \overline{Y_{\text{c}}}(\mu,k) > \widehat{B}_{\text{c}}(\mu,k-1)\, T_{\text{snr}}, \\ \Delta_{\text{noise}}, & \text{else.} \end{cases} \tag{49}$$

The value for the fast increment $\Delta_{\text{inc-fast}}$ is chosen to be at about 40 dB/s. For the speech case, Δ_{speech} has to very slow and is chosen to be at about 0.5 dB/s (where T_{snr} is the SNR threshold for speech presence), and finally, the Δ_{noise} is chosen to be at about 6 dB/s.

Decrementing the noise estimate The choice of a decrementing constant does not have to be as explicit as the incrementing case. This is because of lesser ambiguity when $\overline{Y_{\text{c}}}(\mu,k)$ is smaller than $\widehat{B}_{\text{c}}(\mu,k-1)$ as a decrease in power usually settles down to the background noise power level. Here, the noise estimator chooses a decremental constant Δ_{dec} by default. The value for falling edge is chosen to be at about -20 dB/s. For a subband μ, only one of the above two stated conditions is chosen. From either of the two conditions, a final multiplicative constant is determined

$$\Delta_{\text{final}}(\mu,k) = \begin{cases} \Delta_{\text{inc}}(\mu,k), & \text{if } \overline{Y_{\text{c}}}(\mu,k) > \widehat{B}_{\text{c}}(\mu,k-1), \\ \Delta_{\text{dec}}, & \text{else.} \end{cases} \tag{50}$$

4.5.5 Combining all detection schemes

The input spectrum consists of only background noise when no speech-like activity is present. During this time, the best estimate is to set the noise estimate equal to the input spectrum. When the estimated noise is lower than

the input spectrum, the noise estimate and the input spectrum are combined with a certain weight. The weights are computed according to Eq. (52). A pre-estimate $\widehat{B}_{\text{pre}}(\mu,k)$ is obtained for computing the weights. The pre-estimate is used in combination with the input spectrum. It is obtained by multiplying the estimate from the previous frame with the multiplicative constant $\Delta_{\text{final}}(\mu,k)$ and the trend constant $\Delta_{\text{tr}}(\mu,k)$

$$\widehat{B}_{\text{pre}}(\mu,k) = \Delta_{\text{final}}(\mu,k)\ \Delta_{\text{tr}}(\mu,k)\ \widehat{B}_{\text{pre}}(\mu,k-1). \quad (51)$$

The weighting factor for combining the input spectrum and the pre-estimate is given by

$$W_{\widehat{B}}(\mu,k) = \min\left\{1, \left(\frac{\widehat{B}_{\text{pre}}(\mu,k)}{\overline{Y}_{\text{c}}(\mu,k)}\right)^2\right\}. \quad (52)$$

The final noise estimate is computed by applying this weighting factor as shown in Eq. (43). During the first few frames of the noise estimation algorithm, the input spectrum itself is directly chosen as the noise estimate for faster convergence. The plot in Fig. 11 shows the result of the noise estimation algorithm for subband $\mu = 34$.

4.5.6 Results

The proposed algorithm was evaluated under different automobile noise conditions to test the performance in realistic situations encountered while driving. Noise recordings were performed under different speeds, with the air-condition system turned on and off, opening/closing of a window, accelerating to a high speed, breaking to a low speed, etc. The so-called Harvard sentences [34] were used for mixing speech for different SNRs. Noise recordings from one of the belt microphones were used for the evaluation. Two sentences of male and female speakers in a total length of 20 s were used for the evaluation. Figure 12 shows the log-error distance plot of the proposed noise estimation algorithm as compared to the speech presence probability (SPP) scheme [35] and the minimum statistics method [36] as these were the best among the five estimation schemes that were evaluated in our tests for different SNRs [37–39]. These schemes were evaluated as a noise estimation scheme applied to the belt microphone system. The log-error measure is a way to compute the distance between the true noise $\widehat{B}(\mu,k)$ and the estimated noise $\widehat{B}_{\text{f}}(\mu,k)$ given by [40]

$$\Delta_B(\mu,k) = 20\log_{10}|B(\mu,k)| - 20\log_{10}\left|\widehat{B}_f(\mu,k)\right|,$$
$$Log_{\text{err}} = \frac{1}{L_{\text{seg}}N_{\text{Sbb}}}\sum_{k=0}^{L_{\text{seg}}-1}\sum_{\mu=0}^{N_{\text{Sbb}}-1}|\Delta_B(\mu,k)|, \quad (53)$$

