Matheja *et al. EURASIP Journal on Advances in Signal Processing* 2013, **2013**:191
http://asp.eurasipjournals.com/content/2013/1/191

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

A dynamic multi-channel speech enhancement system for distributed microphones in a car environment

Timo Matheja[1*], Markus Buck[1] and Tim Fingscheidt[2]

Abstract

Supporting multiple active speakers in automotive hands-free or speech dialog applications is an interesting issue not least due to comfort reasons. Therefore, a multi-channel system for enhancement of speech signals captured by distributed distant microphones in a car environment is presented. Each of the potential speakers in the car has a dedicated directional microphone close to his position that captures the corresponding speech signal. The aim of the resulting overall system is twofold: On the one hand, a combination of an arbitrary pre-defined subset of speakers' signals can be performed, e.g., to create an output signal in a hands-free telephone conference call for a far-end communication partner. On the other hand, annoying cross-talk components from interfering sound sources occurring in multiple different mixed output signals are to be eliminated, motivated by the possibility of other hands-free applications being active in parallel. The system includes several signal processing stages. A dedicated signal processing block for interfering speaker cancellation attenuates the cross-talk components of undesired speech. Further signal enhancement comprises the reduction of residual cross-talk and background noise. Subsequently, a dynamic signal combination stage merges the processed single-microphone signals to obtain appropriate mixed signals at the system output that may be passed to applications such as telephony or a speech dialog system. Based on signal power ratios between the particular microphone signals, an appropriate speaker activity detection and therewith a robust control mechanism of the whole system is presented. The proposed system may be dynamically configured and has been evaluated for a car setup with four speakers sitting in the car cabin disturbed in various noise conditions.

Keywords: Distributed microphones; Speaker activity detection; Signal combination; Interfering speaker cancellation; Automotive speech applications

1 Introduction

Applying speech technologies in the car becomes more and more important due to safety and comfort reasons. Relating to automotive environments, many different applications like hands-free telephony, teleconferencing, or speech dialog and recognition are possible. Especially in a car, strong background noises caused by engine, wind, rolling noise, or interfering sound sources may disturb the speech signal and could harm the proper functionality of the mentioned applications. Thus, for the purpose of speech signal enhancement, often, multi-microphone arrangements are used, enabling multi-channel signal processing algorithms. The application of beamforming approaches [1,2] requires a small spacing between the microphones and a predefined geometry in order to get sufficient performance.

In contrast, in this contribution, we want to focus on distributed microphones, where the arrangement is not limited to fixed geometries but where each speaker in the car cabin has a dedicated microphone close to his position. In the case at hand with multiple microphones and multiple speakers to be supported, the sensor signals have to be combined in a beneficial way. In the literature, it is often focussed on setups where multiple microphones are used to capture the speech signal of one single speaker. In this case, multiple spatially distributed microphones may

*Correspondence: timo.matheja@nuance.com
[1] Nuance Communications Deutschland GmbH, Acoustic Speech Enhancement Research, Ulm D-89077, Germany
Full list of author information is available at the end of the article

be mounted in the direct vicinity of just one speaker in order to search for the optimal microphone position.

Hence, the best combination of all microphone signals can be chosen for each bin in the frequency domain, e.g., by applying diversity methods as in [3,4]. It is aimed at obtaining exactly one enhanced and combined output signal out of several input signals. These combined signals can be fed to a hands-free device or a speech recognition system.

In case of noisy speech recognition for in-car situations in [5,6], a fundamentally different approach for the exploitation of spatially distributed distant microphones is introduced. It is proposed to estimate the log speech spectrum at a hypothetical close-talking microphone that should have good quality by multiple regression of the log spectra of several distant microphone signals. In other environments, where the speaker's location is not known before, the microphones are mounted arbitrarily in a living room or in an office to take advantage of the space diversity for distant-talking speech recognition [7] or to process a real-time speaker localization as in [8].

Furthermore, regarding speech enhancement, cross-talk components in the desired signal originating from interfering speakers are a major problem for hands-free as well as for speech recognition systems. Within the scope of this contribution, these components should be suppressed to enhance the combined output signals. Various signal compensation approaches based on Widrow's original work [9] are well known from a range of publications. Due to the risk of signal cancellation, an additional filter helps prevent the cancellation of desired components [10]. This enhanced structure is also picked up by [11] for frequency domain cross-talk cancellation within a call center scenario. It is thought of creating a multiple-input multiple-output system, where each output only includes the speech of the dedicated speaker. Similar techniques can be introduced by blind source separation algorithms [12,13]. These methods are often computationally more expensive. Furthermore, in case of speaker-dedicated microphones, signal compensation approaches can exploit quite good reference signals for compensation of interfering speech components. For further cross-talk suppression, appropriate post-processing schemes exist [14,15].

In this contribution, a generic overall system for speech enhancement of distributed microphone signals is proposed, where each speaker has only one dedicated microphone. An overview is depicted in Figure 1. M microphone signals are transformed to the discrete Fourier transform (DFT) domain and processed, yielding Q mixed output signals for serving a number of applications at the same time. The system allows to configure the resulting number of different output instances designed for different applications during the processing in a generic manner. The core processing part consists of an interfering speaker cancellation (ISC) that compensates the cross-talk components that do not have to be present in the appropriate output instance, a signal enhancement (SE) block performing an extended noise reduction, and a dynamic signal combination (DSC) module. The latter combines a subset of some speaker-related microphone signals to a particular output signal. The whole signal processing is controlled by a control unit based on the comparison of signal powers (see also [11,16]).

In a full-duplex speech communication system, the occurrence of acoustic echoes resulting from the coupling between loudspeakers and microphones has to be avoided. For the proposed system where Q speech applications may be active in parallel, M multi-channel echo cancellation structures are needed each having as many adaptive filters as loudspeaker channels are used by the application. To solve the echo cancellation problem, the related reference signals of the Q different and uncorrelated far-end partners or systems are directly accessible. The topic of stereo- and multi-channel acoustic echo cancellation and the presentation of efficient solutions is not within the scope of this contribution, but further details can be found in [17,18].

The paper is organized as follows: In Section 2, a more detailed overview of the generic system is given. An interfering speaker cancellation is presented in Section 3. The following Section 4 discusses the signal enhancement stage and its submodules. Afterwards, in Section 5, the dynamic signal combination is considered. The robust control of the whole signal processing is introduced in Section 6, and the contribution concludes with an evaluation of the overall system.

2 Generic speech communication system

We propose a highly generic system that allows the handling of several speech applications in the car in parallel. As mentioned, the acoustic echo cancellation problem is not considered in the following. It can be thought, e.g., of two telephone conference calls out of the car in due time, where the two front passengers are communicating with one far-end partner and the backseat passengers with another one within a second application.

The multi-channel system has M microphones and yields a set of Q mixed output signals. Assuming that all the speech sources are uncorrelated, the mth microphone signal $y_m(n)$ can be formulated as the superposition of the clean speech $s_m(n)$, the cross-talk $b_m(n)$, and the background noise component $n_m(n)$ in the time-domain, with n being the sample index:

$$y_m(n) = s_m(n) + b_m(n) + n_m(n). \tag{1}$$

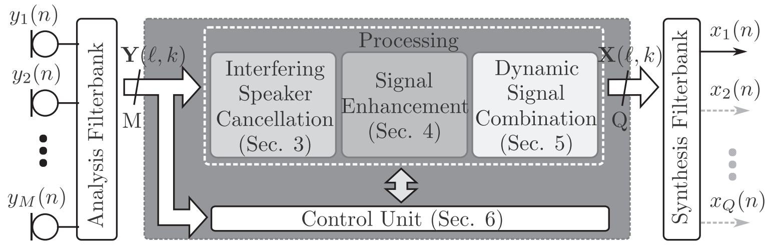

Figure 1 Schematic system overview.

With the time frame index ℓ and the frequency subband index k, the related signal representation in the DFT domain is

$$\mathbf{Y}(\ell,k) = \mathbf{S}(\ell,k) + \mathbf{B}(\ell,k) + \mathbf{N}(\ell,k). \tag{2}$$

The column vector $\mathbf{Y}(\ell,k)$ contains the microphone input signals $Y_m(\ell,k)$ for all microphone channels $m = 1,\dots,M$. According to this formulation, the vector $\mathbf{S}(\ell,k)$ includes the input speech components $S_m(\ell,k)$; the vector $\mathbf{B}(\ell,k)$, the interfering speech components $B_m(\ell,k)$; and the vector $\mathbf{N}(\ell,k)$, the noise components $N_m(\ell,k)$. In general, bold uppercase letters with time frame and/or frequency indices indicate vectors containing M single components for each microphone channel m, or Q components for each output instance q, respectively. The processing is performed at a sampling rate of $f_\mathrm{s} = 16$ kHz. For analysis, a discrete Fourier transform with length of $K = 512$, with a frame-shift of $R = 128$, and a Hann window function is applied. Thus, the subband index k is in the range $k = 0,1,\dots,K-1$. Due to the symmetry properties, only the first $K/2+1$ subbands are effectively processed.

