Brodersen *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2019) 2019:21
https://doi.org/10.1186/s13636-019-0165-9

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Signal enhancement for communication systems used by fire fighters

Michael Brodersen[1,2*], Achim Volmer[1] and Gerhard Schmidt[2]

Abstract

So-called *full-face masks* are essential for fire fighters to ensure respiratory protection in smoke diving incidents. While such masks are absolutely necessary for protection purposes on one hand, they impair the voice communication of fire fighters drastically on the other hand. For this reason communication systems should be used to amplify the speech and, therefore, to improve the communication quality. This paper gives an overview of communication enhancement techniques for masks based on digital signal processing. The presented communication system picks up the speech signal by a microphone in the mask, enhance it, and play back the amplified signal by loudspeakers located on the outside of such masks. Since breathing noise is also picked up by the microphone, it's advantageous to recognize and suppress it – especially since breathing noise is very loud (usually much louder than the recorded voice). A voice activity detection distinguishes between side talkers, pause, breathing out, breathing in, and speech. It ensures that only speech components are played back. Due to the fact that the microphone is located close to the loudspeakers, the output signals are coupling back into the microphone and feedback may occur even at moderate gains. This can be reduced by feedback reduction (consisting of cancellation and suppression approaches). To enhance the functionality of the canceler a decorrelation stage can be applied to the enhanced signal before loudspeaker playback. As a consequence of all processing stages, the communication can be improved significantly, as the results of measurements of real-time mask systems show.

Keywords: Full-face mask communication, Feedback cancellation, Noise reduction

1 Introduction

A so-called *full-face mask* in combination with a self-contained breathing apparatus—in the following abbreviated as *SCBA*—is essential for a fire fighter to ensure respiratory protection in smoke diving incidents and in toxic environments. Such masks exist since the early 20th century.

One of the masks that is already quiet old but sill in use is shown in Fig. 1. The depicted mask (model *Panorama Nova*, Dräger) was (and is) used by fire fighters in several countries since 1980. This mask could be used for an SCBA and also as a so-called *rebreather* for fire fighting, mining, or in industrial applications. A rebreather is a device that absorbs carbon dioxide of a person's breath to permit rebreathing it.

Early masks such as the model depicted in Fig. 1 support voice communication only in a passive manner. If two fire fighters that wear such masks want to communicate they have to shout to each other. As a consequence, the speech intelligibility in a noisy environment is rather limited with such passive masks.

Beside the communication among fire fighters in a direct neighbourhood also other communication channels are of importance. During incidents, fire fighters usually operate in troops of two to four people and they need to communicate in order to act uniformly, while, e.g., crawling for injured people in a building that is on fire. The head of the troop has to report the observed situation of each room to the team leader, who is usually located outside the building. This is done via a so-called *tactical radio unit.*

The *internal* troop communication (the communication among fire fighters) is not that easy due to the high attenuation of the masks (even of most of today's models). To get an impression of this attenuation, the power spectral density of a multitude of speech signals was measured in a

*Correspondence: michael.brodersen@draeger.com
[1]Dräger Safety AG & Co. KGaA, Revalstr, 1, 23560, Lübeck, Germany
[2]Digital Signal Processing and System Theory, Kiel University, Kaiserstr. 2, 24143, Kiel, Germany

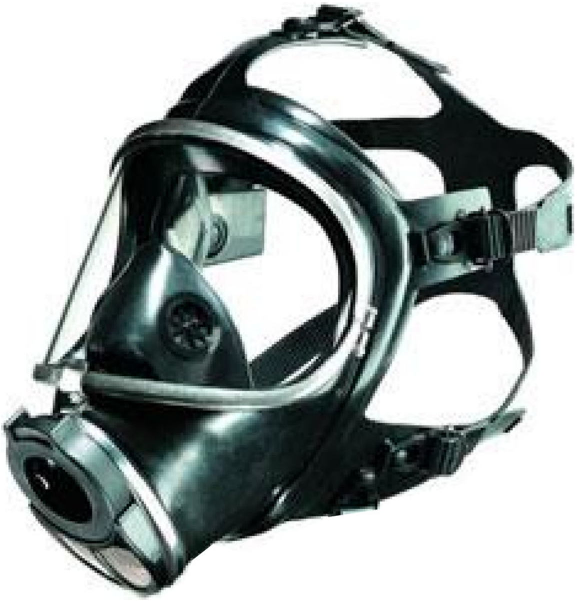


Fig. 1 Full-face mask (model *Panorama Nova*, Dräger [2])

distance of about 1 m in front of the mouth of a torso with an artificial mouth loudspeaker (without any mask). Afterwards, the mask of Fig. 1 was put over the torso and the measurement was repeated. The ratio of the two power spectral densities (depicted in Fig. 2) shows the attenuation due to the mask. This helps to get an impression of the large attenuation and thus for the difficulties that passive masks generate for voice communication. Also the communication via tactical radio is not really a comfortable alternative, because fire fighters have to keep the tactical radio unit in front of their masks and the microphone of this unit picks up the attenuated and distorted speech, which is not very intelligible.

To improve the communication of the masks, signal processing units, which are attached to the masks, have been investigated and are in use since the beginning of this century. These communication units have a microphone inside the mask, which picks up the speech of the fire fighters. This microphone signal is amplified and played back by loudspeakers in front of the mask leading to an enhanced troop communication. Additionally, the speech signal can be supplied to the tactical radio unit to improve the communication with the team leader.

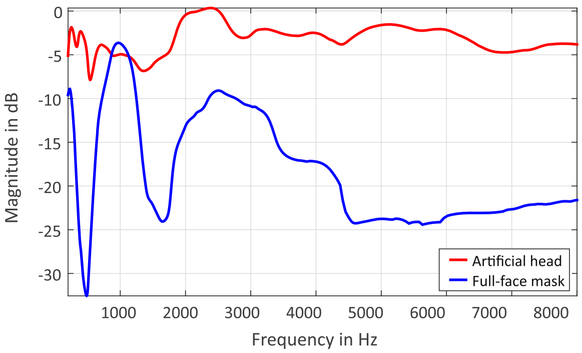


Fig. 2 Frequency response of the full-face mask *Panorama Nova* (blue) and the artificial head without a full-face mask (red)

The signal processing of the first communication systems was purely analog causing potential howling due to the closed-loop behavior of such systems and also breathing noise was not sufficiently suppressed. To solve these problems, the communication systems have continuously been improved in the last years. The results of these developments are described in this contribution. Before going into the details of the involved signal processing units, we would like to mention the special challenges that one faces if working on signal processing for fire fighter masks. Furthermore, we show shortly how this paper is organized and we introduce the notation that will be used in the following.

1.1 Special challenges

When working on communication units of fire fighter masks, special care has to be paid. Fire fighters have to operate in a very large *bandwidth* of conditions. While they are sometimes in very quiet situations, they have to work in the next moment in a very loud environment under very large physical and mental stress. This leads to very large level variations of the speech signals that such communication units have to deal with. Also, the strength of the involved breathing signal can vary a lot from one moment to the next.

Beside the challenges that stem from the involved signals, also the hardware that can be used to implement the invented algorithms has also specific restrictions. Of course, one has to deal with a mobile application, meaning that power consumption as well as power supply are important aspects for communication masks. Even if great technical progress has been achieved in the power supply of mobile phones or hearing aids, the rechargeable batteries of such devices can not be used for communication masks of fire fighters due to temperature requirements. As a consequence only specific battery types can be used which leads to very severe power restrictions. Thus, the algorithms of such communication units have to be designed such that they can operate on fixed-point hardware with limited precision (e.g. 16 bits). Such hardware has—even today—still the lowest power consumption. Also the clock frequencies are usually much smaller (again because of the power/energy restriction) compared to other audio hardware.

Finally, the communication units operate in a closed electro-acoustic loop, since the recorded speech of the fire fighters is played back via small loudspeakers in front of the mouth to improve the communication with, e.g., people that have to be rescued. More details on this issue will be given in the next sections. The consequence is that one faces delay restrictions that are comparable with hearing aids [24], public address systems [18], and so-called *in-car communication systems* [46].

All these boundary conditions make the development of communication units for fire fighters a very challenging (but also interesting) task. However, when keeping in mind that fire fighters rescue our lives day by day, it is pretty clear that engineers should also do their very best to optimize the performance of such communication units.

1.2 Organization of this paper

Our contribution is organized as follows: after the presentation of the properties of the full-face masks in Section 2 the properties of the communication systems will be discussed in Section 3. The signal processing is decomposed into several subunits, which are outlined as follows:

- Preprocessing and analysis filterbanks in Section 3.1
- Feedback cancellation in Section 3.2
- Residual feedback and noise suppression in Section 3.3
- Automatic gain control and equalization in Section 3.4
- Signal decorrelation in Section 3.5
- Post processing and synthesis filterbanks in Section 3.6

Finally the effectiveness of the signal processing algorithms applied on communication systems is presented in Section 4.

1.3 Previous work of the authors related to this contribution

The theory of the step-size control for the feedback cancellation in Section 3.2 was published in 2000 by Mader et al. [36]. The basics of the utilized feedback cancellation in Section 3.2 was published in 2004 by Hänsler and Schmidt [25]. An earlier version of the feedback suppression scheme presented in Section 3.3 was published in 2011 by Lüke et al. [34]. The mask characteristic was published in 2013 by Volmer et al. [52]. Decorrelation schemes for automotive applications were published in 2014 by Withopf et al. [55]. An earlier version of the voice activity detection inside the noise suppression in Section 3.3 was published in 2015 by Brodersen et al. [10].

1.4 Notation

Throughout this contribution the notation will follow some basic rules:

- Scalar quantities such as time-domain signals are written in lowercase, non-bold letters such as $s(n)$ for a signal at time index n.
- Frequency-domain quantities are described by upper case letters such as $X(\mu, k)$. Here, μ indicates the subband or frequency index and k is the frame index.
- Vectors are noted as bold letters, e.g., $\hat{\boldsymbol{H}}(\mu, k)$ represent a vector containing filter coefficients.
- Smoothed signals are noted by over-lined letters such as $\overline{x}(n)$ and estimated signals are written as letters with a hat such as $\hat{x}(n)$.
- All signals are represented in discrete time.

2 Properties of full-face masks

Full-face masks protect the face and the respiratory tract of the mask wearer against toxic gasses and smoke (see [52]). Previous and current masks, as shown in Fig. 3, seal around the face and an SCBA worn on the back is used to ensure a clean air supply. The nose and mouth are covered by a so-called *inner mask*, to direct the (fresh) air stream to the visor while exhaled air is exhausted through a valve to prevent fogging of the mask. Because of the sealing of the masks, the speech of the person wearing the mask is highly attenuated (as already explained before). To overcome this problem, a so-called *speech diaphragm* is placed in front of the mouth. For low and medium frequencies, it acts partly as a resonator. However, frequencies above 2 kHz are largely attenuated. The speech diaphragm has to withstand high chemical exposure to guarantee breathing protection. Typically, the speech diaphragm consists of a thin foil of stainless steel or polyimide which is clamped into a ring [52]. While the diaphragm on the one hand allows speech signals to pass the hermetical sealing of the mask, its mechanical properties simultaneously add distortions (mainly attenuation) to the speech on the other hand.

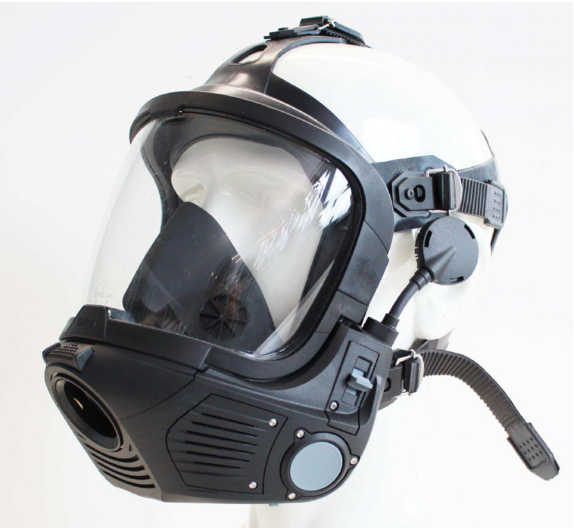


Fig. 3 Full-face mask (FPS 7000, Dräger [3]) with a communication system (FPS-COM 7000, Dräger [4])

The resonance frequencies of typical diaphragms are located around 800 Hz. Unfortunately, this corresponds to the center of the frequency range responsible for proper intelligibility. Furthermore, the sound pressure level (SPL) that is caused by the speech of the fire fighter might be very high (e.g., in situations with large physical stress and/or in loud environments). This might cause a non-linear behavior of the diaphragm.