where L_{seg} is the number of frames of the noisy signal. The errors for the proposed scheme occur mainly when a rising noise has to be followed. The SPP-based estimate follows the noise well but also follows some speech segments, thereby distorting parts of speech. The minimum-statistics-based estimation was not able to follow the rising spectrum parts fast enough. The proposed algorithm performed better in terms of segmental SNR and overall SNR improvement. The estimated noise PSD is used by the postfilter to suppress the noise and residual echoes by using a suppression filter that is presented in the next section.

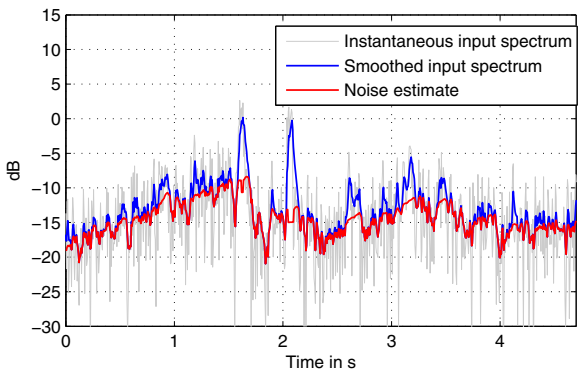

Fig. 11 Plot showing the noise estimate at subband $\mu = 34$

4.6 Residual echo and noise suppression

The output of the adaptive beamformer contains two unwanted components:

- The residual echo from the echo canceler which is caused due to imperfect cancelation
- The background noise of the automobile

Although an estimate of the residual echo $E_u(\mu,k)$ is already available from the echo canceler as per Eq. (13), the application of the adaptive beamformer on the error spectra modifies the estimated residual echo. The residual echo now needs to be re-estimated. Adaptive beamforming can be viewed as spatial filtering applied on the belt microphones. Hence, the filter applied on the error spectra cannot be directly applied on residual echo magnitude estimates. Given this, a good approximate of the residual echo can be obtained from the already available estimate for each microphone on the belt microphone array by applying a maximum over the entire array to get $\widehat{E}_u^2(\mu,k)$. Even if the gain by the beamformer is not included here, it turned out that this approach leads to satisfying results in terms of the echo and so-called double-talk performance of the overall system:

$$\widehat{E}_u(\mu,k) = \max_{l=0\ldots M-1}\left\{\overline{X}_l(\mu,k)\,\beta_{\text{coupl},l}(\mu,k)\right\}. \quad (54)$$

The residual echo estimate along with the background noise estimate $\widehat{B}_{\text{f}}(\mu,k)$ is suppressed by a modified

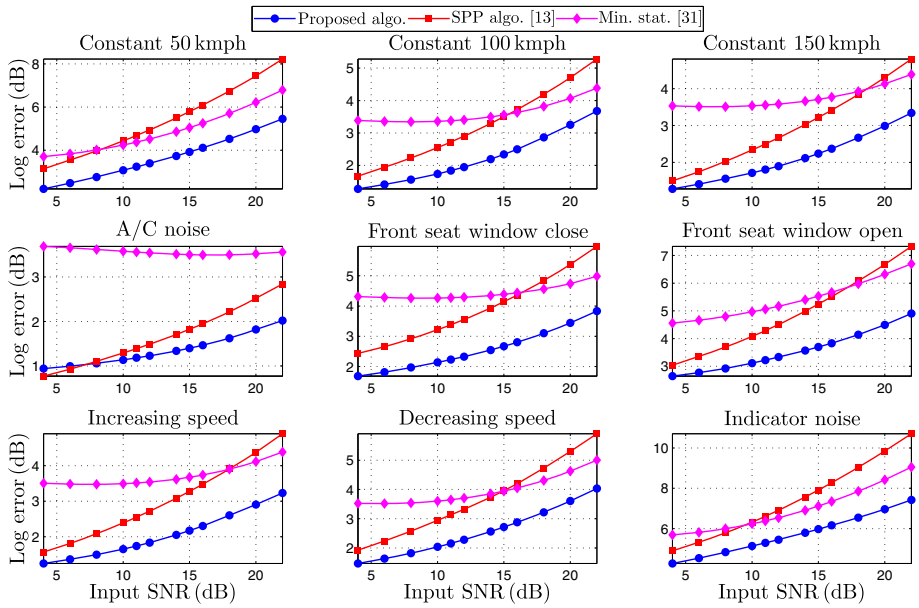

Fig. 12 Log-error plot of the proposed algorithm as compared to SPP [35] and MS [36] under various automobile noise conditions