An overview of the four main parts of the whole distributed microphone processing system is depicted in Figure 2. Bold arrows and characters indicate the availability of multiple channels stacked in vectors. Within the ISC block, interfering speakers can be suppressed in a distant target channel by using their dedicated microphone signals as reference for a noise compensation. An adaptive filter structure uses these references to cancel exactly the cross-talk components in the target signals that do not have to be present later in one of Q output signals. The cross-talk components within those target channels that will be combined to the same mixed output signal later are not cancelled in order to exploit some diversity effects afterwards. $\check{\mathbf{Y}}(\ell,k)$ is the resulting signal vector after filtering the interfering cross-talk speech components $\check{\mathbf{Y}}^\mathrm{c}(\ell,k)$ by $\hat{\mathbf{H}}_{m,m'}(\ell,k)$ and subtracting the results from the input signal spectra $\mathbf{Y}(\ell,k)$. The filter $\hat{\mathbf{F}}_{m',m}(\ell,k)$ realizes a blocking structure to avoid signal cancellation effects within the actual ISC.

The adaptation of the filters is controlled by a speaker activity detection (SAD) measure $\widehat{\mathbf{SAD}}(\ell,k)$, determined in the SAD block of the control unit. To obtain similar signal characteristics in all output channels, an automatic gain control is processed within the signal enhancement (SE) stage that adjusts all signal peak levels to a constant target peak level yielding $\tilde{\mathbf{Y}}(\ell,k)$. During speech activity of one speaker, coupling factors $\hat{\mathbf{K}}(\ell,k)$ between the particular signals can now be computed. Thus, residual cross-talk can be estimated, yielding appropriate filter coefficients $\mathbf{G}^\mathrm{RCS}(\ell,k)$ and maximum attenuations $\boldsymbol{\beta}^\mathrm{RCS}(\ell,k)$ for residual cross-talk suppression (RCS) within an extended noise reduction (ENR). This noise reduction block also has to deal with the preparation of the DSC. Due to the different microphone positions and types, the noise signal characteristics (especially noise level and coloration) may differ strongly across the microphone channels. Since annoying switching artifacts may occur in a combined signal, we propose to adjust all noise power spectral densities (PSDs) $\hat{\boldsymbol{\Phi}}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}}}(\ell,k)$ in each channel to a target reference noise level $\hat{\boldsymbol{\Phi}}^\mathrm{ref}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}}}(\ell,k)$ for the transitions at speaker changes by applying a spectral floor $\boldsymbol{\beta}^\mathrm{DSC}(\ell,k)$ within a Wiener noise reduction filter. The determination of the target values is controlled by the fullband speaker activity detection measure $\widehat{\mathbf{SAD}}(\ell)$ that also controls the subsequent signal combination. The noise-reduced signals $\tilde{\mathbf{X}}(\ell,k)$ are merged to obtain Q mixed signals $\mathbf{X}(\ell,k)$, each being a combination of some processed input channel signals. The quantities $\tilde{\mathbf{X}}(\ell,k)$ still include cross-talk components between those channels that are to be combined to one output signal. Hence, spatial diversity can be exploited.

For controlling the ISC, the SE, and the DSC, some matrices are introduced to determine the behavior of the overall system. $\mathbf{W}^\mathrm{ISC}$ is a symmetric $M \times M$ matrix containing zeros and ones, where each row represents a destination channel m, and each column a source channel m'. By setting a one to a position $\langle m,m'\rangle$, the m'th source will be eliminated from the mth channel. If it is desired in a system with $M = 4$ to cancel channels 3 and 4 from channels 1 and 2 and vice versa, the matrix is defined as

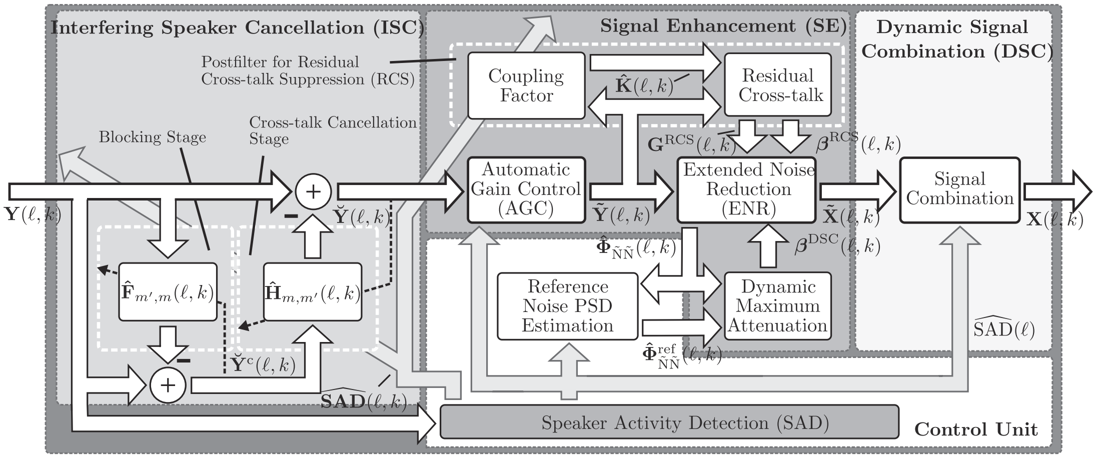
Interfering Speaker Cancellation (ISC)
Signal Enhancement (SE)
Dynamic Signal Combination (DSC)
Postfilter for Residual Cross-talk Suppression (RCS)
Coupling Factor
$\hat{\mathbf{K}}(\ell,k)$
Residual Cross-talk
Blocking Stage
Cross-talk Cancellation Stage
$\mathbf{G}^{\mathrm{RCS}}(\ell,k)$
$\beta^{\mathrm{RCS}}(\ell,k)$
$\mathbf{Y}(\ell,k)$
Automatic Gain Control (AGC)
Extended Noise Reduction (ENR)
Signal Combination
$\check{\mathbf{Y}}(\ell,k)$
$\tilde{\mathbf{Y}}(\ell,k)$
$\tilde{\mathbf{X}}(\ell,k)$
$\mathbf{X}(\ell,k)$
$\hat{\mathbf{\Phi}}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}}}(\ell,k)$
$\beta^{\mathrm{DSC}}(\ell,k)$
$\hat{\mathbf{F}}_{m',m}(\ell,k)$
$\hat{\mathbf{H}}_{m,m'}(\ell,k)$
Reference Noise PSD Estimation
Dynamic Maximum Attenuation
$\widehat{\mathrm{SAD}}(\ell)$
$\check{\mathbf{Y}}^{\mathrm{c}}(\ell,k)$
$\hat{\mathbf{\Phi}}^{\mathrm{ref}}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}}}(\ell,k)$
$\widehat{\mathbf{SAD}}(\ell,k)$
Speaker Activity Detection (SAD)
Control Unit

Figure 2 Overview of the entire distributed microphone processing system. The overall system comprises four main stages: Interfering speaker cancellation (ISC), signal enhancement (SE), dynamic signal combination (DSC), and the control unit.

$$\mathbf{W}^{\mathrm{ISC}} = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}. \tag{3}$$

In a further $Q \times M$ matrix $\mathbf{W}^{\mathrm{DSC}}$, each row represents an output signal q and each column an input channel m. A one at the position $\langle q, m\rangle$ indicates that the channel m has to be present in the qth mixed output signal. Regarding $\mathbf{W}^{\mathrm{ISC}}$ in (3), the related mixing control matrix for $Q = 2$ output signals is

$$\mathbf{W}^{\mathrm{DSC}} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}. \tag{4}$$

In order to implement a generic configuration of the whole system, these control matrices are used for selecting the particular channels to process.

3 Interfering speaker cancellation

In this section, a method to suppress the undesired cross-talk components in each target channel is presented. Interfering speech from a speaker whose dedicated channel signal is to be combined with the considered target channel signal within the qth mixed output signal afterwards is not defined as 'undesired' and is not eliminated in the target channel. This behavior can be configured by the ISC control matrix $\mathbf{W}^{\mathrm{ISC}}$ introduced in (3). Hence, computational costs are saved within the ISC and the possibility of exploiting spatial diversity effects between the microphone channel signals during a later signal combining is kept. The ISC structure is shown in Figure 2 and consists of two parts. The actual cross-talk cancellation stage uses the output of a preceding blocking structure instead of the microphone signals directly for further processing. The blocking stage attenuates the desired signal cancellation effect in order to obtain an improved reference signal within the signal compensation of the undesired components. ISC structures with a blocking stage have been proposed in [10] and are used in [11,19]. Other solutions for the enhancement of the reference signal in noise cancellation structures are, e.g., considered in [20]. However, in this contribution, the blocking structure approach is applied similar to [11] but for the multi-channel case with more than two microphones in a car environment.