3 Methods, experimental, and system overview

Breathing protection is typically used in noisy environments and the intelligibility of the voice is limited if only purely mechanical (passive) systems are used. An improvement is achieved by amplifying the speech by means of a communication system which is attached to a mask (Fig. 3). In more details: a microphone and loudspeakers are attached to modern masks (see Fig. 4). The microphone picks up the speech signal. Afterwards this signal is processed and played back via the loudspeakers (and corresponding amplifiers) that are attached usually to the front part of masks. As already mentioned in Section 1.1 the requirements for the microphone and the loudspeakers are really high, because they have to be heat resistant for flame retardant, watertight for cleaning, and shock resistant for the case that fire fighters accidently collide in an incident [52]. Thus, a suitable system has to manage the challenge between acoustical performance and sufficient robustness.

Fig. 4 depicts an overview about the individual components of a mask with a communications system. The (mouth) loudspeakers are located in the front of the mask to optimize direct communication. The so-called *ear loudspeakers* are located close to the ears (without fully covering them). They are intended to perceive the signals of a so-called *tactical radio*. The radio unit is usually realized as an external device which establishes the communication to the control center as well as to other fire fighters via a so-called *team talk*.

The microphone of the communication unit is located in front of the speech diaphragm (outside the sealing volume) and it records the distorted signal. The signal quality would be better if the microphone would be placed inside the mask, but then the opening of the speech diaphragm would decrease and—as a result—the attenuation of the mechanical system would also increase. This modification is not possible because investigations showed that the speech transmission index (STI) [30, 48, 49] decreases and the mechanical system does not satisfy the North America standard NFPA 1981:2013 [38] any more. This restriction leads to a microphone position in front of the speech diaphragm.

The recorded signal of the microphone contains speech, feedback from the loudspeakers, as well as breathing and background noise. The breathing noise has usually the highest sound pressure level and can be removed by appropriate voice activity detection (VAD) combined with an attenuation unit—as explained in the following sections.

It takes about 0.3 ms for the acoustic signal to trace from the (mouth) loudspeakers to the microphone. This leads

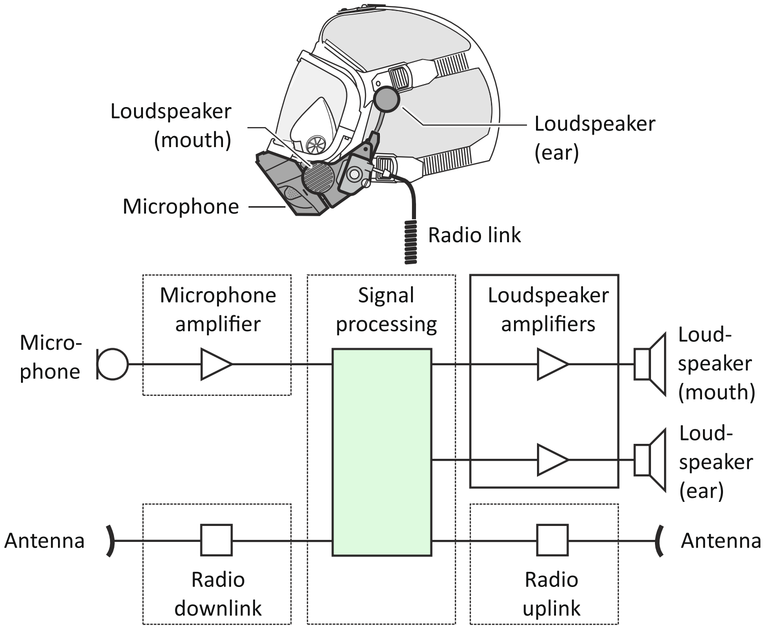


Fig. 4 Structure of the communication system

to potential feedback situations, which can be avoided by applying appropriate feedback suppression algorithms. The stability of the system is essential, because fire fighters use these systems in awkward situations and they have to rely on good communication possibilities. A restriction for the communication system is battery lifetime, because the energy supply is implemented with special heat resistant batteries. Therefore, a very power-efficient signal processing is necessary as already mentioned at the beginning of this paper. Another restriction is the delay of the system, which has to be very low, because otherwise the loudspeaker output signals are perceived as disturbing echoes [7]. An overview of the signal processing is shown in Fig. 5.

The microphone signal will be denoted in the following sections as $x_{\text{mic}}(n)$. The second input signal is the signal received by the radio unit, $x_{\text{radio}}(n)$. This signal is usually called *downlink* signal of the radio unit. After several stages of processing, three types of output signals are generated. On the one hand, the two types of loudspeaker signals are as follows: the signal $y_{\text{mouth}}(n)$ that is played back in front of the mask to improve communication with partners being in the direct neighbourhood of the fire fighter and the ear loudspeaker signal $y_{\text{ear}}(n)$ that is the received radio signal (after enhancement). On the other hand, an enhanced microphone signal $y_{\text{radio}}(n)$ is computed as a last output, being the so-called *uplink* signal. The amount of processing in the two main processing chains (see Fig. 5) is rather different. While only a small amount of processing (mainly automatic gain control, equalization, and limitation) is usually done in the path that connects the radio downlink signal $x_{\text{radio}}(n)$ with the ear loudspeaker output $y_{\text{ear}}(n)$, much more is done in the path that connects the microphone signal $x_{\text{mic}}(n)$ with the mouth loudspeaker output and the uplink signal of the radio unit.

As a first stage, feedback cancellation by means of adaptive filters is applied. Since the performance of such algorithms is usually not sufficient to completely remove the feedback, suppression of remaining feedback as well as suppression of background noise is performed in a separate unit. Beside stationary noise also non-stationary breathing noise is suppressed in this unit. To detect such breathing periods a separate detection unit (in Fig. 5 denoted as *voice activity detection*) is utilized. To adjust different speaking and (noise-dependant) playback levels automatic gain control algorithms are applied in different *flavors*. To overcome the problem that adaptive filters have with correlated excitation and distortion signals (seen from the perspective of an adaptive filter), a decorrelation stage is necessary (or at least beneficial) before playing back the enhanced microphone signals. Since computational complexity must be kept low, all processing stages of the signal path that connects the microphone of the mouth loudspeakers and the radio uplink are embedded in low-delay versions of analysis and synthesis filterbanks. Most of the before mentioned signal processing components will be described in the next sections. For those components that are more or less state-of-the art, we will give only short explanations with references to good descriptions in the literature.

3.1 Preprocessing and analysis filterbanks

Since—as mentioned before—the signal processing must be designed such that only a minimum amount of processing load is required, most of the algorithmic parts are processed in the subband domain. The conversion is achieved by appropriately designed analysis and synthesis filterbanks. However, to use the hardware precision in best manner, so-called *preem phase filters* are applied before entering the subband domain. This is achieved by a simple two-tap FIR filter:[1]

$$x_{\text{pre,mic}}(n) = x_{\text{mic}}(n) - \beta_{\text{pre-de}}\, x_{\text{pre,mic}}(n-1). \tag{1}$$

The filter can also be interpreted as a prediction error filter that has *whitening* properties. The decorrelation coefficient $\beta_{\text{pre-de}}$ is usually chosen in the range:

$$0.95 > \beta_{\text{pre-de}} > 0.99. \tag{2}$$

Afterwards the conversion to the subband domain is performed [26]. This conversion is done here by the fast fourier transformation (FFT). The output of the preem phase filter $x_{\text{pre}}(n)$ is processed with the frame size R and the discrete time index n. The input signal is windowed by an analysis window $h_{\text{ana}}(n)$ before the DFT to improve the aliasing properties (within the subband domain) [9]:

$$X_{\text{mic}}(\mu,k) = \sum_{n=0}^{N_{\text{DFT}}-1} h_{\text{ana}}(n)\, x_{\text{pre,mic}}(n+kR)\, e^{-j\frac{2\pi}{N_{\text{DFT}}}\mu n}, \tag{3}$$

where $X_{\text{mic}}(\mu,k)$ is the resulting spectrum with the subband index μ and the frame index k. Figure 6 shows an overview about the analysis filterbank and the preprocessing.

The key element of filterbanks is the utilized window function. Beside a perfect reconstruction property, two further criteria are important for this application. The aliasing components that appear after the analysis stage should be kept below a certain limit in order to allow for unconstrained adaptive system identification approaches (in the subband domain) and the delay should be kept rather low in order to avoid self perception of the mask wearer. We will use here a DFT order of $N_{\text{DFT}} = 256$

[1] Please note that we perform the preprocessing and the analysis filter bank for the microphone input and for the mouth loudspeaker signals. However, to keep the description short, we will only show the equations of the microphone path. The other path is performed equivalently and results in the short-term spectra $X_{\text{mouth}}(\mu,k)$.

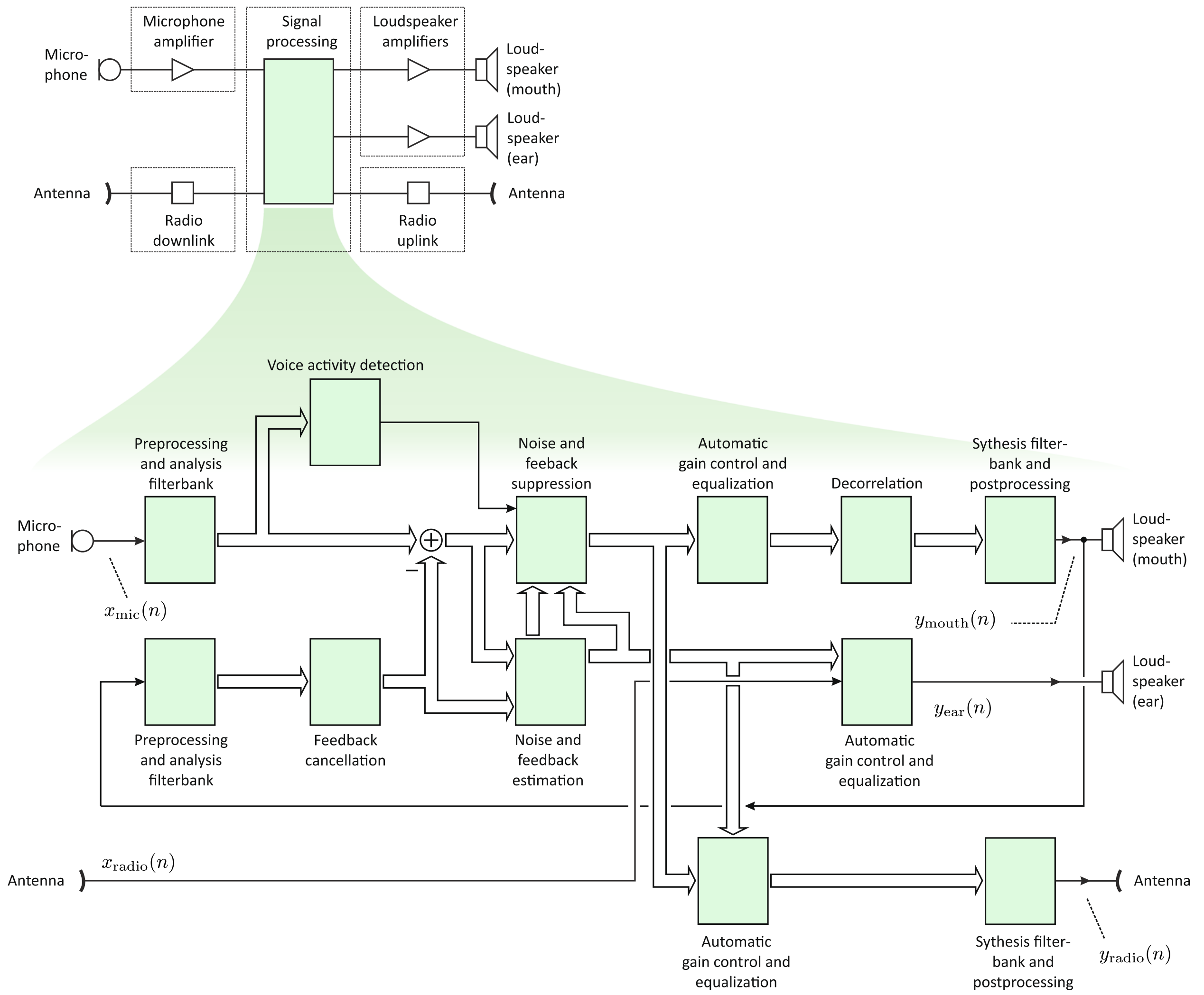


Fig. 5 Signal processing overview

and a basic frameshift of $R = 64$. For the window, an appropriately scaled Hann window is used. Due to the conjugate complex symmetry in the spectral domain only $N_{\mathrm{DFT}}/2 + 1 = 129$ subbands need to be processed. At a sample rate of $f_s = 16$ kHz this results in an overall delay of the analysis/synthesis system (see Section 3.6 for details of the synthesis filterbank) of 16 ms. If this is too high, the filter design method of [54] can be used, resulting in a delay of just 8 ms (with the same frameshift and DFT size). In that case the aliasing, properties are slightly worse, resulting in a maximum feedback reduction (due to cancellation) of about 25 dB. However, this setup is still sufficient for the signal processing approaches described in the following sections.