Wiener filter [41] in the subband domain given by

$$\tilde{\zeta}_{\mathrm{i}}(\mu,k) = 1 - \frac{\Lambda_{\mathrm{res}}\,\widehat{E}^2_{\mathrm{u}}(\mu,k) + \Lambda_{\mathrm{ns}}\,\widehat{B}^2_{\mathrm{f}}(\mu,k)}{|Y_{\mathrm{c}}(\mu,k)|^2}, \tag{55}$$

where $\tilde{\zeta}(\mu,k)$ is the instantaneous Wiener filter. The filter is usually limited by a floor value (usually about -8 to -15 dB) to control the maximum attenuation applied by the filter:

$$\zeta(\mu,k) = \max\left\{\tilde{\zeta}(\mu,k),\ \zeta_{\mathrm{floor}}\right\}. \tag{56}$$

Equation (55) also contains two parameters Λ_{res} and the Λ_{ns} which are overestimation factors applied for the residual echo and the background noise estimate, respectively. This ensures that estimated values are compensated for incorrect estimates especially when the filter does not attenuate sufficiently. The output of the noise suppression filter is the estimated clean speech for each belt. It is obtained by

$$\widehat{S}_{\mathrm{e}}(\mu,k) = \tilde{\zeta}(\mu,k)\,Y_{\mathrm{c}}(\mu,k). \tag{57}$$

Whenever the estimated residual echo is larger than the background noise level reduced by the spectral floor, the output signal is replaced by so-called comfort noise at an appropriate level. Therefore, the following condition is checked:

$$\widehat{E}_{\mathrm{u}}(\mu,k) > \widehat{B}_{\mathrm{f}}(\mu,k)\,\zeta_{\mathrm{floor}}. \tag{58}$$

4.7 Belt microphones in multi-seat scenarios: speech mixer

In a setup which involves multiple seat belts fitted with belt microphones, all channels have to be mixed before they can be sent out via the single channel phone uplink. A first solution for mixing the channels is to simply add all of them. This would lead to a simple mixing scheme, but the overall broadband noise would increase. The solution presented here performs what is called *noise equalization* among the various seats before mixing them with individual time-varying weights. The weights are determined mainly on whether the channels (after beamforming and posfiltering) are active, meaning that the passenger on the specific seat is speaking. Before the channels are added, the noise in every channel should be the same. This is done by means of noise equalization, which is described at the end of this section. In general, the output of the speech mixer signal is computed by

$$\widehat{S}(\mu,k) = \sum_{p=0}^{P-1} \widehat{S}^{(p)}_{\mathrm{e,noise\text{-}eq}}(\mu,k)\,w^{(p)}(k), \tag{59}$$

where $\widehat{S}(\mu,k)$ is the output signal and $\widehat{S}^{(p)}_{\mathrm{e,noise\text{-}eq}}(\mu,k)$ are the noise equalized input channels to be mixed. Since we will combine now the outputs of the individual seats, the index (p) is no longer omitted. The time-varying weights $w^{(p)}(k)$ are combinations of two factors

$$w^{(p)}(k) = a^{(p)}(k)\,b(k). \tag{60}$$

The activity weight $a^{(p)}(k)$ is based on whether the channel has been detected to have activity and the background noise weight $b(k)$ *normalizes* the background noise to stay constant assuming all noise components from the individual seat being mutually orthogonal. Figure 13 shows the signal flow diagram of the speech mixer with the activity and the background noise weights.