3.1 Blocking stage

Within the first stage, adaptive filtering is performed by the blocking structure, where the $M-1$ microphone signals are filtered by $\hat{\mathbf{F}}_{m',m}(\ell,k)$ and subtracted from the signal spectrum in channel m'. This yields the signal component $\check{Y}^{\mathrm{c}}_{m'}(\ell,k)$ to be effectively used as a reference signal for cross-talk cancellation in the mth channel. With the Hermitian operator $(\cdot)^{\mathrm{H}}$, the output results in

$$\check{Y}^{\mathrm{c}}_{m'}(\ell,k) = Y_{m'}(\ell,k) - \sum_{\substack{m\in\{1,\ldots,M\}\\ m\neq m'}} \left(\hat{\mathbf{F}}_{m',m}(\ell,k)\right)^{\mathrm{H}} \mathbf{Y}_m(\ell,k), \tag{5}$$

with the related column vectors for filtering

$$\hat{\mathbf{F}}_{m',m}(\ell,k) = \left[\hat{F}_{m',m,0}(\ell,k), \ldots, \hat{F}_{m',m,L_{\mathrm{FIR}}-1}(\ell,k)\right]^{\mathrm{T}},$$
$$\mathbf{Y}_m(\ell,k) = \left[Y_m(\ell,k), \ldots, Y_m(\ell - L_{\mathrm{FIR}}+1,k)\right]^{\mathrm{T}}. \tag{6}$$

Here, L_{FIR} indicates the length of the adaptive filters, and $(\cdot)^{\mathrm{T}}$ denotes the transpose of the vectors. To exclude the desired speech components from the ISC reference and therewith to avoid signal cancellation within the ISC, the filter coefficients $\hat{\mathbf{F}}_{m',m}(\ell,k)$ are only updated if solely the particular mth ISC target channel shows speaker activity ($\widehat{\mathrm{SAD}}_m(\ell,k)$=1, as introduced in Section 6.1 and determined by (82)). Thus, the resulting speech component and therewith the effective cross-talk in channel m' calculated in (5) will equal to 0 during these situations. The filter coefficients are adapted by the NLMS algorithm (e.g., [21]):

$$\hat{\mathbf{F}}_{m',m}(\ell+1,k)=\hat{\mathbf{F}}_{m',m}(\ell,k)+\alpha_m^{\mathrm{bs}}(\ell,k)\frac{\check{Y}_{m'}^{\mathrm{c}\,*}(\ell,k)\mathbf{Y}_m(\ell,k)}{\parallel \mathbf{Y}_m(\ell,k)\parallel^2}, \tag{7}$$

where $(\cdot)^*$ is the conjugate complex operator. The related step size can be expressed as

$$\alpha_m^{\mathrm{bs}}(\ell,k)=\begin{cases}\alpha, & \text{if } \widehat{\mathrm{SAD}}_m(\ell,k)=1,\\ 0, & \text{else.}\end{cases} \tag{8}$$

Alternatively, for a two-channel scenario, an additional control mechanism based on an optimal step size has been proposed by the authors in [22]. The preferred values for the implementation are $L_{\mathrm{FIR}}=3$ and $\alpha=0.3$.

3.2 Cross-talk cancellation stage

As depicted in Figure 2, secondly, the cross-talk cancellation stage follows within the ISC. With the filter vector $\hat{\mathbf{H}}_{m,m'}(\ell,k)$ and the cross-talk component vector $\check{\mathbf{Y}}_{m'}^{\mathrm{c}}(\ell,k)$ introduced by

$$\hat{\mathbf{H}}_{m,m'}(\ell,k)=\left[\hat{H}_{m,m',0}(\ell,k),\ldots,\hat{H}_{m,m',L_{\mathrm{FIR}}-1}(\ell,k)\right]^{\mathrm{T}},$$
$$\check{\mathbf{Y}}_{m'}^{\mathrm{c}}(\ell,k)=\left[\check{Y}_{m'}^{\mathrm{c}}(\ell,k),\ldots,\check{Y}_{m'}^{\mathrm{c}}(\ell-L_{\mathrm{FIR}}+1,k)\right]^{\mathrm{T}}, \tag{9}$$

the cross-talk cancelled signal is obtained:

$$\check{Y}_m(\ell,k)=Y_m(\ell,k)-\sum_{m'=1}^{M}W_{m,m'}^{\mathrm{ISC}}\cdot\hat{\mathbf{H}}_{m,m'}^{\mathrm{H}}(\ell,k)\check{\mathbf{Y}}_{m'}^{\mathrm{c}}(\ell,k). \tag{10}$$

Here, the combination of the last two factors constitutes the filtered cross-talk components originating from all channels m' and used for cancellation of interfering speakers' signals by subtraction from the target signals $Y_m(\ell,k)$ (see structure in Figure 2). The single elements $W_{m,m'}^{\mathrm{ISC}}$ of the ISC control matrix $\mathbf{W}^{\mathrm{ISC}}$ ensure that only those cross-talk components are eliminated that are desired to be cancelled. Due to forced zeros on the main diagonal of $\mathbf{W}^{\mathrm{ISC}}$, the contribution of the desired signal itself is excluded. For the filter update with the NLMS algorithm, we have

$$\hat{\mathbf{H}}_{m,m'}(\ell+1,k)=\hat{\mathbf{H}}_{m,m'}(\ell,k)+\alpha_{m'}(\ell,k)\frac{\check{Y}_m^*(\ell,k)\check{\mathbf{Y}}_{m'}^{\mathrm{c}}(\ell,k)}{\parallel\check{\mathbf{Y}}_{m'}^{\mathrm{c}}(\ell,k)\parallel^2}. \tag{11}$$

The step size

$$\alpha_{m'}(\ell,k)=\begin{cases}\alpha, & \text{if } \widehat{\mathrm{SAD}}_{m'}(\ell,k)=1,\\ 0, & \text{else,}\end{cases} \tag{12}$$

controls the ISC adaptation, showing that the cross-talk cancellation filters $\hat{\mathbf{H}}_{m,m'}(\ell,k)$ are only to be updated if interfering speech activity is indicated for the m'th channel. The signals are continuously filtered, and the cross-talk components are attenuated without causing much distortion of the desired speech.

4 Signal enhancement (SE)

To further enhance the speech signals, an automatic gain control (AGC) and an ENR follow within the SE block. In addition to a stationary noise reduction, still existing residual cross-talk components are suppressed, and the AGC and the ENR care for the adjustment of the signal characteristics prior to the subsequent signal combination. The determination of all these parts is discussed in the following. Suitable parameters for the implementation of the SE part are depicted in Table 1.

4.1 Automatic gain control

Due to varying distances between the speakers and the microphones, the related microphone speech signal levels differ among the channels. To care for a compensation of these differences, an AGC is performed. Based on the input signal $\check{Y}_m(\ell,k)$, the related peak level $\check{Y}_m^{\mathrm{P}}(\ell,k)$ is estimated, and a fullband amplification factor $a_m(\ell)$ is determined to adapt the current peak level to a target peak level $\check{Y}^{\mathrm{ref}}$ that can be defined beforehand. A method for peak level estimation is proposed in [23] based on a simple speech activity detector. But, here, the speaker activity detector presented in Appendix 2 is used, and instead of processing a time domain signal for peak tracking, a root-mean-square measure over all subbands is applied. The actual peak level is estimated whenever single-talk

Table 1 Preferred parameter values for the implementation of the signal enhancement part

Parameter	Value
γ_{a}	0.8
$\check{Y}^{\mathrm{ref}}$	0.5
γ_{WF1}	20
γ_{WF2}	2
β	0.25
β^{min}	0.3
β^{max}	3.6
$\gamma_{\mathrm{inc}}^{K}$	1.05
$\gamma_{\mathrm{dec}}^{K}$	0.95

is detected for the related channel. Single-channel speech activity $\widehat{\text{STD}}_m(\ell) \in \{0, 1\}$ is indicated by

$$\widehat{\text{STD}}_m(\ell) = \begin{cases} 1, \text{ if } \widehat{\text{SAD}}_m(\ell) = 1 \wedge \widehat{\text{DTD}}(\ell) = 0, \\ 0, \text{ else.} \end{cases} \tag{13}$$