3.2 Feedback reduction

Feedback is the main problem of the communication unit, because the microphone picks up a large amount of the signals emitted by the mouth loudspeakers. To allow significant amplification of the loudspeaker, signal feedback cancellation (in combination with feedback suppression, described in the next section) is required (see Fig. 5). Such algorithms require an estimation of the loudspeaker signal component that is included in the microphone signal. This is done by a convolution of the estimated impulse responses of the electro-acoustic system with the loudspeaker signal. A subtraction of this estimated signal from the microphone signal cancels the feedback and, thus, allows a higher gain at the loudspeakers. The estimation is performed in the attempt described here using the normalized least mean square (NLMS) algorithms [27, 44], where the estimation of the transfer function works best in case of fully decorrelated signals. The term *decorrelated* means that the mouth loudspeaker signal should be decorrelated from the signal that is recorded in the mask (the speech signal of the fire fighter). Of course, that is

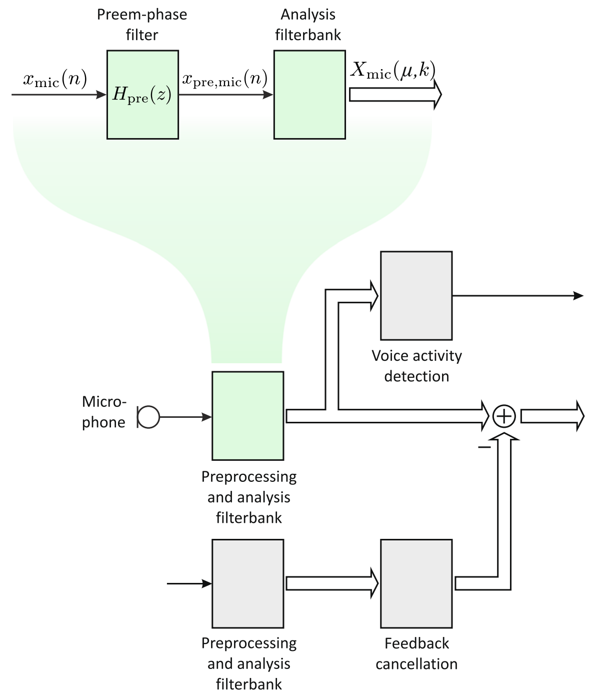


Fig. 6 Structure of the preprocessing andanalysis filterbank stage

the same signal in an ideal case, but from the perspective of system identification attempts, a non-uniqueness and a robustness problem arises from that correlation. For details on that problem (appearing in hearing aids, public address systems, etc.), see [5, 39, 53] for excellent overviews and problem descriptions. Anyhow, a decorrelation of the loudspeaker signal is necessary, which is realized in the approach described here by a frequency shift. The results of the feedback cancellation and suppression attempts are shown in Section 4.2, where the transmission characteristic of the communication system including all described algorithms is illustrated in comparison to a purely mechanical approach.

3.2.1 Feedback cancellation

As already mentioned above, feedback cancellation approaches are estimating the transfer function from loudspeaker to microphone. The output of the convolution of the loudspeaker signal and the transfer function is subtracted from the microphone signal. If the transfer function is estimated correctly, the enhanced microphone signal after the spectral subtraction only includes the speech signal but not the feedback. To save computational complexity, the feedback cancellation is realized in the subband domain as is shown in Fig. 7.

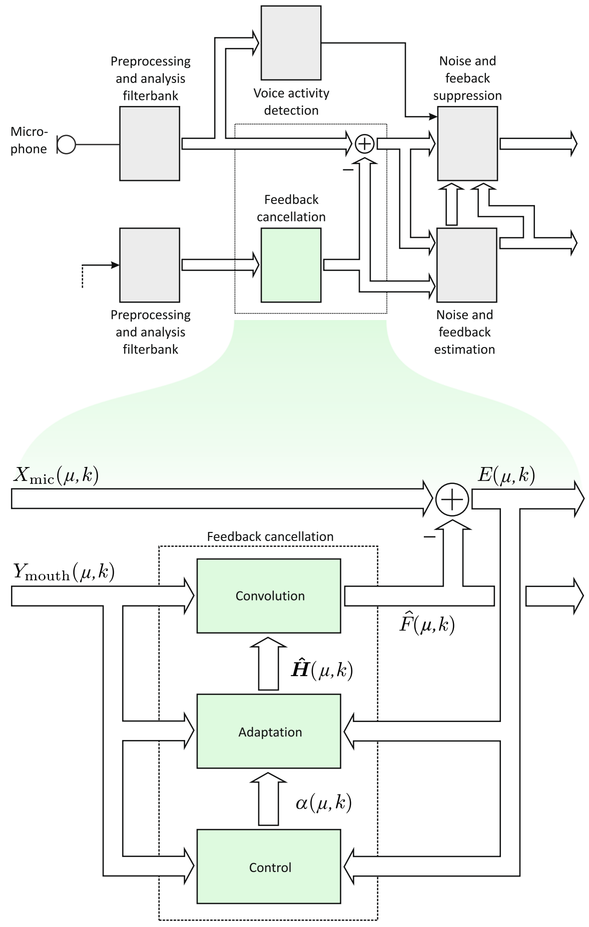

Preprocessing and analysis filterbank
Voice activity detection
Noise and feeback suppression
Micro-phone
−
Feedback cancellation
Preprocessing and analysis filterbank
Noise and feedback estimation
$X_{\text{mic}}(\mu,k)$
$E(\mu,k)$
−
Feedback cancellation
$Y_{\text{mouth}}(\mu,k)$
Convolution
$\hat{F}(\mu,k)$
$\hat{\boldsymbol{H}}(\mu,k)$
Adaptation
$\alpha(\mu,k)$
Control

Fig. 7 Structure of an feedback cancellationsystem operating in subbands (according to [25])

The signal $x_{\text{mic}}(n)$ contains speech $s(n)$, background noise $b(n)$ and the coupled signal from the loudspeakers $f(n)$ (feedback):

$$x_{\text{mic}}(n) = s(n) + b(n) + f(n). \tag{4}$$

This signal is transformed into the frequency domain to perform the complex-valued spectral subtraction of $X_{\text{mic}}(\mu,k)$ and $\hat{F}(\mu,k)$, which is computed by a convolution of the mouth loudspeaker signals with the estimated transfer functions in each subband:

$$\hat{F}(\mu,k) = \hat{\boldsymbol{H}}^{\text{H}}(\mu,k)\, \boldsymbol{Y}_{\text{mouth}}(\mu,k), \tag{5}$$

with $^{\text{H}}$ denoting complex conjugation and transposition and $\boldsymbol{Y}_{\text{mouth}}(\mu,k)$ being a vector containing the last N_{canc} frames

$$\boldsymbol{Y}_{\text{mouth}}(\mu,k) = \Big[Y_{\text{mouth}}(\mu,k),\ ...,\ Y_{\text{mouth}}(\mu,k-N_{\text{canc}}+1)\Big]^{\text{T}}. \tag{6}$$

The enhanced microphone spectrum (also called the error spectrum) $E(\mu,k)$ is created by subtracting the estimated feedback spectrum from the microphone spectrum:

$$E(\mu,k) = X_{\text{mic}}(\mu,k) - \hat{F}(\mu,k). \tag{7}$$

The coefficients of the subband impulse responses $\hat{H}(\mu,k)$ are updated using the NLMS algorithm [25]

$$\hat{\boldsymbol{H}}(\mu,k{+}1) = \hat{\boldsymbol{H}}(\mu,k) + \alpha(\mu,k)\,\frac{\boldsymbol{Y}_{\text{mouth}}(\mu,k)E^*(\mu,k)}{\left\|\boldsymbol{Y}_{\text{mouth}}(\mu,k)\right\|^2}, \tag{8}$$

where the step size has to be chosen close to the pseudo-optimal value for the NLMS algorithm (for a derivation see [36]):

$$\alpha_{\text{opt}}(\mu,k) = \frac{\mathrm{E}\left\{\left|E_{\text{u}}(\mu,k)\right|^2\right\}}{\mathrm{E}\left\{\left|E(\mu,k)\right|^2\right\}}. \tag{9}$$

The problem here is that so-called *undistorted error spectrum* $E_{\text{u}}(\mu,k)$ cannot be measured directly. Thus, it must be estimated (at least its short-term power). It is defined as the error spectrum without the local signal components

$$E_{\text{u}}(\mu,k) = E(\mu,k) - S(\mu,k) - B(\mu,k), \tag{10}$$

where $S(\mu,k)$ and $B(\mu,k)$ are the speech and the noise spectra, respectively. In our approach, we use short-term power smoothing with IIR (infinite impulse response) filters of first order with the smoothing constant β to obtain the required quantities for the step size. The short-term power of the error signal is obtained as

$$P_e(\mu,k) = \beta\, P_e(\mu,k-1) + (1-\beta)\left|E(\mu,k)\right|^2. \tag{11}$$

For the short-term power of the undisturbed error signal, we utilize the so-called *coupling factor method*. Here a delayed version of the reference signal is squared and smoothed

$$P_y(\mu,k) = \beta\, P_y(\mu,k-1) + (1{-}\beta)\left|Y_{\text{mouth}}(\mu,k{-}\Delta)\right|^2 \tag{12}$$

and finally multiplied with a coupling factor $c(\mu,k)$:

$$P_{e_{\text{u}}}(\mu,k) = P_y(\mu,k)\,c(\mu,k). \tag{13}$$

The parameter Δ takes the delay of the electro-acoustic feedback path into account. The coupling factor $c(\mu,k)$ can be set either to a fixed value that stems from a desired echo reduction of, e.g., 15 dB. A better way is to estimate the coupling adaptively by tracking the ratio $P_e(\mu,k)/P_y(\mu,k)$ in such a way that decreasing values are followed much faster than increasing ones. Since feedback cancellation filters are (in contrast to echo cancellation filters) in a permanent double-talk period, the update of the coupling factors is performed only during falling signal edges of the microphone signal power. In such situations the local speech activity is usually smaller compared to the loudspeaker output. For further details about this method, the reader is referred to [36] for an extended derivation and to [12] for details on the proposed method. Using both short-term power estimations the optimal step size can be approximated as

$$\alpha_{\text{opt}}(\mu,k) \approx \alpha(\mu,k) = \frac{P_{e_{\text{u}}}(\mu,k)}{P_e(\mu,k)}. \tag{14}$$

3.2.2 Residual feedback estimation

In order to further expand the system gain, the residual feedback can be analyzed and estimated. With the estimation of the (short-time) power spectral density of the residual feedback, it is possible to suppress this undesired signal component with, e.g., a Wiener filter (described in Section 3.3.5). The feedback estimation is split into an attenuation and a coupling part. The attenuation part comprises reverberation estimated by the reverberation time T_{60} in seconds [34]:

$$T_{60}(\mu) = -\frac{3000\,R}{\log\left(\alpha_{\text{feedb}}(\mu)\right) f_{\text{s}}}, \tag{15}$$

which yields to the to the attenuation factor

$$\alpha_{\text{feedb}}(\mu) = 10^{-\frac{3000\,R}{T_{60}(\mu) f_{\text{s}}}}. \tag{16}$$

The reverberation time T_{60} describes the time needed until the signal is attenuated by 60 dB. This parameter can be extracted in a frequency selective manner out of the filter coefficients of the feedback cancellation filter (under the assumption of a sufficient degree of convergence). This is shown in Fig. 8 for the full-face mask with a communication system attached.