The activity weight $a^{(p)}(k)$ controls the opening and closing of a given channel, and hence, the value of this weight lies between 0 dB and A_{att}. The background noise weight $b(k)$ is computed based on these activity weights:

$$b(k) = \frac{1}{\sqrt{\sum_{p=0}^{P-1} \left(a^{(p)}(k)\right)^2}}. \tag{61}$$

The background noise weight is derived by assuming the remaining noise output in terms of its power spectral density should be the same as the noise level of all inputs after equalization:

$$\begin{aligned} &\mathrm{E}\left\{\left|\widehat{S}(\mu,k)\right|^2\right\}\Big|_{\text{during speech pauses}} \\ &\quad \stackrel{!}{=} \mathrm{E}\left\{\left|\widehat{S}^{(p)}_{\text{e,noise}-\text{eq}}(\mu,k)\right|^2\right\}\Big|_{\text{during speech pauses}} \quad \forall p, \end{aligned} \tag{62}$$

where $\stackrel{!}{=}$ indicates "should be equal to." Starting from that condition, using mutual orthogonality among all seat outputs, and restricting the weight $b(k)$ to be positive, leads to the solution given in Eq. (61).

4.7.1 Activity detection

The acoustic arrangement of the belt microphones in a car is such that when only one talker is active, his/her voice is also captured by belt microphones of other passenger seats which could result in incorrect detection for the non-speaking passengers. If there is no mechanism to differentiate this, then the result will be an addition of all the channels. The problem here is to identify the "true" passenger microphone where the talker is active. The approach adopted here is to find the loudest channel among the available channels. The activity decision for the loudest channel is set to 1, and the other channels are decayed with a falling constant up to A_{att}. Before a loudest channel search is performed, a simple VAD is performed as a pre-detection. The pre-detection ensures that channels where the signal is not loud enough is omitted from the loudness search. For example, the belt microphone immediately behind the driver's seat can be omitted from the loudness search when only the driver is active. The loudest channel search is performed over the subbands with a minimum SNR T_{loud} as compared to the background noise $\widehat{B}_{\text{f}}(\mu,k)$

$$\tilde{\chi}^{(p)}(\mu,k) = \begin{cases} 1, \text{ if } \left|\widehat{S}^{(p)}_{\text{e}}(\mu,k)\right| > T_{\text{loud}}\,\widehat{B}^{(p)}_{\text{f}}(\mu,k), \\ 0, \text{ else,} \end{cases} \tag{63}$$

where the computation of the noise level $\widehat{B}^{(p)}_{\text{f}}(\mu,k)$ was presented in Eq. (43). The individual subbands are summed to obtain a value per passenger

$$\chi^{(p)}(k) = \sum_{\mu=0}^{N_{\text{FFT}}-1} \tilde{\chi}^{(p)}(\mu,k). \tag{64}$$

A maximum search is performed over the $\chi^{(p)}(k)$ value to obtain the loudest passenger index $p_{\max}(k)$ given by

$$p_{\max}(k) = \begin{cases} p_{\max}(k-1), \\ \quad \text{if } \max\limits_{p}\left\{\chi^{(p)}(k)\right\} < \chi_{\min}, \\ \operatorname{argmax}\limits_{p}\left\{\chi^{(p)}(k)\right\}, \\ \quad \text{else.} \end{cases} \tag{65}$$

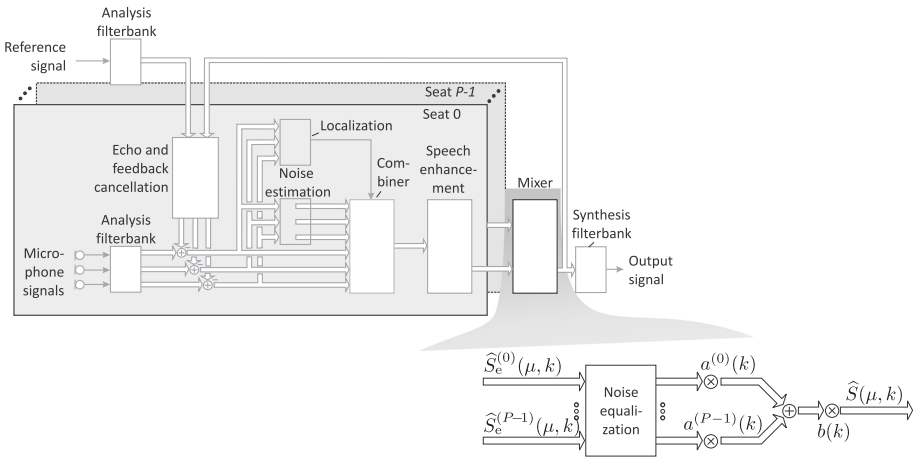

Fig. 13 Input-output signal flow of the mixer (depicting the two types of weights) and its position in the entire processing structure