For an introduction to the fullband speaker activity detector $\widehat{\text{SAD}}_m(\ell) \in \{0, 1\}$ and the double-talk detector $\widehat{\text{DTD}}(\ell) \in \{0, 1\}$, please refer to Section 6.1 and Appendix 1. The equalized output for each channel results in

$$\tilde{Y}_m(\ell, k) = a_m(\ell)\check{Y}_m(\ell, k), \tag{14}$$

with the recursively averaged frequency-independent gain factors [24]

$$a_m(\ell) = \gamma_{\text{a}} \cdot a_m(\ell - 1) + (1 - \gamma_{\text{a}}) \cdot \frac{\check{Y}^{\text{ref}}}{\check{Y}^{\text{P}}_m(\ell)}. \tag{15}$$

4.2 Extended noise reduction

With the objective of obtaining an overall extended noise reduction including a postfilter for residual cross-talk suppression (RCS) and a dynamic maximum attenuation to realize a dynamic combination of the microphone signals later, two approaches are combined to one noise reduction characteristic. For the filtering of the noisy signal $\tilde{Y}_m(\ell, k)$ follows

$$\tilde{X}_m(\ell, k) = G^{\text{ENR}}_m(\ell, k) \cdot \tilde{Y}_m(\ell, k). \tag{16}$$

The filter coefficients $G^{\text{ENR}}_m(\ell, k)$ are determined by restriction of the cross-talk suppression filter coefficients $G^{\text{RCS}}_m(\ell, k)$ to a time- and frequency-dependent maximum attenuation $\beta^{\text{ENR}}_m(\ell, k)$ to keep a certain level of residual background noise and mask artifacts like musical tones:

$$G^{\text{ENR}}_m(\ell, k) = \max\left\{G^{\text{RCS}}_m(\ell, k),\ \beta^{\text{ENR}}_m(\ell, k)\right\}. \tag{17}$$

The maximum attenuation includes two factors:

$$\beta^{\text{ENR}}_m(\ell, k) = \beta^{\text{RCS}}_m(\ell, k) \cdot \beta^{\text{DSC}}_m(\ell, k), \tag{18}$$

where the first factor is the spectral floor conditioned by the cross-talk suppression postfilter in Section 4.2.1, and the second one is the additional maximum attenuation for DSC determined in Section 4.2.4.

4.2.1 Postfilter for residual cross-talk suppression

For suppression of the still existing residual cross-talk components $\tilde{B}_m(\ell, k)$ present in the cross-talk compensated and equalized signal $\tilde{Y}_m(\ell, k)$, a postprocessing can be applied that complements the reduction of stationary background noise similar to the approach in [14]. Generally, different spectral weighting filter characteristics can be chosen for noise reduction. Instead of applying the basic Wiener filter [23] in this contribution, the application of a recursive Wiener filtering [25] is proposed to reduce musical tones in the noise-reduced output signal $\tilde{X}_m(\ell, k)$. With the maximum noise overestimation factor γ_{WF1} and the fixed overestimation γ_{WF2}, the filter coefficients for the residual cross-talk suppression postfilter characteristic result in

$$G^{\text{RCS}}_m(\ell, k) = 1 - \min\left\{\gamma_{\text{WF1}}, \frac{\gamma_{\text{WF2}}}{G^{\text{ENR}}_m(\ell - 1, k)}\right\} \cdot \frac{\hat{\Phi}'_{\tilde{\text{N}}\tilde{\text{N}},m}(\ell, k)}{\hat{\Phi}_{\tilde{\text{Y}}\tilde{\text{Y}},m}(\ell, k)}, \tag{19}$$

where $G^{\text{ENR}}_m(\ell - 1, k)$ is the limited quantity $G^{\text{RCS}}_m(\ell, k)$ of the previous frame (see (17)). Furthermore, $\hat{\Phi}'_{\tilde{\text{N}}\tilde{\text{N}},m}(\ell, k)$ is a modified noise PSD that is a combination of an AGC weighted stationary noise part and the residual cross-talk component:

$$\hat{\Phi}'_{\tilde{\text{N}}\tilde{\text{N}},m}(\ell, k) = \hat{\Phi}_{\tilde{\text{N}}\tilde{\text{N}},m}(\ell, k) + \hat{\Phi}_{\tilde{\text{B}}\tilde{\text{B}},m}(\ell, k), \tag{20}$$

where the stationary term is determined by weighting a continuously estimated noise PSD $\hat{\Phi}_{\check{\text{N}}\check{\text{N}},m}(\ell, k)$ by the squared AGC gain factors (15) as

$$\hat{\Phi}_{\tilde{\text{N}}\tilde{\text{N}},m}(\ell, k) = a^2_m(\ell) \cdot \hat{\Phi}_{\check{\text{N}}\check{\text{N}},m}(\ell, k). \tag{21}$$

To obtain $\hat{\Phi}_{\check{\text{N}}\check{\text{N}},m}(\ell, k)$, e.g., the improved minimum recursive averaging approach [26] can be chosen. Regarding (19), it has to be ensured that the cross-talk components are effectively suppressed. Thus, the residual cross-talk suppression component $\beta^{\text{RCS}}_m(\ell, k)$ of the overall spectral floor in (18) has to be adjusted. In addition to a constant spectral floor β, here, a dynamic time- and frequency-dependent component realizes the attenuation of the residual cross-talk down to the same level as the stationary background noise. Including β, we have

$$\beta^{\text{RCS}}_m(\ell, k) = \beta \cdot \sqrt{\frac{\hat{\Phi}_{\tilde{\text{N}}\tilde{\text{N}},m}(\ell, k)}{\hat{\Phi}_{\tilde{\text{N}}\tilde{\text{N}},m}(\ell, k) + \hat{\Phi}_{\tilde{\text{B}}\tilde{\text{B}},m}(\ell, k)}}. \tag{22}$$

4.2.2 Residual cross-talk

For realization of the residual cross-talk suppression, an estimate for the residual cross-talk component $\hat{\Phi}_{\tilde{\text{B}}\tilde{\text{B}},m}(\ell, k)$ used in (20) and (22) has to be determined. Due to the signal model described in (2), it follows for the processed signal PSD estimates after the ISC and AGC:

$$\hat{\Phi}_{\tilde{\text{Y}}\tilde{\text{Y}},m}(\ell, k) = \hat{\Phi}_{\tilde{\text{S}}\tilde{\text{S}},m}(\ell, k) + \hat{\Phi}_{\tilde{\text{B}}\tilde{\text{B}},m}(\ell, k) + \hat{\Phi}_{\tilde{\text{N}}\tilde{\text{N}},m}(\ell, k), \tag{23}$$

with $\hat{\Phi}_{\tilde{S}\tilde{S},m}(\ell,k)$ including all desired speech components - direct and cross-talk components - that are not to be cancelled. The overall residual cross-talk in channel m can be expressed as the sum of all relevant components resulting from each channel m' to the desired one m:

$$\hat{\Phi}_{\tilde{B}\tilde{B},m}(\ell,k) = \sum_{m'=1}^{M} W^{\mathrm{ISC}}_{m,m'} \cdot \hat{\Phi}_{\tilde{B}\tilde{B},m,m'}(\ell,k). \tag{24}$$

Due to forced zeros on the main diagonal of $\mathbf{W}^{\mathrm{ISC}}$, the contribution of the desired signal itself is always excluded. The residual cross-talk quantity $\hat{\Phi}_{\tilde{B}\tilde{B},m,m'}(\ell,k)$ in channel m resulting from the m'th channel cannot be observed. It may be estimated by weighting a remote speaker's signal PSD in channel m' by an estimated instantaneous acoustic coupling factor $\tilde{K}_{m,m'}(\ell,k)$ between each channel m' and the channel m:

$$\hat{\Phi}_{\tilde{B}\tilde{B},m,m'}(\ell,k) = \tilde{K}_{m,m'}(\ell,k) \cdot \hat{\Phi}_{\tilde{S}\tilde{S},m'}(\ell,k). \tag{25}$$

Alternatively, the residual cross-talk PSD can be written only during single-talk activity in the m'th channel ($\hat{\Phi}_{\tilde{S}\tilde{S},m}(\ell,k) = 0$ and $\hat{\Phi}_{\tilde{B}\tilde{B},m,m'}(\ell,k) = \hat{\Phi}_{\tilde{B}\tilde{B},m}(\ell,k)$) by directly observable quantities. Rearranging (23) and simplifying and including (24) results in

$$\hat{\Phi}_{\tilde{B}\tilde{B},m,m'}(\ell,k) = \hat{\Phi}_{\tilde{Y}\tilde{Y},m}(\ell,k) - \hat{\Phi}_{\tilde{N}\tilde{N},m}(\ell,k). \tag{26}$$

For the single-talk speech component PSD in channel m' follows accordingly:

$$\hat{\Phi}_{\tilde{S}\tilde{S},m'}(\ell,k) = \hat{\Phi}_{\tilde{Y}\tilde{Y},m'}(\ell,k) - \hat{\Phi}_{\tilde{N}\tilde{N},m'}(\ell,k). \tag{27}$$

However, with this expression and a long-term estimate $\hat{K}_{m,m'}(\ell,k)$ for the coupling factor, the residual cross-talk PSD can be estimated by the weighted sum of all considered remote speech components in channel m'. After including (25) in (24), we have

$$\hat{\Phi}_{\tilde{B}\tilde{B},m}(\ell,k) = \sum_{m'=1}^{M} W^{\mathrm{ISC}}_{m,m'} \cdot \hat{K}_{m,m'}(\ell,k) \cdot \hat{\Phi}_{\tilde{S}\tilde{S},m'}(\ell,k). \tag{28}$$

Note that if no single speech activity occurs in channel m', then $\hat{\Phi}_{\tilde{S}\tilde{S},m'}(\ell,k) = 0$. Within the computation of the overall considered cross-talk quantity in channel m again, the coefficients of the ISC control matrix $\mathbf{W}^{\mathrm{ISC}}$ force to neglect eliminating cross-talk components originating from a channel that has to be merged with the currently considered one afterwards.