The reverberation time of the full-face mask depends on one the hand on the mask itself but also on the environment in which the mask is used. Our experiments show that in most situations about 100 ms to 300 ms are estimated. This is very long compared to other audio devices; for example the T_{60} time in a car is about 50 ms (see [34]). The relevant frequencies are located above 1 kHz, because the loudspeakers can only transmit in that range due to the small housing.

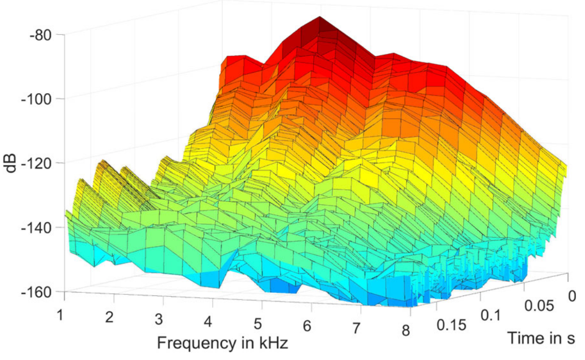


Fig. 8 Cumulative spectral decay of the full-face mask with activated communication system

The coupling part is described by the coupling factor $c_{\mathrm{feedb}}(\mu,k)$, which is calculated based on the measured/estimated coupling from the loudspeaker to the error signal (including the attenuation caused by the cancellation filter). The feedback reduction due to the feedback cancellation is estimated by the factor $c(\mu,k)$ (see the last section for details). This parameter, however, covers only the attenuation between the microphone and the error signal, but not the coupling from the loudspeaker to the microphone of the mask. This coupling, denoted by $c_{\mathrm{ls,mic}}(\mu)$ is not changing over time and is mainly determined by the mechanical setup of the mask. It can be measured off-line and stored as a fixed (frequency selective) parameter. Thus, the entire coupling factors $c_{\mathrm{feedb}}(\mu,k)$ are determined as

$$c_{\mathrm{feedb}}(\mu,k) = c(\mu,k)\, c_{\mathrm{ls,mic}}(\mu) \tag{17}$$

For the computation of the (short-term) power, spectral density of the residual feedback the instantaneous squared magnitude of the loudspeaker spectrum is computed as a first step:

$$P_{y_{\mathrm{m}}}(\mu,k) = \left|Y_{\mathrm{mouth}}(\mu,k)\right|^2. \tag{18}$$

By using a coupling-based estimation scheme and the average feedback attenuation per frame according to Eq. (16) we can estimate the short-term power spectral density of the residual feedback as:

$$\begin{aligned} P_f(\mu,k) =& \alpha_{\mathrm{feedb}}(\mu)\, P_f(\mu,k-1) + \ldots \\ &\ldots + c_{\mathrm{feedb}}(\mu,k)\, P_{y_{\mathrm{m}}}(\mu,k-\Delta). \end{aligned} \tag{19}$$

Again Δ is the delay of the acoustic path between the loudspeaker in front of the mask and the microphone (see previous section) in frames. The coupling factor $c_{\mathrm{feedb}}(\mu,k)$ is an estimation of the acoustic attenuation from the mouth loudspeaker to the microphone on the one hand and of the attenuation of the feedback cancellation unit on the other hand. The estimated short-term power will be used within a Wiener-like attenuation characteristic that will be described in more detail in Section 3.3.5. With the feedback cancellation and the suppression of the residual feedback, a gain increase of about 5 to 10 dB can be achieved. Here, one should keep in mind that until now, no decorrelation approach was performed. When activating such units an additional gain increase can be realized, mainly by means of an improved performance of the cancellation filter. The decorrelation is described in Section 3.5.

3.3 Residual feedback and noise suppression

Beside feedback, also background and breathing noise are distortions that appear when using full-face communication masks. The breathing noise reduction relies mainly on a so-called *voice activity detection* (VAD) scheme, which is realized using a pattern recognizer approach that distinguishes between five classes:

- Side talkers
- Pause
- Breathing out
- Breathing in
- Speech

The spectrogram in Fig. 9 represents a microphone signal with the classes breathing out, pause, speech, and breathing in. The spectral content of the different classes is clearly visible. The characteristic of the class side talkers is spectrally close to the speech but with less energy. It usually includes talkers and speech from a so-called *land mobile radio* located in front of the mask. It is necessary to reduce the breathing noise because of its very high-sound pressure level. Background noise is some stationary noise like a water pump that can be estimated by conventional noise estimation schemes [21, 35] such as presented in Section 3.3.4. After estimating the power spectral densities the undesired signal components are suppressed by a time-varying spectral suppression method such as a Wiener filter (described in Section 3.3.5).

The differentiation among the individual classes is necessary since this decision will be used when controlling a so-called *comfort noise injection* (see Eq. (29)) on the one hand. Here, the signal is replaced by artificial noise whenever the classes breathing in, breathing out, or side-talkers are detected. Details will be described in Section 3.3.5. On the other hand, detailed knowledge about, e.g., the duration of breathing in and out phases as well as the signal power and the spectral contents of the individual phases can be used for health monitoring purposes. However, this aspect is not in the scope of this contribution and will not be described here, but it should motivate why two breathing classes instead of one are used.

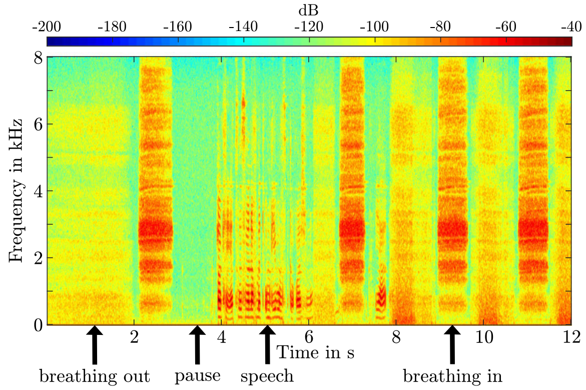


Fig. 9 Spectrogram of a microphone signal picked up by the masks communication system

3.3.1 Voice activity detection

Voice activity detection (VAD) is a widely explored field in speech processing. Several overviews show this, see [16, 22, 47] for example. In our case, we need a scheme that decides on the one hand rather fast and on the other hand is optimized for specific distortions (breathing noise) that are usually not part of *conventional* VAD schemes. The VAD scheme that we used here is implemented by means of a neural network [10]. This neural network distinguishes between the five signal classes mentioned in the last section. The aim of the VAD and the related units for signal manipulation is to keep all speech components and transmit them to the loudspeaker outputs, while suppressing all other signal components. The pattern recognition scheme presented in the next section uses mainly spectral envelope properties to classify between the classes; hence, the feature extraction elaborates properties from the frequency domain input signal $X_{\text{mic}}(\mu, k)$.

3.3.2 Feature extraction

The feature extraction is used to work out the most relevant properties out of the microphone spectrum $X_{\text{mic}}(\mu, k)$. These properties should represent the significant information utilizing a minimum set of features to reduce the computational complexity. The input signal is transformed into a set of absolute spectral values. Afterwards, a mel filterbank [6] with 10 mel bands is applied to reduce the amount of frequency supporting points while mimicking the frequency perception of the human ear [41, 50]. A simple loudness approximation is achieved by a logarithmic characteristic [17, 19]. As a result, a subset

$$\boldsymbol{X}_{\text{in}}(k) = \Big[X_{\text{in}}(0, k), \ldots, X_{\text{in}}(N_{\text{in}} - 1, k)\Big]^{\text{T}} \qquad (20)$$

of $N_{\text{in}} = 10$ features is generated out of the N_{DFT} complex spectral values $X_{\text{mic}}(\mu, k)$. Subsequently, the extracted features are processed by a neural network for classification.

3.3.3 Neural network

The neural network is based on the function of the brain, which consists of neurons [8, 13]. These neurons are connected together to weight the transition paths by w_{ij}, meaning the transition path from the neuron i to the neuron j [33]. The used neural network (see Fig. 10) comprises one input layer, one hidden layer, and one output layer. The input layer normalizes the features to the range of -1 to 1. The normalized input $\tilde{X}_{\text{in}}(m, k)$ is distributed from the corresponding neuron to every neuron of the hidden layer and multiplied with the associated weight $w_{\text{hid}}(i, n)$. Each weighted incoming signal is individually biased by $B_{\text{hid}}(i)$ before summing. Subsequently, $f_{\text{act}}(x)$ is a linear transfer function, with max and min limitations at $+1$ and -1. The low computational complexity of this functions has clear advantages in the fixed-point implementation. Furthermore, no significant performance degradation could be observed when comparing the recognition results when sigmoid and other typical activation functions (with larger computation load) were used. With this limited linear activation function

$$f_{\text{act}}(x) = \begin{cases} 1, & \text{if } x > 1, \\ \text{-1}, & \text{if } x < \text{-1}, \\ x, & \text{else}; \end{cases} \qquad (21)$$

the output of the hidden layer is generated:

$$X_{\text{hid}}(i, k) = f_{\text{act}}\left(B_{\text{hid}}(i) + \sum_{n=0}^{N_{\text{in}}-1} \tilde{X}_{\text{in}}(n, k)\, w_{\text{hid}}(i, n)\right), \quad \text{for } 0 \le i < N_{\text{hid}}. \qquad (22)$$

Here, N_{hid} is the number of neurons in the hidden layer, $w_{\text{hid}}(i, n)$ is the weight from the input n to the hidden layer i. The computation of the output vector

$$\boldsymbol{X}_{\text{out}}(k) = \Big[X_{\text{out}}(0, k), \ldots, X_{\text{out}}(N_{\text{out}} - 1, k)\Big]^{\text{T}} \qquad (23)$$

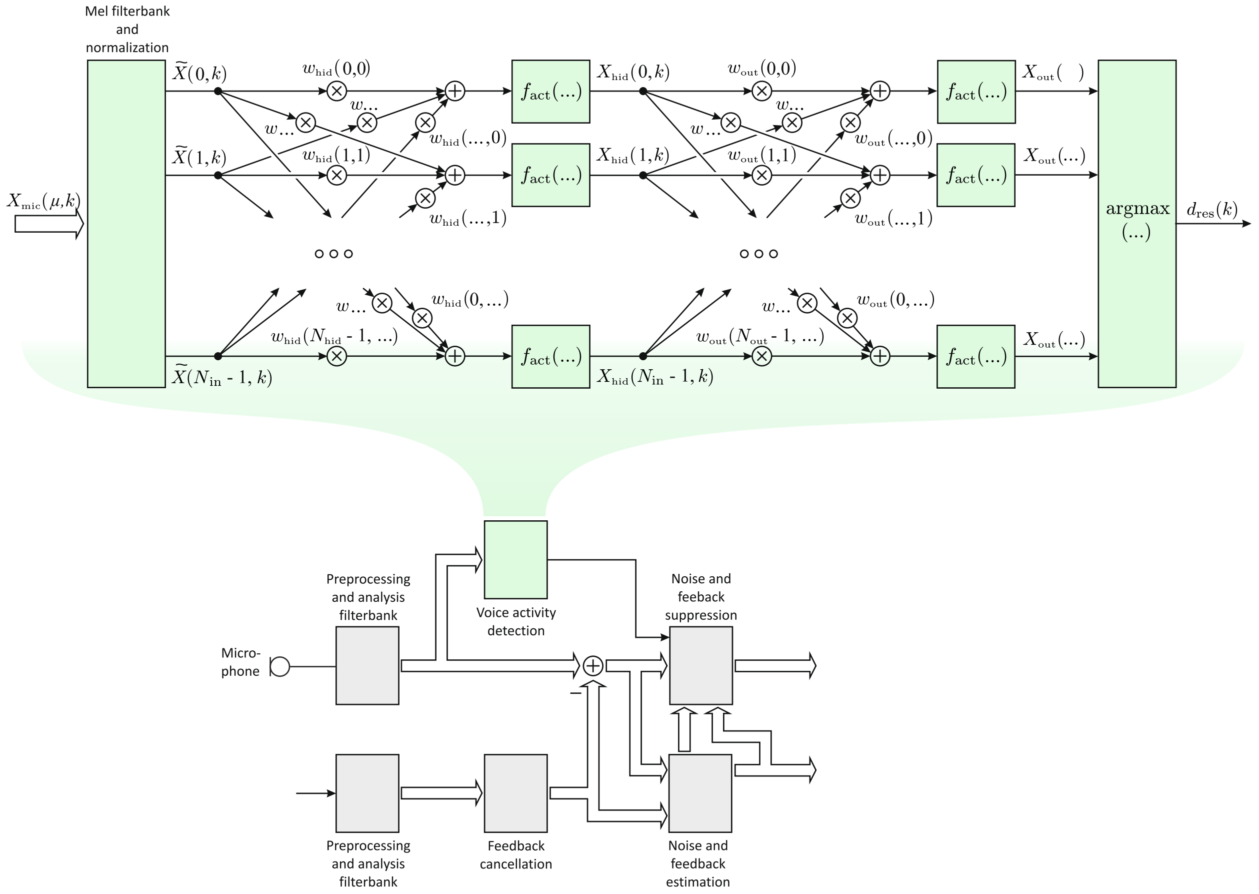
with the number of output elements N_{out}, is similar to the previous stage and is given by

$$X_{\text{out}}(j, k) = f_{\text{act}}\left(B_{\text{out}}(j) + \sum_{i=0}^{N_{\text{hid}}-1} X_{\text{hid}}(i, k)\, w_{\text{out}}(j, i)\right), \quad \text{for } 0 \le j < N_{\text{out}}. \qquad (24)$$