However, the index of the loudest passenger is only updated if a sufficient amount of subbands show a large SNR. The weights resulting for the activity decision are changing on the bases of multiplicative time-constants. This ensures a smooth transition when the estimated talker's activity is switching from one passenger to another. To compute the activity weights $a^{(p)}(k)$, first, a default time-constant $\Delta_{\text{act-def}}$ is applied to the activity weights. This is then followed by a faster incremental time constant if the current channel is found to be the loudest; else, it is decremented with a slow time constant. The activity decisions are then set according to

$$a_{def}^{(p)}(k) = \min\left\{a^{(p)}(k-1)\,\Delta_{\text{act-def}},\, A_{\text{att}}\right\}, \tag{66}$$

with

$$a^{(p)}(k) = \begin{cases} \min\left\{a_{\text{def}}^{(p)}(k)\,\Delta_{\text{act-inc}},\, 1\right\}, \\ \quad \text{if } p = p_{\max}(k), \\ \max\left\{a_{\text{def}}^{(p)}(k)\,\Delta_{\text{act-dec}},\, A_{\text{att}}\right\}, \\ \quad \text{else.} \end{cases} \tag{67}$$

4.7.2 Noise equalization

Before the channels can be added, the noise for every channel has to be the same. This is achieved by a slowly changing average-noise-gain tracker $G^{(p)}(\mu,k)$ multiplied with every channel. The noise gains are aimed to bring all channels to an average noise level. This is performed by a two-stage procedure. First, the gains of the last frame are updated in a preliminary fashion ($G^{(p)}(\mu,k-1) \rightarrow \widetilde{G}^{(p)}(\mu,k)$). Afterwards, the preliminary gains are limited, resulting in the final gains for the current frame ($\widetilde{G}^{(p)}(\mu,k) \rightarrow G^{(p)}(\mu,k)$). An average noise $\widehat{B}_{\text{f,avg}}(\mu,k)$ of all channels is computed by

$$\widehat{B}_{\text{f,avg}}(\mu,k) = \frac{1}{P}\sum_{p=0}^{P-1}\widehat{B}_{\text{f}}^{(p)}(\mu,k). \tag{68}$$

The background noise estimates are the ones that were used in the postfilter section (see Eq. (43)). The noise gains per channel are tracked with the help of slow time constants:

$$\widetilde{G}^{(p)}(\mu,k) = \begin{cases} G^{(p)}(\mu,k-1)\,\Delta_{gain-inc}, \\ \text{if } \widehat{B}_{\text{f}}^{(p)}(\mu,k)\,G^{(p)}(\mu,k-1) \\ \qquad < \widehat{B}_{f,avg}(\mu,k), \\ G^{(p)}(\mu,k-1)\,\Delta_{\text{gain-dec}}, \text{else.} \end{cases} \tag{69}$$

The gains $G^{(p)}(\mu,k)$ are finally limited by G_{max} and $G_{\min}$:

$$G^{(p)}(\mu,k) = \max\left\{G_{\min},\, \min\left\{G_{\max},\, \widetilde{G}^{(p)}(\mu,k)\right\}\right\}. \tag{70}$$

These noise gains are applied to the respective input channels resulting in

$$\widehat{S}_{\text{e,noise-eq}}^{(p)}(\mu,k) = \widehat{S}_{\text{e}}^{(p)}(\mu,k)\,G^{(p)}(\mu,k). \tag{71}$$

Although the usefulness of the speech mixer for a multi-seat conference is easy to see, there are no readily available evaluation methods for the mixer. One way to evaluate the mixer is to measure how fast the activity detection switches for continuously changing speakers in different seat positions. The noise equalizer can be evaluated by creating different noise situations at each seat position. Due to time and space constraints, the speech mixer presented here was subjectively evaluated by a remote speaker for scenarios involving conversation with passengers seated in different seat positions in real driving conditions. During the first part of the conversation, the mixer was turned off, and for second part of the conversation, the mixer was turned on. The remote speakers always positively acknowledged the improved speech situation from the car when the mixer was turned on.