4.2.3 Coupling factor

The principle of an acoustic coupling factor is already introduced for the acoustic echo cancellation problem by [23]. Firstly, using (26) and (27), the instantaneous coupling factor within (25) can be expressed during single-talk as

$$\tilde{K}_{m,m'}(\ell,k) = \frac{\hat{\Phi}_{\tilde{Y}\tilde{Y},m}(\ell,k) - \hat{\Phi}_{\tilde{N}\tilde{N},m}(\ell,k)}{\hat{\Phi}_{\tilde{Y}\tilde{Y},m'}(\ell,k) - \hat{\Phi}_{\tilde{N}\tilde{N},m'}(\ell,k)}. \tag{29}$$

The long-term estimate of the coupling factor applied in (28) is updated during periods of single-talk whenever frequency-selective speech activity is detected in the m'th channel:

$$\hat{K}_{m,m'}(\ell,k) = \begin{cases} \hat{K}_{m,m'}(\ell-1,k), & \text{if } \widehat{\mathrm{SAD}}_{m'}(\ell,k)=0 \vee \widehat{\mathrm{DTD}}(\ell)=1, \\ \gamma^{\mathrm{K}}_{m,m'}(\ell,k)\cdot\hat{K}_{m,m'}(\ell-1,k), & \text{else}, \end{cases} \tag{30}$$

with the time- and frequency-dependent constant $\gamma^{\mathrm{K}}_{m,m'}(\ell,k)$ determined by comparing the instantaneous with the long-term estimated coupling factor:

$$\gamma^{\mathrm{K}}_{m,m'}(\ell,k) = \begin{cases} \gamma^{\mathrm{K}}_{\mathrm{inc}}, & \text{if } \tilde{K}_{m,m'}(\ell,k) > \hat{K}_{m,m'}(\ell-1,k), \\ \gamma^{\mathrm{K}}_{\mathrm{dec}}, & \text{if } \tilde{K}_{m,m'}(\ell,k) < \hat{K}_{m,m'}(\ell-1,k), \\ 1, & \text{else}. \end{cases} \tag{31}$$

For increasing and decreasing, the appropriate constants $\gamma^{\mathrm{K}}_{\mathrm{inc}}$ and $\gamma^{\mathrm{K}}_{\mathrm{dec}}$ are chosen. The fullband speaker activity detection $\widehat{\mathrm{SAD}}_m(\ell)$, the frequency-selective one $\widehat{\mathrm{SAD}}_{m'}(\ell,k)$, and the double-talk detector $\widehat{\mathrm{DTD}}(\ell)$ are explained in Section 6.1, Appendix 1, Appendix 2, and Appendix 3.

4.2.4 Dynamic maximum attenuation

The noise signal characteristics may differ strongly across the microphone channels, depending on the position or type of the microphone and the kind of background noise. However, as a preprocessing step for the realization of a dynamic combination of the microphone signals (Section 5), equal power and spectral shape of the background noise have to be provided for all related channels during transitions between different active speakers if a switching between them is performed. Thus, annoying switching artifacts are to be avoided by a dynamic maximum attenuation that can be applied within the noise reduction regarding (17) and (18). The dynamic spectral floor factor [24]

$$\beta^{\mathrm{DSC}}_{m}(\ell,k) = \sqrt{\frac{\hat{\Phi}^{\mathrm{ref}}_{\tilde{N}\tilde{N},m}(\ell,k)}{\hat{\Phi}_{\tilde{N}\tilde{N},m}(\ell,k)}} \tag{32}$$

used in (18) adjusts the estimated noise PSD for each microphone channel signal to a reference PSD $\hat{\Phi}^{\mathrm{ref}}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}},m}(\ell,k)$ in such a way that no discontinuities within the signal characteristics are noticeable across the microphones. The important reference PSD is determined by (50) later in Section 6.2, and $\hat{\Phi}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}},m}(\ell,k)$ is obtained by (21). Regarding the maximum attenuation, it might be advantageous to introduce a limit $\beta^{\mathrm{DSC}}_m(\ell,k) \in [\beta^{\min},\beta^{\max}]$ with $\beta^{\min} \le \beta \le \beta^{\max}$ [24] for an adequate performance of the DSC.

5 Dynamic signal combination

Finally, the signals of the single-microphone channels have to be combined to mixed output signals. Applying the AGC in (14) and the extended noise reduction in (16) with the dynamic maximum attenuation in (32), this can be performed without noticeable switching artifacts within the signal characteristics. In [24], the authors presented a solution for this challenge but without considering diversity and with only one desired output signal $Q = 1$. Within the presented generic system here, diversity effects are exploited, similar to [27]. Frequency-selective switching shall be applied, and it shall be possible to serve several speech applications in parallel. Hence, selected microphone channels are to be combined to Q separate mixed output signals. The microphone channels to be combined to one output signal instance can be selected by the DSC control matrix $\mathbf{W}^{\mathrm{DSC}}$ introduced in (4). As depicted in Figure 2, after the signal combination, the vector $\mathbf{X}(\ell,k)$ includes Q output signals, where each is a combination of some appropriately processed microphone signals $\tilde{\mathbf{X}}(\ell,k)$. If speech activity is detected only in channels that are combined to the qth output signal later and no speech activity is detected in other channels at this time instance, we call it output-related single-talk. Then, the mixed signal can be calculated by a combination of all M available signals $\tilde{X}_m(\ell,k)$ to exploit the spatial diversity by considering the cross-talk components occurring in each of all channels. For the appropriate output-related single-talk detection measure $\widehat{\mathrm{STM}}_q(\ell)$, we have

$$\widehat{\mathrm{STM}}_q(\ell) = \begin{cases} 0, & \text{if } \widehat{\mathrm{DTD}}_{\overline{q}}(\ell) = 1, \\ \sum\limits_{m=1}^{M} W^{\mathrm{DSC}}_{q,m} \cdot \widehat{\mathrm{SAD}}_m(\ell), & \text{else,} \end{cases} \tag{33}$$

where $\overline{q} \in \{1,\ldots,Q\}$ with $\overline{q} \neq q$. Therewith, $\widehat{\mathrm{DTD}}_{\overline{q}}(\ell) \in \{0,1\}$ is a double-talk detector related to the specific $\overline{q}$th output signal. It takes effect if speech is detected not only for the currently observed qth output signal but also in microphone channels related to other output signals $\overline{q}$. For details concerning the robust fullband SAD detector $\widehat{\mathrm{SAD}}_m(\ell)$, please refer to Section 6.1 and Appendix 2.