The weights from the hidden layer i to the output layer j are described by $w_{\text{out}}(j, i)$. This neural network is used to classify the features into $N_{\text{out}} = 5$ classes, which are summarized in the set $C = \{$side talkers, pause, breathing out, breathing in, speech$\}$. A distinction is made between 5 classes, since in the future post processing can be implemented, which can distinguish between the states and, for example, attenuating some classes more than others, or a time dependency between the classes could become interesting. With the recognition of inhalation and exhalation, it is also recognized in non-speech passages whether the wearer is still breathing and in which frequency one is breathing. This information can be provided to the team leader and he has an overview of the vitality of the troop members. If these variabilities are not needed, then only a distinction could be made between speech and non-speech. In order to detect the most likely class, the index of the maximum entry of the vector $\boldsymbol{X}_{\text{out}}(k)$ is determined by

$$d_{\text{res}}(k) = \underset{j \in C}{\operatorname{argmax}} \Big\{X_{\text{out}}(j, k)\Big\}. \qquad (25)$$

If speech is detected with $d_{\text{res}}(k)$, the signal will be processed normally and the background noise will be attenuated in the noise reduction. If a small inhalation passage of



Mel filterbank and normalization
$\widetilde{X}(0,k)$
$\widetilde{X}(1,k)$
$\widetilde{X}(N_{\text{in}}-1,k)$
$X_{\text{mic}}(\mu,k)$
$w_{\text{hid}}(0,0)$
$w_{\text{hid}}(1,1)$
$w_{\text{hid}}(\ldots,0)$
$w_{\text{hid}}(\ldots,1)$
$w_{\text{hid}}(0,\ldots)$
$w_{\text{hid}}(N_{\text{hid}}-1,\ldots)$
$w\ldots$
$f_{\text{act}}(\ldots)$
$X_{\text{hid}}(0,k)$
$X_{\text{hid}}(1,k)$
$X_{\text{hid}}(N_{\text{in}}-1,k)$
$w_{\text{out}}(0,0)$
$w_{\text{out}}(1,1)$
$w_{\text{out}}(\ldots,0)$
$w_{\text{out}}(\ldots,1)$
$w_{\text{out}}(0,\ldots)$
$w_{\text{out}}(N_{\text{out}}-1,\ldots)$
$X_{\text{out}}(\ )$
$X_{\text{out}}(\ldots)$
argmax (...)
$d_{\text{res}}(k)$
Preprocessing and analysis filterbank
Voice activity detection
Noise and feeback suppression
Micro-phone
Preprocessing and analysis filterbank
Feedback cancellation
Noise and feedback estimation

Fig. 10 Feed-forward neural network [20]

less than 0.5 sec occurs during speech, it is slightly attenuated so that the speech intelligibility of the sentence is improved. If the inhalation passage is longer than 0.5 sec, it is assumed that the sentence is not completed. When wearing a mask with SCBA it is not normal to be exhaled during speech and thus this condition needs to be applied only for inhalation. When pause is detected, this state is used to estimated the background noise. In the other cases, if no speech is detected with $d_{\text{res}}(k)$, the signal is completely muted for the all classes in the noise suppression and the amplifiers of the speakers are turned off to save power.

If the recognition of speech and exhalation in $d_{\text{res}}(k)$ is very similar and exhalation is detected, in the future, the perceived exhalation can be attenuated less in the noise suppression; thus, potentially less dropouts are generated in the speech passages and a better speech intelligibility can be achieved.

For the training a large database containing speech signals recorded in typical fire fighter environments was created and labeled. This database includes

- For the classes side talkers and pause data of approximately 1 h
- For the main classes breathing in, breathing out, and speech data of approximately 5 h

of data. Then, the training set is generated in a fixed-point feature extraction (bit exact as on the digital signal processor) and the results are bundled in a class wise way. One example of such a signal is depicted in Fig. 9. The training of the network was done using the back propagation algorithm [13, 20], in which the mean square error serves as the cost function:

$$E = \frac{1}{Q+1}\sum_{k=0}^{Q}\left[t(k)-a(k)\right]^2, \tag{26}$$

with the sets $\{\boldsymbol{X}_{\text{in}}(0), t(0)\}$, ..., $\{\boldsymbol{X}_{\text{in}}(Q), t(Q)\}$ in which $\boldsymbol{X}_{\text{in}}(k)$ is the input of the network and $t(k)$ indicates the associated target class for the training vector $\boldsymbol{X}_{\text{in}}(k)$ and the network output of the specific class $a(k)$. The optimization is performed by the gradient descent method.

Results of the VAD

The classification performance of the neural network is shown in Table 1. The input classes (the results detected by the neural network) are on the left of the table while the target classes are located on the top. The recognition rates of the target classes are shown in % and for each target class the overall recognition rate for the input class is given. The last row shows the average recognition rate. For proper recognition rates it is important to have at least a small confusion between breathing out and speech, because otherwise fricatives for example might be classified as breathing out and thus would be attenuated.

The recognition rate of 53% for the side talkers is not particularly good, but only 8% of side talkers is recognized as speech. The recognition of pauses is given by 70% and only 7% are wrongly classified as speech. Breathing in is recognized by 99% which is nearly perfect. The most important parts are breathing out and speech, where the recognition rate for breathing out is 83% and 12% of breathing out is recognized as speech. For the input class speech the recognition rate is 95% and 2% of speech is recognized as breathing out. It is essential that no speech is wrongly recognized, which would lead to attenuation of speech. The fact that a few parts of breathing out are audible is acceptable, as the breathing out signal has less energy than breathing in or speech, which can be seen in Fig. 9.

In the communication system, non-speech parts are muted always and only the speech is audible. The result of the VAD can be seen in Fig. 11 (upper part), showing the spectrogram of recorded microphone data of the communication system. The lower part of Fig. 11 shows the spectrogram of recorded data with the processing of the VAD and muting non-speech parts. The comparison of the figures shows that only the speech parts are retained.

3.3.4 *Background noise estimation*

Fire fighters often work in areas with background noise. Examples for noise sources are the engine of a fire truck or pumps that are necessary to bring the water to the desired places. This noise is picked up by the microphone, amplified, and played back via the loudspeakers. Hence, it leads to an increased overall noise level at the ears of the fire fighter. To avoid this, the noise must be reduced before playback. The noise suppression itself is performed using a Wiener-type filter. However, this filter requires an estimate of the power spectral density (PSD) of the noise signal $P_b(\mu, k)$. We target here only stationary noise sources, since the main non-stationary noise sources (breathing in and out) are already covered by the network approach. The stationary estimation is done by a simple two-stage procedure. First, we compute the squared magnitude of the input spectrum $X_{\text{mic}}(\mu, k)$ and afterwards we smooth it over the time, which is done by a first-order IIR filter [51]:

$$P_x(\mu, k) = \alpha_{\text{sm}} \left| X_{\text{mic}}(\mu, k) \right|^2 + (1 - \alpha_{\text{sm}})\, P_x(\mu, k-1). \tag{27}$$

The smoothing constant is set to $\alpha_{\text{sm}} = 0.1482$. As a second stage, the computation of the estimated noise PSD $P_b(\mu, k)$ is done by comparing the smoothed PSD estimation $P_x(\mu, k)$ with $P_b(\mu, k-1)$ and by updating the estimated PSD appropriately afterwards. When $P_x(\mu, k)$ is larger than $P_b(\mu, k-1)$, the estimated PSD is multiplied by an increase constant Δ_{inc} and otherwise by a decrease constant Δ_{dec}:

$$P_b(\mu, k) = \begin{cases} \Delta_{\text{inc}}\, P_b(\mu, k-1), \\ \quad \text{if } P_x(\mu, k) > P_b(\mu, k-1), \\ \Delta_{\text{dec}}\, P_b(\mu, k-1), \\ \quad \text{if } P_x(\mu, k) \leq P_b(\mu, k-1). \end{cases} \tag{28}$$

Thereby, Δ_{dec} is chosen larger than Δ_{inc}, because the noise estimation only follows temporally stationary noise and if Δ_{inc} is large the estimator also follows for example speech. In the other case when Δ_{dec} is too large, the noise estimation decreases too fast and the estimation level is too low. The increment constant Δ_{inc} for this setup was set to 1.0005 (which corresponds to an increase of about 1 dB per second) and the decrement constant Δ_{dec} to 0.9986 (corresponding to a decrease of about 3 dB per second).

Table 1 Confusion matrix of the classification approach based on a neural network

		Target						
		Side talker	Pause	Breathing out	Breathing in	Speech	Recognition	
Input	Side talker	53%	23%	14%	2%	8%	53%	47%
	Pause	14%	70%	8%	1%	7%	70%	30%
	Breathing out	2%	2%	83%	1%	12%	83%	17%
	Breathing in	< 1%	< 1%	< 1%	99%	< 1%	99%	1%
	Speech	2%	< 1%	2%	< 1%	95%	95%	5%
				Average recognition rates			79%	21%

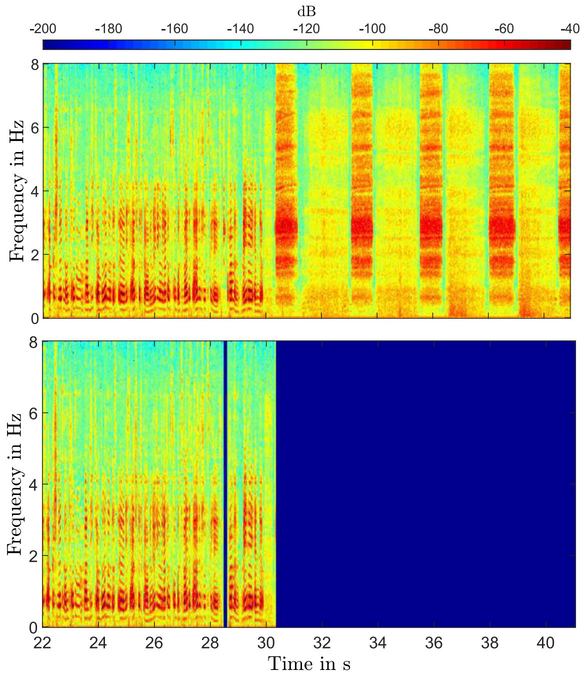


Fig. 11 Spectrogram of recorded microphone data from the communication system without (upper part) and with VAD (lower part), that mutes non-speech passages

3.3.5 Short-term spectral attenuation and comfort noise

Finally, after having for all types of distortions PSD estimations available and by using the VAD results the suppression of background noise and feedback is performed by either multiplying the error spectrum $E(\mu,k)$ obtained in the feedback cancellation stage by the frequency-dependent attenuation factor $H_{\text{att}}(\mu,k)$ or by replacing it with so-called *comfort noise* to get the enhanced spectrum

$$X_{\text{enh}}(\mu,k) = \begin{cases} E(\mu,k)\, H_{\text{att}}(\mu,k), \\ \quad \text{if speech or pause was detected,} \\ C(\mu,k), \\ \quad \text{else.} \end{cases} \tag{29}$$

The computation of $H_{\text{att}}(\mu,k)$ is performed by a modified Wiener filter [25]

$$H_{\text{att}}(\mu,k) = \max\Bigg\{H_{\min}, 1 - \ldots \tag{30}$$

$$\ldots \frac{\beta_b\, P_b(\mu,k) + \beta_f\, P_f(\mu,k)}{|E(\mu,k)|^2}\Bigg\},$$

where $P_b(\mu,k)$ and $P_f(\mu,k)$ are the estimated PSDs of the background noise (Eq. (28)) and feedback (see Eq. (19)), respectively. They are weighted with overestimation factors β_b for the background noise and β_f for the feedback. These factors adjust the bias of the estimation and thus control the *aggressivity* of the characteristic. The result of the summation is divided by the (instantaneous) PSD of the microphone signal after feedback cancellation $|E(\mu,k)|^2$. When the filter only attenuates some subbands for a short time, so-called *musical tones* can appear which is an unwanted distortion. A maximum attenuation factor $H_{\min}$ and appropriate overestimation by means of the factors β_b and β_f, respectively, are used to avoid this phenomenon. In this setup, the maximum attenuation factor $H_{\min}$ is set to − 9 dB, the overestimation factor β_b to 2 dB and the overestimation factor β_f to 1 dB. The stationary parts of the background noise can be suppressed by 6 to 12 dB and the feedback suppression leads to an additional amplification gain of about 2 dB.