5 Summary

The paper highlighted various positions in which microphones can be placed inside an automobile for capturing speech and compared conventional sensors to belt microphones. The natural SNR advantage of the belt microphones is hindered by various properties of the seat belt. For example, movement relative to the local speaker and the loudspeaker lead to acoustic echo path changes. The work presented in this paper has tried to solve these problems using appropriate signal processing schemes. First, the problem of continuously changing echo paths is managed by combining existing methods in step-size control with improvements through a delay estimate for faster adaptation. To tackle a large change in the echo path, a shadow-filter-based approach was presented. By using these additional filter, the coupling factors between the reference and error spectra are triggered to new values whenever a room change was detected. The results showed that the detection of the room change and the coupling trigger helps in quickly re-adapting to the changed frequency response. The next stage of signal processing were three different approaches for acquiring the local speech in the form of choosing a single microphone with the best SNR, an SNR-based microphone combination, and finally an adaptive beamformer where a first SNR-based beamformer is used in combination with an adaptive blocking matrix and an interference canceler. The adaptive blocking matrix is used to generate a reference for the interference canceler. Real measurement in cars showed that even with this highly time-varying setup, still some benefit could be achieved with beamforming techniques. For the estimation of the overall background noise of the vehicle, a simple but

reliable and effective noise estimation method based on switching multiplicative constants was presented. The switching is based mainly on a long-term activity measure, the previous noise estimates, and the smoothed input spectrum. This results in an overall improved performance for the automobile noise situations compared to two other well-established noise estimation schemes. As a last step, a speech mixer was presented that combines the different belts from various seats to a single output. The mixer is designed in such a way that it differentiates between single-talker and multiple-talker situations by controlling the attenuation applied for every belt. The mixer also performs a simple noise equalization to maintain a constant background noise even if the individual belts might have different individual noise levels. The work shows that much of the problems presented initially for the belt microphones have been successfully handled through the various signal processing techniques developed. These techniques are supported with results using measures commonly applied by the research community.

6 Conclusions

This paper has shown that belt microphones can be used as an alternative to traditionally placed microphones. The algorithms presented in this paper are applied in real systems, and the results were rather promising. The expected problems, such as short echo bursts after movements of the passengers, can be avoided by appropriate control schemes. The presented methods have also taken into account multiple inputs, and an extension to multiple output is rather straightforward. The methods have also been tested against industry standards like the ITU [25] specifications.

The next steps towards improving the system can be made in several directions: since the echo cancelation schemes involve(s) several adaptive filters and control units, methods to reduce complexity by reducing processing bands, finding a correlation between the loudspeaker and the different microphone arrays can be explored. On the same line, the overall complexity of the adaptive beamformer could be reduced by further using the same attempts. In terms of noise estimation and suppression, other methods like a time-varying attenuation floor as presented in [42] can be examined. These methods also have the potential to be extended further to other systems like high-quality multichannel audio-video conferencing systems. Finally, the belt microphones itself can be used to measure the noise level as heard by the passengers since they are close to the ears. This information can be used to control the gain of the level control unit for the remote-side signals. The same holds for the adjustment of the playback level of music signals.

Endnotes

[1]Again, we dropped the superscript (p) for better readability here, since the same processing scheme is applied to all seats. The subscript l, however, is required now—in contrast to the previous sections.

[2]In order to be independent of the sample rate and frameshift, all time-constants are denoted in dB/s.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
The authors would like to thank all the colleagues at Digital Signal Processing and System Theory, University of Kiel, and paragon AG who helped in different ways while conducting various driving tests and different recordings, and special thanks for reviewing the manuscript and providing valuable inputs at various stages of development.

Author details
[1]Digital Signal Processing and System Theory, University of Kiel, Kaiserstrasse 2, Kiel, Germany. [2]Division Acoustics, paragon AG, Schwalbenweg 29, Delbrueck, Germany.