During the described output-related single-talk, the magnitude and phase are treated differently and independently within the signal combination process. The spectral magnitude of the channel signal showing the best signal-to-noise ratio (SNR) is selected by the real-valued weights $w_{q,m}(\ell,k) \in \{0,1\}$, and the phase $\Phi^{\mathrm{mix}}_q(\ell,k)$ of the last active channel within the qth mixed output signal is appended (e.g., similar to [3]). Hence, for the qth output signal, we obtain

$$X_q(\ell,k) = \begin{cases} \sum\limits_{m=1}^{M} w_{q,m}(\ell,k) \cdot \left|\tilde{X}_m(\ell,k)\right| \cdot e^{j\phi^{\mathrm{mix}}_q(\ell,k)}, & \text{if } \widehat{\mathrm{STM}}_q(\ell) > 0, \\ \sum\limits_{m=1}^{M} W^{\mathrm{DSC}}_{q,m} \cdot w_{q,m}(\ell) \cdot \tilde{X}_m(\ell,k), & \text{else.} \end{cases} \tag{34}$$

The last line applies if speech activity is detected in other than the qth output related signals, or an overall noise period occurs. No frequency-selective channel switching is adopted but rather a fullband decision controlled by the weights $w_{q,m}(\ell) \in \{0,1\}$. With the Kronecker delta $\delta_{m,u(\ell,k)}$ selecting the channel with the maximum SNR, the temporary frequency-selective weights result in

$$w'_{q,m}(\ell,k) = \delta_{m,u(\ell,k)} \;, \tag{35}$$

where $u(\ell,k) \in \{0,\ldots,M\}$ denotes the channel showing the maximum SNR:

$$u(\ell,k) = \underset{m\in\{1,\ldots,M\}}{\operatorname{argmax}} \left\{ \hat{\xi}_m(\ell,k) \right\}. \tag{36}$$

In Section 6.1, the estimation of the SNR $\hat{\xi}_m(\ell,k)$ is given by (43). The final resulting frequency-selective weight is determined by

$$w_{q,m}(\ell,k) = \begin{cases} w'_{q,m}(\ell,k), & \text{if } \left|\tilde{X}_{u(\ell,k)}(\ell,k)\right| > \left|\tilde{X}_m(\ell,k)\right|, \\ w_{q,m}(\ell), & \text{else.} \end{cases} \tag{37}$$

This implies that the maximum SNR channel is only selected if the absolute value of the noise-reduced signal in this channel is larger than the absolute value of the signal within the currently observed mth channel. Otherwise, the fullband weight $w_{q,m}(\ell)$ is used. Therefore, it is searched for fullband activity of the mth speaker corresponding to the qth output signal. If no single one or more than one speakers are active per output instance, the previous decision is kept:

$$w_{q,m}(\ell) = \begin{cases} \widehat{\mathrm{SAD}}_m(\ell), & \text{if } \widehat{\mathrm{STM}}_q(\ell) = 1, \\ w_{q,m}(\ell-1), & \text{else.} \end{cases} \tag{38}$$

Regarding the phase in (34), always the phase value of the last active channel in the present output instance indicated by $\nu_q(\ell)$ is used:

$$\phi_q^{\mathrm{mix}}(\ell,k) = \phi_{\nu_q(\ell)}(\ell,k). \tag{39}$$

With (38), it follows for the last active channel index:

$$\nu_q(\ell) = \underset{m\in\{1,\ldots,M\}}{\operatorname{argmax}} \left\{ w_{q,m}(\ell) \right\}. \tag{40}$$

6 Control Unit

The energy-based control mechanism for the proposed speech communication system with distributed speaker-dedicated microphones is introduced below as well as the reference noise PSD estimation that is important for the signal combination process.

6.1 Robust speaker activity detection

A robust differentiation between several speakers has to be achieved. For this SAD, an energy-based approach relying on the evaluation of signal power ratios between the microphone signals is applied. A similar overall SAD system was already introduced by the authors in [28]. An overview of the whole SAD block is given in Figure 3. The enhanced fullband detector $\widehat{\mathbf{SAD}}(\ell)$ as an improvement of a basic fullband detector $\widetilde{\mathbf{SAD}}(\ell)$ as well as a frequency-selective detector $\widehat{\mathbf{SAD}}(\ell,k)$ is obtained. As depicted in Figure 2, the fullband SAD measure is used for general control, whereas the frequency-selective value is of interest for the ISC and especially for controlling the adaptive filters. Besides relying on the signal power ratio (SPR), these detectors are based on the SNR as a further energy-based measure.

As a first step, the SPR has to be defined. Regarding the signal model in (2), we obtain for the signal PSD estimate $\hat{\Phi}_{\Sigma\Sigma,m}(\ell,k)$ including the direct speech component as well as the cross-talk components

$$\hat{\Phi}_{\Sigma\Sigma,m}(\ell,k) = \max\left\{ \hat{\Phi}_{\mathrm{YY},m}(\ell,k) - \hat{\Phi}_{\mathrm{NN},m}(\ell,k), 0 \right\}. \tag{41}$$

The estimate $\hat{\Phi}_{\mathrm{YY},m}(\ell,k)$ is determined by smoothing the squared magnitudes of the microphone signal spectra $Y_m(\ell,k)$. The noise PSD $\hat{\Phi}_{\mathrm{NN},m}(\ell,k)$ can be estimated, e.g., by the improved minimum controlled recursive averaging approach [26]. In a system with $M \geq 2$ microphones, the SPR is expressed similar to [29] for each channel m as

$$\widehat{\mathrm{SPR}}_m(\ell,k) = \frac{\max\left\{ \hat{\Phi}_{\Sigma\Sigma,m}(\ell,k), \epsilon \right\}}{\max\left\{ \underset{\substack{m'\in\{1,\ldots,M\}\\ m'\neq m}}{\max} \left\{ \hat{\Phi}_{\Sigma\Sigma,m'}(\ell,k) \right\}, \epsilon \right\}}, \tag{42}$$

with the very small value ϵ ensuring the validity of the expression. Due to the fact that each speaker has a dedicated microphone and due to the assumption that always one microphone captures the speech best, the active speaker can be identified by the evaluation of the SPR among the available microphones. Basically, speech activity of speaker m is detected if the related logarithmic SPR is larger than 0 dB. For computational details of such a basic fullband detector $\widetilde{\mathbf{SAD}}(\ell)$, please refer to Appendix 1. In order to consider the SPR only in significant regions during the determination of the SAD, the channel-related SNR $\hat{\xi}_m(\ell,k)$ is included. It is estimated similar to [30] by

$$\hat{\xi}_m(\ell,k) = \frac{\max\left\{ \min\left\{ \hat{\Phi}_{\mathrm{YY},m}(\ell,k), |Y_m(\ell,k)|^2 \right\} - \hat{\Phi}'_{\mathrm{NN},m}(\ell,k), 0 \right\}}{\hat{\Phi}'_{\mathrm{NN},m}(\ell,k)}, \tag{43}$$

with the PSD estimate $\hat{\Phi}_{\mathrm{YY},m}(\ell,k)$ and the related modified noise estimate $\hat{\Phi}'_{\mathrm{NN},m}(\ell,k)$ for the determination of a reliable SNR value. Using the preferred factor $\gamma_{\mathrm{SNR}} = 4$, it follows

$$\hat{\Phi}'_{\mathrm{NN},m}(\ell,k) = \gamma_{\mathrm{SNR}} \cdot \hat{\Phi}_{\mathrm{NN},m}(\ell,k). \tag{44}$$

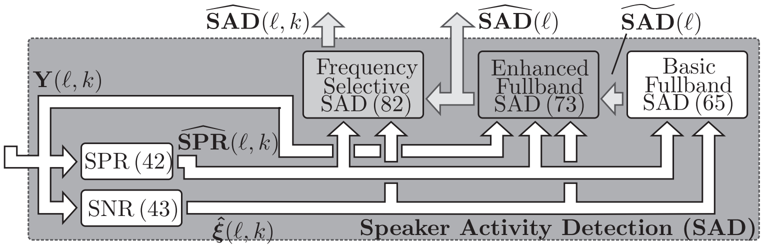
$\widehat{\mathbf{SAD}}(\ell,k)$
$\widehat{\mathbf{SAD}}(\ell)$
$\widetilde{\mathbf{SAD}}(\ell)$
$\mathbf{Y}(\ell,k)$
Frequency Selective SAD (82)
Enhanced Fullband SAD (73)
Basic Fullband SAD (65)
SPR (42)
$\widehat{\mathbf{SPR}}(\ell,k)$
SNR (43)
$\hat{\boldsymbol{\xi}}(\ell,k)$
Speaker Activity Detection (SAD)

Figure 3 Speaker activity detection stage. Overview of the speaker activity detection block in Figure 2 within the control unit.

By simply evaluating the power ratios, the presented basic fullband detection of the active speaker can be performed. But, transient interferers like indicator noise, outside crossing cars, and speech from interfering speakers may be wrongly assigned to one speaker's activity, e.g., during interfering backseat passengers in a system with only two microphones in the front. The robustness for these and for other situations in general can be increased by applying an enhanced fullband detector $\widehat{\mathbf{SAD}}(\ell)$ based on the exploitation of SPR patterns as was first introduced by the authors in [29]. Therewith, the characteristics of the room acoustics shall be involved and evaluated. Due to the distinguishing room acoustics, a sharp decline of the energy may occur in some special subbands of the speaker's dedicated mth microphone signal. This causes a lower amount of energy in the speaker's closest microphone compared to the distant ones. Hence, for the active mth speaker, the related observable signal power ratio $\widehat{\mathrm{SPR}}_m(\ell,k)$ is smaller than one or at least very low only in some special subbands. These subbands may be called *multipath-induced fading* subbands related to the multipath propagation effects. The number and location of these subbands are assumed to be characteristic for each sound source at a different location in the car. Thus, appropriate patterns representing this effect may indicate the position of a speaker if they match a reference pattern set. For further details, please refer to Appendix 2.