The maximum attenuation $H_{\min}$ would not be sufficient to suppress distortions that originate from breathing in or out. Thus, the attenuation based mechanism is supported by the injection of stationary noise. This noise is usually produced in the complex subband domain by white Gaussian noise generators with zero mean and unit variance, mutually independent for the real and imaginary part of each subband. Before inserting the noise, its power is adjusted by multiplication with $\sqrt{P_b(\mu,k)} \cdot H_{\min}$. This adjustment should match the statistical properties of the artificial noise to the residual noise after applying maximum attenuation.

3.4 Gain control and equalization

After suppressing noise and feedback typically two further frequency-domain weighting schemes can be applied (see Fig. 12). On one hand an equalization characteristic can be applied by means of weighting the input spectrum with appropriate gain factors $H_{\text{eq}}(\mu)$. This can be either achieved in the time-domain (usually for narrow notch characteristics, details in Section 3.7) or in a more smooth way in the frequency domain. On the other hand. the background noise power can be mapped via a noise-to-gain characteristic onto an amplification of the input spectrum. Such a characteristic is usually applied to the signal coming from the radio unit before being played back via the ear loudspeakers. Furthermore, the peak power of the microphone signal can be tracked. An amplification/compression characteristic can be applied to normalize the output signal. This is usually done before emitting the signals of the fire fighters via the tactical radio link (in the uplink path). Quite often such characteristics are only time but not frequency-dependent, leading here to gain/attenuation factors $H_{\text{agc}}(k)$. Since such characteristics are well established units, we will not go into the detail of how to adjust them and refer here to the literature: [56] gives a very good overview about such characteristics.

In the path that connects the microphone signal with the mouth loudspeakers the spectral modification can be

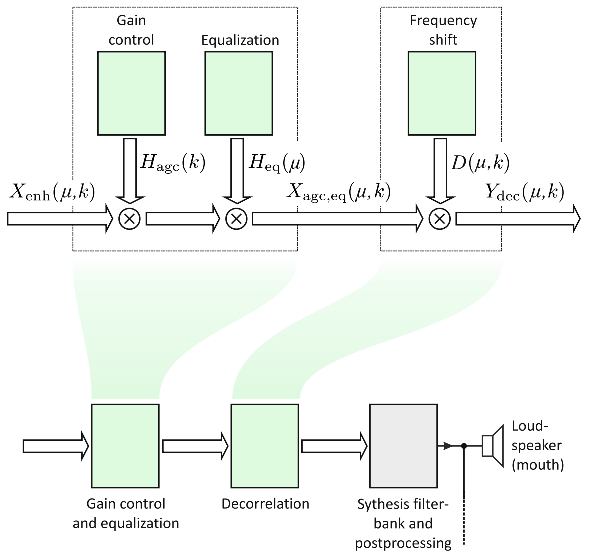

Gain control
Equalization
Frequency shift
$H_{\mathrm{agc}}(k)$
$H_{\mathrm{eq}}(\mu)$
$D(\mu,k)$
$X_{\mathrm{enh}}(\mu,k)$
$X_{\mathrm{agc,eq}}(\mu,k)$
$Y_{\mathrm{dec}}(\mu,k)$
Loud-speaker (mouth)
Gain control and equalization
Decorrelation
Sythesis filter-bank and postprocessing

Fig. 12 Structure of the gain control and equalization as well as of the decorrelation

described as

$$X_{\mathrm{agc,eq}}(\mu,k) = X_{\mathrm{enh}}(\mu,k)\, H_{\mathrm{agc}}(k)\, H_{\mathrm{eq}}(\mu). \qquad (31)$$

In the other pathes, similar schemes are applied. Since the reduction of large signal peaks has also impact on the energy consumption we will describe time-domain range compression in Section 3.7.1.

3.5 Decorrelation

Before converting the enhanced spectra back to the time-domain a decorrelation stage should be applied (see Fig. 12). Such decorrelation is necessary in order to help the feedback cancellation filters to converge to the desired solution [39]. Beside this effect, also the maximum stable gain that can be achieved by the communication system can be improved with the scheme presented now.

The decorrelation is implemented by a frequency shift of each subband in the frequency domain [23, 28, 43, 55]. The shift in Hz is given by $f_{\mathrm{shift}}(\mu)$ and the phase adjustment is realized by $e^{j2\pi k f_{\mathrm{shift}}(\mu)}$. This phase adjustment depends on the frame index k; hence, the phase adjustment $D(\mu,k)$ has to be updated each frame:

$$D(\mu,k) = D(\mu,k-1)\, e^{j2\pi f_{\mathrm{shift}}(\mu)}. \qquad (32)$$

The update is done by a complex multiplication with constant factors. The initialization of the phase adjustment is performed according to

$$D(\mu,0) = 1. \qquad (33)$$

with $X_{\mathrm{agc,eq}}(\mu,k)$ representing the input of the frequency shift and $Y(\mu,k)$ being the output it yields to following equation:

$$Y_{\mathrm{dec}}(\mu,k) = X_{\mathrm{agc,eq}}(\mu,k)\, D(\mu,k), \qquad (34)$$

where the output signal is shifted by $f_{\mathrm{shift}}(\mu)$ Hz. For the decorrelation it is only necessary to shift a few Hz, but if a larger shifting is possible, the resonance of the feedback is shifted more and a higher stable gain can be chosen. We suggest to apply the following shifts:

- Below 1 kHz: shift about 5 Hz,
- 1 to 3 kHz: shift about 10 Hz,
- above 3 kHz: shift about 20 Hz.

Since all the shift frequencies are smaller than the distance between the center frequencies of neighboring subbands, we need no subband index increment when shifting. However, for larger shifts (or filter banks with more channels) this has to be taken into account. With the setup presented above, an additional gain of up to 20 dB can be achieved—mainly due to improved performance of the feedback cancellation approach.

3.6 Synthesis filterbank and postprocessing

In the last signal processing stage, we go back to the time-domain using synthesis filterbanks and perform some final time-domain postprocessing. Such postprocessing includes deem-phase filters (the inverse of the preem phase filter that was applied before the analysis filterbank), time-domain equalization, dynamic range compression, and a limiter for each output channel in order to equalize the frequency response, compress the signals into the desired dynamic range, and afterwards to limit the signals. The detailed signal processing structure is depicted in Fig. 13.

As in the previous sections, we will describe the algorithms for the path that connects the microphone input signal $x_{\mathrm{mic}}(n)$ with the mouth loudspeaker output signal $y_{\mathrm{mouth}}(n)$. The second synthesis filterbank and postprocessing unit (see Fig. 5) is computed equivalently.

3.6.1 Synthesis filterbank

The synthesis filterbank is mainly the inverse structure of the analysis filter bank, that was described in Section 3.1. We have used here an overlap-add based structure with an appropriated-scaled Hann window that leads for the parameter setup to a delay of 16 ms. This value can easily be reduced to 8 ms if the method according to [54] is used.

Using overlap-add-based structures instead of overlap-save based ones have the advantage, that artifacts that appear due to gain changes from frame to frame are smoothed due to the overlapped adding stage of the windowed frames. A second advantage is the better aliasing properties. This allows to avoid projection stages within the adaptive filters. However, the price to be paid for this are the smaller frameshift which removes most of the computational savings in the adaptive filter stage.

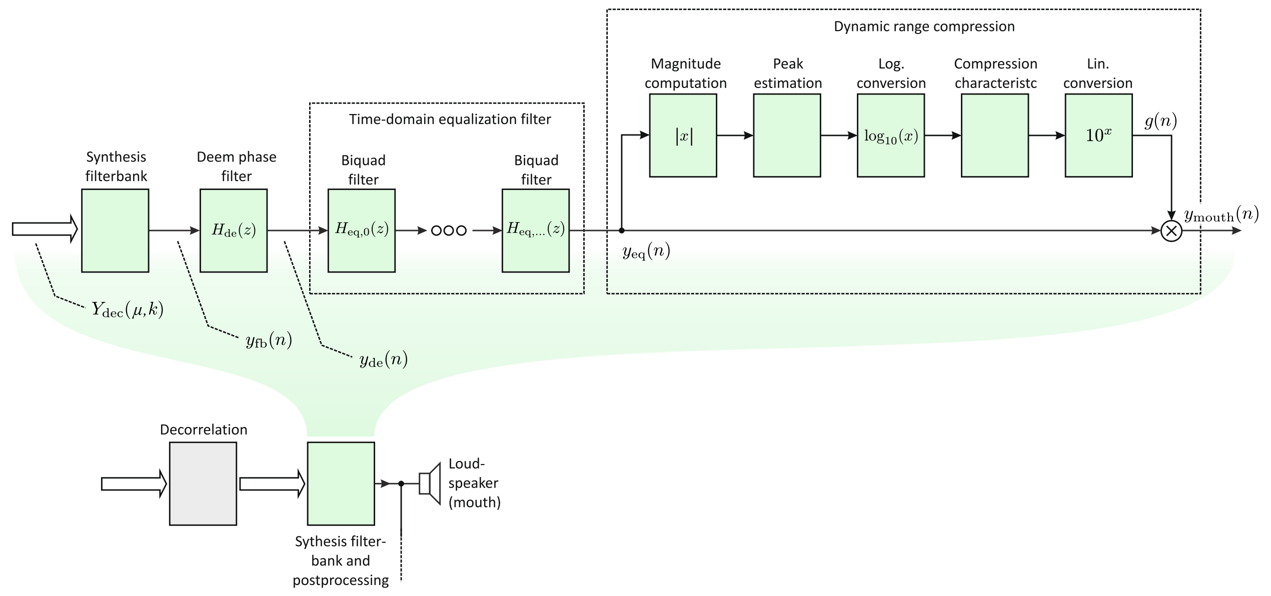


Fig. 13 Postprocessing—overview

The resulting time-domain signal of the synthesis filterbank will be denoted here as $y_{\mathrm{fb}}(n)$.

3.6.2 Deem-phase filter

The first time-domain postprocessing part within the postprocessing framework is the deem-phase filter that inverts its preem-phase counterpart described in Section 3.1. Since this was a first-order FIR filter, we have to use now a first-order IIR filter according to

$$y_{\mathrm{de}}(n) = y_{\mathrm{fb}}(n) + \beta_{\mathrm{pre\text{-}de}}\, y_{\mathrm{de}}(n-1). \qquad (35)$$

3.7 Time-domain equalization

The time-domain equalizer is computed as a cascade of second order recursive filters (so-called *biquad filters*) according to [31, 32, 37, 56]:

$$H_{\mathrm{eq}}(z) = \prod_{i=0}^{N_{\mathrm{eq}}-1} H_{\mathrm{eq},i}(z). \qquad (36)$$

Each subfilter has the following transfer function:

$$H_{\mathrm{eq},i}(z) = \frac{b_{0,i} + b_{1,i}\, z^{-1} + b_{2,i}\, z^{-2}}{1 + a_{1,i}\, z^{-1} + a_{2,i}\, z^{-2}}. \qquad (37)$$

These biquad filters are used instead of a finite impulse response-filter (FIR filter), because the IIR filter structure allows for a significantly lower order compared to FIR structures for comparable filter effects. The disadvantage of the IIR filter is the frequency selective group delay.