Received: 12 July 2015 Accepted: 26 February 2016
Published online: 15 March 2016

References

1. E Hänsler, The Hands-free telephone problem: an annotated bibliography. Signal Process. **27**(3), 259–271 (1992)
2. E Hänsler, The hands-free telephone problem: an annotated bibliography update. Ann. Telecommun. Annales des Télécommun. **79**(7-8), 360–367 (1994)
3. G Schmidt, T Haulick. Signal processing for in-car communication systems.Signal Process. **86**(6), 1307–1326 (2006)
4. J Withopf, C Lüke, H Özer, G Schmidt, in *5th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems*. Signal Processing for In-car Communication Systems (Kiel, Germany, p. 2011. http://www.dss.tf.uni-kiel.de/en/events/workshops/dsp-in-vehicles-2011/the-5th-biennial-workshop-on-digital-signal-processing-for-in-vehicle-systems
5. H Höge, S Hohenner, B Kämmerer, N Kunstmann, S Schachtl, M Schönle, P Setiawan, in *Automatic Speech Recognition on Mobile Devices and Over Communication Networks Advances in Pattern Recognition*, ed. by Z Tan, B Lindberg. Automotive Speech Recognition, vol. 2008 Spinger, Berlin, Germany, 2008), pp. 347–373
6. M Brandstein, *D Ward, Microphone Arrays*. (Springer, Berlin, Germany, 2001)
7. J Freudenberger, Microphone diversity combining for in-car applications. EURASIP J. Adv. Signal Process. **2010** (2010). http://asp.eurasipjournals.springeropen.com/articles/10.1155/2010/509541
8. AG Paragon. http://www.paragon.ag. Accessed 25 Feb 2016
9. J Withopf, G Schmidt, in *Speech Communication; 10. ITG Symposium; Proceedings of*. Suppression of Instationary Distortions in Automotive Environments, (Germany, 2012). http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6309586&queryText=Suppression%20of%20Instationary%20Distortions%20in%20Automotive%20Environments&newsearch=true
10. JB Allen, Short-term spectral analysis, synthesis, and modification by discrete fourier transform. IEEE Trans. Acoust. Speech Signal Process. **25**(3), 235–238 (1977)
11. D Mauler, R Martin, in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. Optimization of switchable windows for low-delay spectral analysis-synthesis, (2010), pp. 4718–4721. doi:10.1109/ICASSP.2010.5495181, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5495181&newsearch=true&queryText=Optimization%20of%20Switchable%20Windows%20for%20Low-Delay%20Spectral%20Analysis-Synthesis
12. J Withopf, L Jassoume, G Schmidt, A Theiss, in *DAGA 2012 Ű 38. Deutsche Jahrestagung für Akustik*. Modified Overlap-Add Filter Bank With Reduced Delay (Darmstadt, Germany, p. 2012. https://www.dega-akustik.de/publikationen/daga-tagungen/verzeichnisse.html