After the determination of the robust fullband SAD, a frequency-selective detection $\widehat{\mathbf{SAD}}(\ell,k)$ of the active speaker has to be carried out. Due to the occurring multipath-induced fading subbands, it is not reliable to distinguish between the active speakers, depending on whether a positive or negative logarithmic SPR occurs in a frequency subband as presented in [11]. In case of speech activity of one speaker, the related SPR may show negative values for a small number of the multipath-induced fading subbands due to the room acoustics. The detection of speech activity might be missed in these subbands for the corresponding speaker. Thus, we want to avoid the decision based on a hard thresholding and propose an approach that exploits a modeling of the power ratios as was similarly proposed by the authors in [31]. The details of the specific version used in this contribution is presented in Appendix 3. Finally, it should be noted that due to the sparseness of speech activity, double-talk does not have to be detected in a frequency-selective manner but rather on a frame basis as a fullband measure.

6.2 Reference noise power spectral density estimation

For signal combination, the spectra of the residual background noise after noise reduction are aligned among the Q output signals. This allows for selecting channels within the dynamic signal combination unit without getting switching artifacts. The spectral alignment to a reference noise spectrum is done by dynamic modification of a frequency-dependent spectral floor parameter within the noise reduction (16). The computation of this dynamic spectral floor is proposed in (32), where we need to know an appropriate reference noise PSD. In order to determine a reference background noise out of all those different microphone signals that are to be mixed to one output signal, it has to be decided which speaker is the dominant one at a time instance. Corresponding dominance weights can be determined by evaluating the duration for which a speaker has been detected. While a speaker is active alone, his dominance increases until it reaches a maximum value and therewith full dominance. Then, the target noise level has to be controlled by this channel alone. If a different relevant speaker within the subset of microphone signals to be combined to one output instance becomes active, the dominances of all the other related channels decrease. In order to determine dominance weights, firstly, we define the channel-dependent dominance counters [24]

$$c_m(\ell) = \max\left\{\min\left\{c_m(\ell-1) + \Delta c_m(\ell), c_{\max}\right\}, c_{\min}\right\}, \tag{45}$$

where the limitation of the counters to a minimum $c_{\min}$ and a maximum value $c_{\max}$, respectively, defines the range between the minimum and full dominance of a speaker. The parameter $\Delta c_m(\ell)$ controls the increase or decrease of the counters and is dependent on the single-talk speaker activity detection $\widehat{\mathrm{STD}}_m(\ell) \in \{0,1\}$ introduced in (13). With the increasing and decreasing step sizes c_{inc} and c_{dec}, respectively, it follows

$$\Delta c_m(\ell) = \begin{cases} c_{\mathrm{inc}}, & \text{if } \widehat{\mathrm{STD}}_m(\ell) = 1, \\ -c_{\mathrm{dec},m}, & \text{else.} \end{cases} \tag{46}$$

After speaking for a period t_{inc}, a speaker m should get full dominance. This determines the step size for increasing [24]

$$c_{\mathrm{inc}} = \frac{c_{\max} - c_{\min}}{t_{\mathrm{inc}}} \cdot T_{\mathrm{frame}}\,, \tag{47}$$

with the period T_{frame} between two consecutive time frames. The dominance counter of the previous active speaker has to reach $c_{\min}$ after the time the currently active speaker achieves full dominance and therewith counters the value $c_{\max}$. Therefore, the decreasing constant has to be recomputed for each channel m every time a speaker in any other channel $m'(m \neq m')$ corresponding to the same output signal subset becomes active:

$$c_{\mathrm{dec},m} = \begin{cases} \frac{c_m(\ell)-c_{\min}}{c_{\max}-c_{m'}(\ell)+\epsilon} \cdot c_{\mathrm{inc}}, & \text{if } \widehat{\mathrm{STD}}_{m'}(\ell) = 1 \wedge W^{\mathrm{ISC}}_{m,m'} = 0, \\ 0, & \text{else.} \end{cases} \tag{48}$$

with the very small value ϵ. The matrix $\mathbf{W}^{\mathrm{ISC}}$ avoids a decrease of the dominance of the mth speaker if a speaker related to a different output signal other than the currently considered output signal becomes active. To characterize the dominance of a speaker, finally, the counters have to be mapped to the speaker dominance weights by normalization of each counter to the sum of all counters similar to [24]

$$g_m^{\mathrm{DW}}(\ell) = \frac{c_m(\ell)}{\sum\limits_{m'=1}^{M} \left|1 - W^{\mathrm{ISC}}_{m,m'}\right| \cdot c_{m'}(\ell)}. \tag{49}$$

With the help of the dominance weights, an output signal-dependent reference noise PSD $\hat{\Phi}^{\mathrm{ref}}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}},m}(\ell,k)$ used for the dynamic spectral floor computation in (32) can be determined. Note that for input channels corresponding to the same output instance, this reference noise PSD has to be identical. Applying the dominance weights and the control matrix $\mathbf{W}^{\mathrm{ISC}}$ for involving only the noise PSDs of the relevant channels, we then have

$$\hat{\Phi}^{\mathrm{ref}}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}},m}(\ell,k) = \sum_{m'=1}^{M} \left|1 - W^{\mathrm{ISC}}_{m,m'}\right| \cdot g_{m'}^{\mathrm{DW}}(\ell) \cdot \hat{\Phi}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}},m'}(\ell,k), \tag{50}$$

where $\hat{\Phi}_{\tilde{\mathrm{N}}\tilde{\mathrm{N}},m'}(\ell,k)$ is the noise PSD estimate as introduced in (21). Figure 4 shows the dominance weights and the adjustment of the signal characteristics for a scenario, where four passengers in a car speak one after another and all signals are combined to one output signal instance $Q = 1$. Due to a slightly opened window at the front right passenger, the background noise is higher there compared to the other channels. The noise and speech signal show smooth transitions at speaker changes compared to hard switching between the channels.

Preferred parameters of the implementation of the reference value computation can be found in Table 2.

7 Evaluation

For evaluation purposes, a measurement database has been recorded in an Audi A6 with four distributed speaker-dedicated microphones. The driver and the front passenger each have a dedicated microphone located in the A-pillar. The microphones for the two backseat passengers are located in the ceiling in front of each seat. Speech and noise signals have been recorded separately to be combined to noisy signals afterwards. Based on this scenario, instrumental quality measures can be determined by evaluating the components before and after the processing. Clean speech signal components of eight speakers (four females and four males) speaking four different test utterances have been recorded for all four available seating positions in the car. To cause the Lombard effect, car noise with an average sound pressure level of around 65 dB(A) has been played back via headphones during the recording. Thus, the database includes 128 test sentences (eight speakers × four positions × four utterances). The noise signal components have been recorded for six different speeds (50, 80, 100, 130, 160, and 180 km/h) with all windows closed. Additionally noise scenarios with a slightly opened front right window were recorded for the first five speeds. In order to obtain realistic noisy microphone signals, the signal components

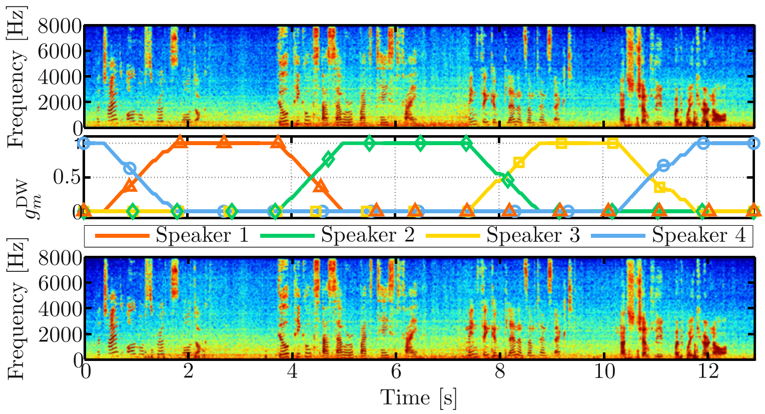

Figure 4 Speaker dominance weights and mixing example. Speaker dominance weights (middle) in case of a car driving at 160 km/h. Four passengers speaking one after another each with an SNR = 5 dB. The front right window is slightly opened. The spectrograms of the resulting noise-reduced ($\beta = -10$ dB) signal after hard switching between active channels (top) and of the combined signal after DSC (bottom) are shown.