The time-domain equalizer can be adjusted to form a high-pass, low pass, peak, notch, and/or shelving filter. In our setup, each output channel may comprise up to $N_{\mathrm{eq}} =$ 8 filters. The filters are used especially for the voice-amplification loudspeakers in the front part of the mask, for example, to boost higher frequencies or to set a notch filter attenuating a frequency where feedback occurs. The loudspeakers in front of the mask should preferably play frequencies that are attenuated by the mechanical mask. Hence, frequencies above 2.5 kHz are boosted by a shelving filter, because of the high attenuation caused by the mask. Frequencies below 1 kHz cannot be transmitted, because of size constraints of the loudspeaker and the resonance volume.

3.7.1 Dynamic range compression and limitation

Dynamic range compression (DRC) is an algorithm that maps the dynamic range of an input signal to a (usually) smaller range for the output signal [15, 42, 56]. DRC is necessary for a communication system, because the dynamic range has to limit the sound pressure level of the ear speaker. This ensures that the ear is not damaged and that quiet passages are amplified to a desired loudness. The DRC algorithm can be split into five parts:

- Compute the absolute magnitude of the input signal
- Estimate the peak level
- Compute the logarithm
- Get the gain from the compression characteristics
- Apply the gain to the input signal

The signal flow graph is shown in the right part of Fig. 13. The peak estimation is implemented by a first-order IIR filter with a time-variant smoothing constant $\alpha_{\mathrm{sm}}(n)$ that is applied to the magnitude of the input signal $|y_{\mathrm{eq}}(n)|$:

$$\overline{y}_{\text{mag}}(n) = \alpha_{\text{sm}}(n)\,\left|y_{\text{eq}}(n)\right| + ... \qquad (38)$$
$$... + \left(1 - \alpha_{\text{sm}}(n)\right)\overline{y}_{\text{mag}}(n-1).$$

The smoothing constant $\alpha_{\text{smo}}(n)$ is defined by the so-called *attack time constant* α_{att}, if the magnitude signal $|y_{\text{eq}}(n)|$ is larger than the previous smoothed magnitude signal $\overline{y}_{\text{mag}}(n-1)$. Otherwise, the so-called *release time constant* α_{rel} is applied:

$$\alpha_{\text{sm}}(n) = \begin{cases} \alpha_{\text{att}}, & \text{if } \left|y_{\text{eq}}(n)\right| > \overline{y}_{\text{mag}}(n-1), \\ \alpha_{\text{rel}}, & \text{else.} \end{cases} \qquad (39)$$

The attack time is much shorter than the release time; hence, this time-variant filter can be interpreted as a peak tracker. For the communication system the attack time α_{sm} is set to 20 ms and the release time α_{rel} to 200 ms. Finally, the logarithm is computed to generate the input value for the compressor characteristic:

$$\overline{y}_{\text{log}}(n) = 20\,\log_{10}\left(\overline{y}_{\text{mag}}(n)\right). \qquad (40)$$

Please note that the computation of the logarithm as well as the following steps can be computed in a subsampled manner, in order to save computational complexity. Due to the smoothing according to Eq. (39), this has nearly no impact on the signal quality.

Figure 14 shows the compressor characteristic of the ear-loudspeakers and Fig. 15 shows the compressor characteristic of the loudspeakers in front of the mask. It depicts the input signal $\overline{y}_{\text{log}}(n)$ in dB on the x axis and the desired output signal $\overline{y}_{\text{des,log}}(n)$ on the y axis. The blue curve represents the desired compressor characteristic and the bisectrix colored in gray represents a linear transmission.

The gain $g_{\text{log}}(n)$ in dB is computed by subtracting the input level from the output level to get the *gap* between both:

$$g_{\text{log}}(n) = y_{\text{des,log}}(n) - \overline{y}_{\text{log}}(n). \qquad (41)$$

The desired output signal has to be multiplied by this gain in the linear domain; hence, it is transformed to a linear gain factor:

$$g(n) = 10^{g_{\text{log}}(n)/20}. \qquad (42)$$

The input signal $y_{\text{eq}}(n)$ is multiplied by the linear gain $g(n)$ to get the desired output signal $y_{\text{mouth}}(n)$:

$$y_{\text{mouth}}(n) = y_{\text{eq}}(n)\cdot g(n). \qquad (43)$$

As overshoots could occur, it is necessary to apply a limiter additionally to ensure a limited sound pressure level at the output to protect the fuse of the device.

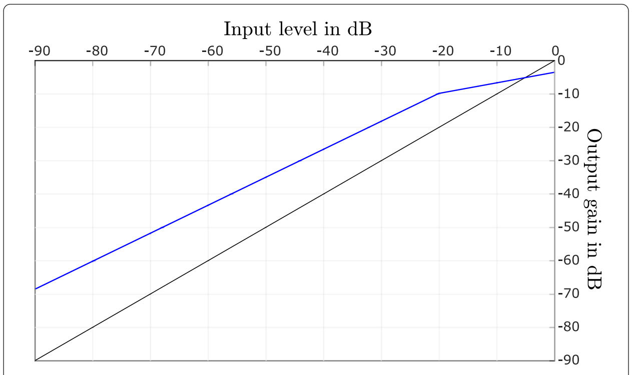


Fig. 14 Compressor characteristic of the ear loudspeakers

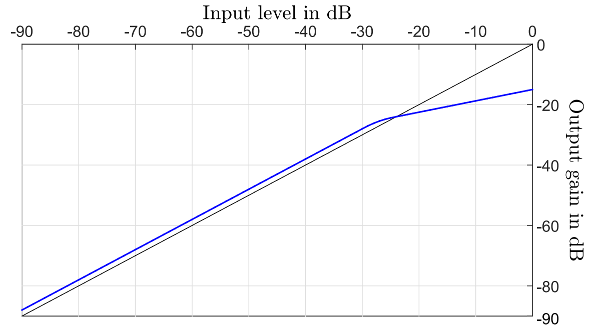


Fig. 15 Compressor characteristic of the loudspeakers in front of the mask

4 Results and discussion

For the evaluation of the communication system, a subjective test as well as objective measurements were performed. In addition, the computational complexity was analyzed and measured when all algorithms are active. The subjective test is realized as a modified rhyme test. The objective measurements are determined by measuring the transmission characteristic of the passive mask in comparison to the mask with an activated communication system.

4.1 Modified rhyme test

The data for the modified rhyme test (MRT) was recorded with two artificial heads, one simulating the speaker and the other one the listener [14, 29]. The heads are surrounded by ambient loudspeakers as shown in Fig. 16.

On the left side, one can see an artificial head from GRAS (KEMAR 45), which simulates the listener. The two ear microphones produce binaural recordings. Furthermore, the left and right loudspeakers are installed to generate ambient noise. On the other side of the GRAS artificial head, a second torso is placed, which is also an artificial head, a DRÄGER Quaestor head [1], which fits to the contours of typical masks. Conventional heads such as the KEMAR (but also others) are a bit too small for typical mask sizes.

In the test, a mask with a communication system was mounted on the Quaestor head, such that in the test the

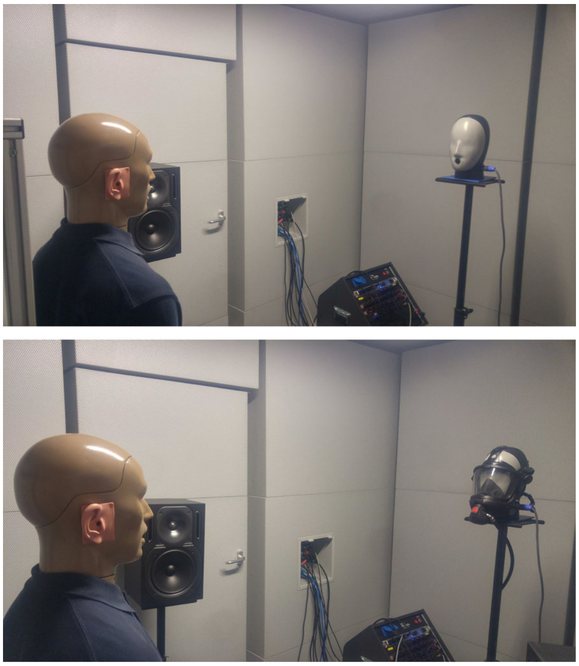


Fig. 16 Evaluation setup with and without mask and communication system

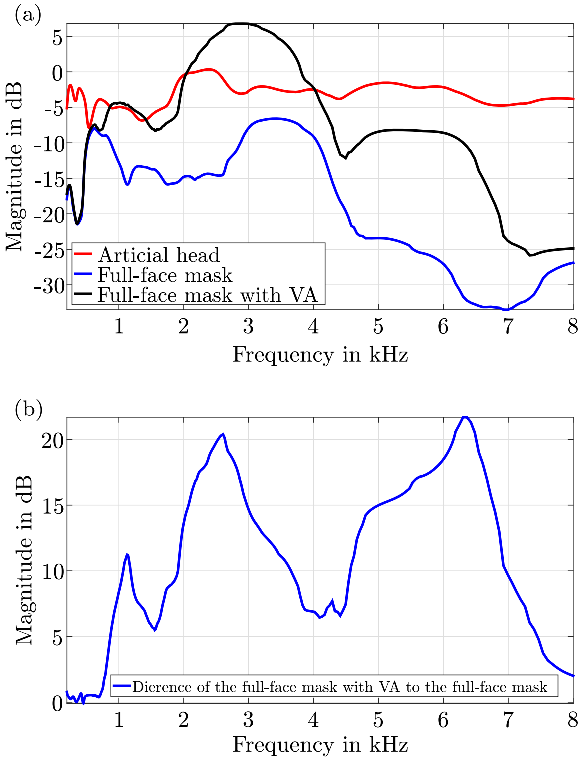


Fig. 17 Evaluation results. **a** Power spectral densities of the full-face mask FPS 7000 without a communication system (blue), with an activated voice amplifier of the communication system FPS-COM 7000 (black), and the artificial head without a full-face mask (red). **b** Difference of the power spectral densities from the full-face mask FPS 7000 with an activated voice amplifier of the communication system FPS-COM 7000 to the full-face mask FPS 7000 without a communication system

passive mask and the mask with activated communication system can be compared in a fair manner. The ambient noise is represented by white noise, which sounds very similar to a so-called *c-pipe*, a protective fan and similar equipment. The SNR was adjusted such that with the passive mask a value of 0 dB at the ears of the listeners was achieved. When the communication system is activated the SNR increases according to the transmission characteristic as shown in Fig. 17.

The modified rhyme test was performed according to [14], where in this test eight samples for each of the seven rhyme classes are used. Thus, 56 samples have been evaluated for each variant. The listening test was attended by 15 people and an error rate (ER) for the passive mask of 23.27% was achieved. With the activated communication system an ER of 17.06% was achieved. Thus, the activated communication system improves the ER by 5.31%. This is only a small improvement, but this can be crucial in terms of clarity in use. The participants have indicated that sibilants are better understood. These can be for example very important in distinguishing between words that differ only by the (plural) “s” at the end of the word. With this difference, the chief of operations has to decide how much support she/he coordinates to the individual fire fighters.

4.2 Transmission characteristic of the communication system

The described full-face mask without a voice-amplification unit has a different frequency response compared to with attached communication system—as shown in Fig. 17. To show the differences, power spectral densities were measured. The microphone for those measurements was located 1 m in front of the artificial head. The red curve, shown in part a of Fig. 17, was measured for a setup without any mask. This power spectral density can be interpreted as a reference. The measurements were made with white noise excitation signal, sampling rate of 48 kHz and had a length of 64 k samples. The impulse response was obtained by means of an NLMS algorithm. However, keep in mind, that with this reference setup also no protection is present for the fire fighter.