13. MM Sondhi, The history of echo cancellation. IEEE Signal Process. Mag. **23**(5), 95–102 (2006)
14. S Haykin, *Adaptive Filter Theory*, 3rd ed. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996)
15. C Breining, P Dreiseitel, E Hänsler, A Mader, B Nitsch, H Puder, T Schertler, G Schmidt, J Tilp. IEEE Signal Process. Mag. **16**(4), 42–69 (1999)
16. C Antweiler, J Grunwald, H Quack, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Approximation of optimal step size control for acoustic echo cancellation, vol. 1, (1997), pp. 295–298. doi:10.1109/ICASSP.1997.599627, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=599627&newsearch=true&queryText=Approximation%20of%20optimal%20step%20size%20control%20for%20acoustic%20echo%20cancellation,%20acoustics,%20speech,%20and%20signal
17. MA Iqbal, SL Grant, in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. Novel variable step size NLMS algorithms for echo cancellation, (2008), pp. 241–244. doi:10.1109/ICASSP.2008.4517591, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4517591&newsearch=true&queryText=Novel%20variable%20step%20size%20nlms%20algorithms%20for%20echo%20cancellation
18. C Paleologu, J Benesty, S Ciochina, A variable step-size affine projection algorithm designed for acoustic echo cancellation.Audio Speech Lang. Process. IEEE Trans. **16**(8), 1466–1478 (2008)
19. C Carlemalm, A Logothetis, in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. On detection of double talk and changes in the echo path using a Markov modulated channel model, vol. 5, (1997), pp. 3869–3872. doi:10.1109/ICASSP.1997.604743, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=604743&newsearch=true&queryText=On%20detection%20of%20double%20talk%20and%20changes%20in%20the%20echo%20path%20using%20a%20Markov%20modulated%20channel%20model
20. WC Lee, KH Jeong, DH Youn, in *Circuits and Systems, 1997. Proceedings of the 40th Midwest Symposium on*. A robust stereophonic subband adaptive acoustic echo canceller, vol. 2, (1997), pp. 1350–1353. doi:10.1109/MWSCAS.1997.662332, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=662332&newsearch=true&queryText=A%20robust%20stereophonic%20subband%20adaptive%20acoustic%20echo%20canceller
21. F Amano, HP Meana, A de Luca, G Duchen, A multirate acoustic echo canceler structure. IEEE Trans. Commun. **43**(7), 2172–2176 (1995)
22. MA Iqbal, SL Grant, in *Region 5 Technical Conference, 2007 IEEE*. A novel normalized cross-correlation based echo-path change detector, (2007), pp. 249–251. doi:10.1109/TPSD.2007.4380390, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4380390&newsearch=true&queryText=A%20Novel%20Normalized%20Cross-Correlation%20Based%20Echo-Path%20Change%20Detector
23. W Armbrüster, in *Signal Processing*. Wideband Acoustic Echo Canceller with Two Filter Structure (Elsevier, Belgium, 1992), pp. 1611–1614. http://www.sciencedirect.com/science/article/pii/B9780444895875501055, https://s100.copyright.com/AppDispatchServlet?publisherName=ELS&contentID=B9780444895875501055&orderBeanReset=true
24. K Ochiai, T Araseki, T Ogihara, Echo canceller with tow echo path models. IEEE Trans. Commun. COM-25, 589–595 (1977)
25. ITU: ITU-T P.1110 – Wideband hands-free communication in motor vehicles (2015). http://www.itu.int/. Accessed 4 July 2015
26. C Knapp, G Carter, The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. **24**(4), 320–327 (1976)
27. GC Carter, AH Nuttall, P Cable, The smoothed coherence transform. IEEE Trans. Acoust. Speech Signal Process. **61**(10), 1497–1498 (1973)
28. M Krini, K Rodemer, *Seat Belt-Microphone Systems and their Application to Speech Signal Enhancement*. (Oldenburg, Germany, p. 2014
29. M Krini, VK Rajan, K Rodemer, G Schmidt, *Adaptive Beamforming for Microphone Arrays on Seat Belts*. Nuremberg, Germany, 2015). https://www.dega-akustik.de/publikationen/daga-tagungen/verzeichnisse.html
30. LJ Griffiths, CW Jim, An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. **30**(1), 24–34 (1982)
31. DH Johnson, DE Dudgeon. 1st ed., in *Englewood Cliffs, NJ, USA*. Array Signal Processing: Concepts and Techniques (Prentice Hall, 1993)
32. H Cox, RM Zeskind, MM Owen, Robust adaptive beamforming. IEEE Trans. Acoust. Speech Signal Process. **35**(10), 1365–1375 (1987)
33. O Hoshuyama, A Sugiyama, A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters. IEEE Trans. Signal Process. **61**(10), 1497–1498 (1973)
34. IEEE, Recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoustics. **17**(3), 225–246 (1969)
35. T Gerkmann, RC Hendriks, Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1383–1393 (2012)
36. R Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)
37. I Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003)
38. G Doblinger, Computationally efficient speech enhancement by spectral minima tracking in subbands. Proc. Eurospeech. **2**, 1513–1516 (1995)
39. HG Hirsch, C Ehrlicher, in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. Noise estimation techniques for robust speech recognition, vol. 1, (1995), pp. 153–156. doi:10.1109/ICASSP.1995.479387, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=479387&queryText=Noise%20estimation%20techniques%20for%20robust%20speech%20recognition&newsearch=true
40. PC Loizou, *Speech Enhancement: Theory and Practice. Signal Processing and Communications*, 1st ed. (CRC press, 2007). https://www.crcpress.com/Speech-Enhancement-Theory-and-Practice/Loizou/9780849350320
41. N Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series: With Engineering Applications*. (MIT Press, 1949). http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6284744, http://ieeexplore.ieee.org/xpl/bkabstractplus.jsp?bkn=6267356
42. VK Rajan, C Baasch, M Krini, G Schmidt, in *Speech Communication; 11. ITG Symposium; Proceedings of*. Improvement in Listener Comfort Through Noise Shaping Using a Modified Wiener Filter Approach (Nuremberg, Germany, 2014), pp. 1–4. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6926063&newsearch=true&queryText=Improvement%20in%20Listener%20Comfort%20Through%20Noise%20Shaping%20Using%20a%20Modified%20Wiener%20Filter%20Approach