Using a mask without a communication unit significant attenuation (blue curve) can be measured. It starts around 1.5 kHz and the attenuation of frequencies above 2 kHz is really strong. The black curve shows the full-face mask with an activated communication system including all the processing stages presented in this contribution. The communication system starts to transmit above 1 kHz having the highest amplification of approx. 22 dB at 3 kHz.

This amplification is only possible by applying the algorithms described before.

4.3 Computational complexity of the communication system

The computational complexity is measured with respect to a 120 MHz digital signal processor with fixed-point arithmetic. All algorithms together have a computational complexity in total of 50% on this processor, meaning that the entire system runs in real-time and requires about 60 Mips (million [fixed-point] instructions per second). In more detail the filterbank uses approx. 1.5%, the VAD 1%, the automatic gain control and equalization less than 1%, the feedback cancellation approx. 25% , the noise and feedback estimation 1%, the decorrelation less than 1%, the time-domain equalization filter (8 biquad stages) requires 1% per channel, and the dynamic range compression uses 1% per channel. In terms of the efficiency, some algorithms could still be optimized, but they are sufficient in terms of the workload of the digital signal processor and have not been considered further.

5 Summary and conclusion

Full-face masks are used to ensure clean air supply for fire fighters. However, when wearing such masks the speech signals of the wearer are strongly attenuated due to the hermetical sealing of the masks—the communication is impaired. For this reason, communication systems can be used to clean and amplify the speech signal, which leads to an enhancement of the communication if applied appropriately. The presented communication system is able to improve the communication drastically. However, this is only possible by applying a *cocktail* of enhancement stages that have to be adjusted and optimized in a mutual manner.

Further improvements can still be made in different areas. This could be, for example, the extension of the pattern recognition scheme. Long short-term memory approaches with deep neural networks could be useful here—this has not yet been analyzed for reasons of complexity and training data.

With respect to the loudspeakers of the ears and the mouth, the dynamic range compression could become frequency-dependent. This allows to respond adaptively and frequency selectively to background noise variations [45]. This might improve the speech intelligibility in different situations. To estimate the background noise for this purpose, a microphone would be needed that is located on the outside of the communication system.

Another improvement could be the usage of non-linear signal processing to reconstruct the harmonics of the speech, which are attenuated by the mask characteristics. This approach could improve the speech intelligibility on the mouth loudspeaker and the ear loudspeakers [11].

Finally, the performance of the feedback cancellation can be improved. This could be achieved either by improving the quality of the involved transducers (in order to reduce non-linear effects) and by extending the processing structures to non-linear approaches, such as Volterra filters. However, this would lead to a significant increase of computational complexity. Anyhow, the computational power of embedded hardware is permanently increasing. Thus, it might just be a question of time until such solutions would be realizable.

Abbreviations

DFT: Discrete Fourier transformation; DRC: Dynamic range compression; FFT: Fast Fourier transformation; FIR: Finite impulse response; IIR: Infinite impulse response; MRT: Modified rhyme test; NLMS: Normalized least mean square; PSD: Power spectral density; SCBA: Self-contained breathing apparatus; SNR: Signal-to-noise ratio; SPL: Sound pressure level; STI: Speech transmission index; VAD: Voice activity detection

Acknowledgements

The authors would like to thank all participants of the subjective tests.

Authors' contributions

MB has conducted the research and analyzed the data. MB and GS authored the paper. All authors read and approved the final manuscript.

Funding

No funding was given for the presented contents of this contribution.

Availability of data and materials

The data that support the findings of this study are available from Dräger Safety AG & Co. KGaA but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Dräger Safety AG & Co. KGaA.

Competing interests

The authors declare that they have no competing interests.

Received: 30 December 2018 Accepted: 7 November 2019
Published online: 12 December 2019

References

1. https://www.draeger.com/en_uk/Products/Quaestor-7000. Accessed 11 Mar 2018
2. https://www.draeger.com/en_uk/Products/Panorama-Nova. Accessed. 11 Mar 2018
3. https://www.draeger.com/en_uk/Products/FPS-7000. Accessed 11 Mar 2018
4. https://www.draeger.com/en_uk/Products/FPS-COM-7000. Accessed 11 Mar 2018
5. J. Benesty, D. R. Morgan, M. M. Sondhi, A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation. IEEE Trans Speech Audio Process. **6**(2), 156–165 (1998)
6. J. Benesty, Y. Huang, *Springer Handbook of Speech Processing*. (Springer, 2008)
7. J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. (Springer, Berlin/Heidelberg, 2013)
8. C. M. Bishop, *Pattern Recognition and Machine Learning*. (Springer, 2006)
9. P. Bloomfield, *Fourier Analysis of Time Series: An Introduction*. (Wiley, 2004)
10. M. Brodersen, A. Volmer, M. Romba, G. Schmidt, Sprachaktivitätserkennung mittels eines Mustererkenners für Atemschutzmasken. Proc. DAGA (2015)
11. P. Bulling, K. Linhard, A. Wolf, G. Schmidt, A. Theiß, M. Gimm, Nichtlineare Kennlinien zur Verbesserung der Sprachverständlichkeit in geräuschbehafteter Umgebung. Proc. DAGA (2016)
12. P. Bulling, K. Linhard, A. Wolf, G. Schmidt, Stepsize control for acoustic feedback cancellation based on the detection of reverberant signal periods and the estimated system distance. Proc. Interspeech, 176–180 (2017)

13. A. Cichocki, R. Unbehauen, *Neural Networks for Optimization and Signal Processing*. (Wiley, 1993)
14. N. Dillier, T. Spillmann, Deutsche Version der Minimal Auditory Capability (MAC)-Test-Batterie: Anwendungen bei Hö (1992)
15. M. M. Dimitrios Giannoulis, J. D. Reiss, Digital dynamic range compressor design – a rutorial and analysis. J. Audio Eng. Soc. **60**(6) (2012)
16. M. Espi, S. Miyabe, T. Nishimoto, N. Ono, S. Sagayama, in *Proc. of Spoken Language Technology Workshop (SLT), Berkeley, California, USA*. Analysis on Speech Characteristics for Robust Voicea Ativity Detection (IEEE, 2010). https://doi.org/10.1109/slt.2010.5700838
17. F. Eyben, Real-time speech and music classification by large audio feature space extraction (2015). https://doi.org/10.1007/978-3-319-27299-3
18. N. Faraji, R. C. Hendriks. Noise power spectral density estimation for public address systems in noisy reverberant environments (Proc. IWAENC, Aachen, 2012), pp. 1–4
19. H. Fastl, E. Zwicker, *Psychoacoustics: Facts and Models*. (Springer Science and Business Media, 2007)
20. J. Feldman, R. Rojas, *Neural Networks: A Systematic Introduction*. (Springer, 1996)
21. T. Gerkmann, R. C. Hendriks, Unbiased MMSE-based noise power estimation with low womplexity and low tracking delay. IEEE Trans. Audio. Speech. Lang. Process. **20**(4), 1383–1393 (2012)
22. S. Graf, T. Herbig, M. Buck, G. Schmidt, Features for voice activity detection: a comparative analysis. EURASIP J. Adv. Sig. Process. **91**, 1–15 (2015)
23. M. Guo, S. H. Jensen, J. Jensen, S. L. Grant, in *Acoustic Feedback Cancellation, Prof. EUSIPCO*. On the Use of a Phase Modulation Method for Decorrelation, (2012), pp. 2000–2004
24. V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, U. Rass, Signal processing in high-end hearing aids: state of the art, challenges, and future trends. EURASIP J. Adv. Sig. Process. (2005)
25. E. Hänsler, G. Schmidt, *Acoustic Echo and Noise Control*. (Wiley, 2004)
26. E. Hänsler, G. Schmidt, *Speech and Audio Processing in Adverse Environments*. (Springer, 2008)
27. S. Haykin, *Adaptive Filter Theory, 4th edition*. (Prentice Hall, 2001)
28. J. Herre, H. Buchner, W. Kellermann, Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement. Proc. ICASSP. **1**, 17–20 (2007)
29. D. Howard, J. Angus, *Acoustics and Psychoacoustics*. (Taylor & Francis, 2013)
30. International Electrotechnical Commission, Objective rating of speech intelligibility by speech transmission index. IEC 60268-16 (2011)
31. S. M. Kuo, S. M. K. Bob, H. Lee, *Real-time Digital Signal Processing: Implementations, Applications, and Experiments with the TMS320C55X*. (Tsinghua University Press, 2001)
32. E. Lai, *Practical Digital Signal Processing*. (Elsevier Science, 2003)
33. L. J. Landau, Concepts for neural networks: a survey, perspectives in neural computing (2012)
34. C. Lüke, H. Özer, G. Schmidt, A. Theiß, J. Withopf, in *Proc. 5th Biennial Workshop on DSP for In-Vehicle Systems*. Signal Processing for In-car Communication Systems, (2011)
35. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)
36. A. Mader, H. Puder, G. Schmidt, Step-size control for acoustic echo cancellation filters – an overview. Sig. Process. **80**(9), 1697–1719 (2000)
37. H. Malepati, *Digital Media Processing: DSP Algorithms Using C*. (Elsevier Science, 2010)
38. National Fire Protection Association (NFPA), Standard on open-circuit self-contained breathing apparatus (SCBA) for emergency services, NFPA 1981, edition 2013 (2013)
39. H. Puder, B. Beimel, in *Proc. EUSIPCO, Vienna, Austria*. Controlling the adaptation of feedback cancellation filters – problem analysis and solution approaches, (2004)
40. K. Rao, D. N. Kim, J. J. Hwang, *Fast Fourier Transform: Algorithms and Applications*. (Springer, 2011)
41. D. R. Raichel, *The Science and Applications of Acoustics*. (Springer New York, 2006)
42. J. D. Reiss, A. McPherson, *Audio Effects: Theory, Implementation and Application*. (Taylor and Francis, 2014)
43. L. Romoli, S. Cecchi, F. Piazza, A combined approach for channel decorrelation in stereo acoustic echo cancellation exploiting time-varying frequency shifting. IEEE Sig. Process. Lett. **20**(7), 717–720 (2013)
44. A. H. Sayed, *Fundamentals of Adaptive Filtering*. (Wiley-IEEE Press, 2003)
45. H. Schepker, J. Rennies, S. Doclo, Speech-in-noise enhancement using amplification and dynamic range compression controlled by the speech intelligibility index. **138**(5), 2692–2706 (2015). https://doi.org/10.1121/1.4932168
46. G. Schmidt, T. Haulick, Signal processing for in-car communication systems. Sig. Process. **86**(6), 1307–1326 (2006)
47. M. Van Segbroeck, A. Tsiartas, S. S. Narayanan. A robust Rrontend for VAD: exploiting contextual, discriminative and spectral cues of human voice, (2013)
48. H. J. M. Steeneken, T. Houtgast, A physical method for measuring speech transmission quality. J. Acoust. Soc. Am. **67**(1), 318–326 (1980)
49. H. J. M. Steeneken, T. Houtgast, A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in Auditoria. J. Acoust. Soc. Am. **77**(3), 1069–1077 (1985)
50. S. S. Stevens, J. Volkmann, E. B. Newman, A scale for the measurement of the psychological magnitude pitch. Acoust. Soc. Am. **8**(3), 185–190 (1937). https://doi.org/10.1121/1.1915893
51. P. P. Vaidyanathan, *Multirate Systems and Filter Banks*. (Prentice Hall, 1993)
52. A. Volmer, M. Romba, C. Schmidt, M. H. Harbi. Optimization of speech intelligibility for fire fighters' full face masks, (2013)
53. T. v. Waterschoot, M. Moonen, Fifty years of acoustic feedback control – state of the art and future challenges. Proc. IEEE. **99**(2), 288–327 (2011)
54. J. J. Withopf, L. L. Jassoume, G. G. Schmidt, A. A. Theiß, in *Proc. DAGA, Darmstadt, Germany*. A Modified Overlap-Add Filter Bank With Reduced Delay, (2012)
55. J. Withopf, S. Rohde, G. Schmidt, Application of frequency shifting in in-car communication systems. Proc. ITG Fachtagung Sprachkommunikation (2014)
56. U. Zölzer, *DAFX – Digital Audio Effects*. (John Wiley & Sons, Ltd, 2011). https://doi.org/10.1002/9781119991298

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.