Table 2 Preferred parameter values for the implementation of the reference noise power spectral density estimation

Parameter	Value
$c_{\max}$	100
$c_{\min}$	0
t_{inc}	1.2 s
T_{frame}	8 ms

are combined regarding ITU-T Recommendation P.56 [32] as presented by the authors in [27], with SNR $\in \{-5, 0, 5, 10, 15, 20\}$ dB. It is aimed at generating test signals where four different speakers assigned to the four various seats are speaking different utterances at different noise scenarios and SNRs one after another. Therefore, primarily, one speaker is chosen for each seating position randomly out of the whole measurement database. Therewith, an arrangement in the car with four speakers is simulated. Regarding the current evaluation, four such arrangements are chosen randomly with each speaker speaking four different utterances in the mentioned 11 noise conditions. Hence, we have 176 test signals for each position for six different SNRs.

A special analysis scenario has been picked from the whole dataset, where $M = 4$ speakers are active one after another at 0 dB with background noise of the car driving at 80 km/h. Between the second and third speakers, a short overlapping speech period is present. The two front passengers' signals (speakers 1 and 2) are mixed to one output instance, and the backseat passengers' signals (speakers 3 and 4) to a second one determined by the control matrices defined in (3) and (4). Thus, we have $Q = 2$ output signals. In Figure 5, different spectrograms are visualized. Besides the spectra of the raw microphone signals, several versions of the processed spectra are shown. The AGC has not been considered during these processings. For the spectrograms related to the complete speech enhancement system (excluding AGC), it is obvious that the cross-talk components are robustly suppressed, while chosen particular signals are combined to the appropriate two output signals.

To evaluate the whole system more generally, instrumental quality measures can be computed. Due to the combination of realistic noisy time-domain signals out of the separate signal components, the noisy signal can be processed by the proposed speech enhancement system, whereas the influence on each single component can be observed and evaluated afterwards. The system with $M = 4$ has been configured with $Q = 2$, and the noise reduction applies a maximum attenuation of $\beta = -12$ dB. Again, for the evaluation, the AGC is not included into the whole processing. Beside the speech-to-speech distortion ratio (SSDR) [33], a second measure called direct-to-cross-talk ratio (DCR) is introduced for evaluation. It is common to evaluate such quality measures in segments. Regarding [34] where a typical segment length between 15 and 20 ms is recommended, we choose a length of $N = 320$ at the underlying sampling frequency of $f_{\text{s}} = 16$ kHz which results to 20 ms. In order to measure the speech distortion, the SSDR can be computed based on the clean reference time-domain speech signal component $s_m(n)$ and the processed speech signal component $\tilde{s}_m(n)$. Note that the reference speech component in each mth channel is a combination of the direct component and the cross-talk components occurring in the other channels dependent on the channel selection in (34) in order to avoid a negative influence of exploitation of diversity effects. The SSDR in each frame λ can be written as [33]

$$\text{SSDR}_m(\lambda) = 10 \log_{10} \left[\frac{\sum\limits_{\nu=0}^{N-1} s_m^2(\nu + \lambda N)}{\sum\limits_{\nu=0}^{N-1} e_m^2(\nu + \lambda N)} \right], \tag{51}$$

whereas the speech distortion is defined as comprising the processed speech signal component $\tilde{s}_m(n)$ as

$$e_m(n) = \tilde{s}_m(n) - s_m(n). \tag{52}$$

It has to be ensured that the delay between $\tilde{s}_m(n)$ and $s_m(n)$ is compensated. After limitation of $\text{SSDR}_m(\lambda)$ to a maximum of $\text{SSDR}_{\max} = 30$ dB and a minimum of $\text{SSDR}_{\min} = -10$ dB, the segmental SSDR is proposed to be computed by

$$\text{SSDR}_{\text{seg},m} = \frac{1}{C(\Lambda_m)} \sum_{\lambda \in \Lambda_m} \text{SSDR}_m(\lambda). \tag{53}$$

The term Λ_m represents a subset of all those frames showing fullband voice activity for speaker m and where $\text{SSDR}_m(\lambda) > -10$ dB. $C(\Lambda_m)$ is the number of elements within this subset.

Regarding these subsets, similarly, a measure for the remaining cross-talk is computed. The segmental DCR is defined considering the processed direct signal $\tilde{s}_m(n)$ originating from the exclusively active source belonging to the mth channel and the related processed cross-talk components $\tilde{b}_{m',m}(n)$ occurring in the other distant channels m' and originating from the same source:

$$\text{DCR}_{\text{seg},m,m'} = \frac{1}{C(\Lambda'_m)} \sum_{\lambda \in \Lambda'_m} \text{DCR}_{m,m'}(\lambda), \tag{54}$$

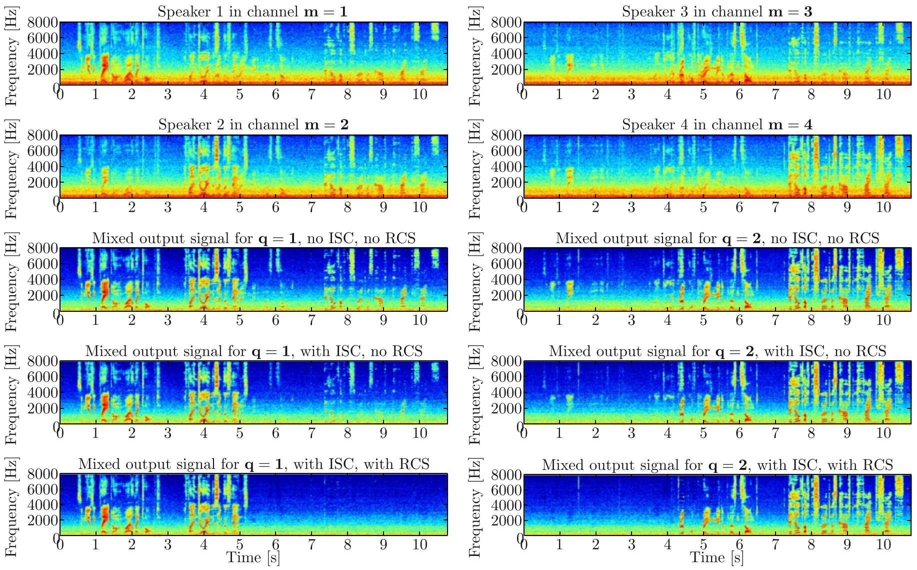

Figure 5 Processed signal spectra for different processings based on an example scenario. Noisy microphone spectra for each of $M = 4$ channels (first two rows) and the output signals for $Q = 2$ are depicted. Three different processings are shown: The processed two output signal spectrograms without ISC and RCS (third row), the processed signal with ISC but again without RCS (fourth row), and finally the outputs based on the overall processing including ISC as well as RCS (last row). The speakers are active one after another (driver, front passenger, rear left passenger, rear right passenger). Noise is superposed from a car driving at 80 km/h. Each passenger speaks at an SNR of 0 dB.

with Λ'_m representing voice active frames where additionally $\mathrm{DCR}_{m,m'}(\lambda) > -10$ dB. The DCR in each frame results in

$$\mathrm{DCR}_{m,m'}(\lambda) = 10\log_{10}\left[\frac{\sum\limits_{\nu=0}^{N-1}\tilde{s}_m^2(\nu+\lambda N)}{\sum\limits_{\nu=0}^{N-1}\tilde{b}_{m',m}^2(\nu+\lambda N)}\right]. \quad (55)$$

The value is limited to a maximum $\mathrm{DCR}_{\max} = 60$ dB and a minimum $\mathrm{DCR}_{\min} = -10$ dB before applying (54). Due to the presence of cross-talk components in multiple distant channels, we consider the mean segmental DCR for the mth channel across all available cross-talk components:

$$\mathrm{DCR}_{\mathrm{seg},m} = \frac{1}{C(\Xi_m)}\sum_{m'\in\Xi_m}\mathrm{DCR}_{\mathrm{seg},m,m'}\,. \quad (56)$$

Here, the number of channels where the cross-talk components are evaluated is specified by $C(\Xi_m)$, and Ξ_m is the subset of channel indices that are not related to the output signal the current mth channel is dedicated to. The mean values for these measures are determined across the whole test set for each SNR.

The results are depicted in Figure 6 showing the mean across all positions. The SNR is represented by the markers increasing from the bottom to the top (-5, 0, 5, 10, 15, and 20 dB). The basic processing without any cross-talk suppression already shows a relatively high DCR due to the attenuation of the active speaker's speech by the acoustic path. Based on this, the ISC performs a further cross-talk cancellation. The overall system with ISC and RCS attenuates the cross-talk components very well, indicated by higher DCR values, whereas the speech distortion remains nearly the same compared across the different processings. The higher the SNR, the lower is the speech distortion (higher SSDR). With exception of the Basic+ISC+RCS processing method, the variation for the DCR results across different SNRs is not as large due to masking effects and room acoustics. Including the RCS, it is obvious that a larger amount of cross-talk components can be suppressed at higher SNRs because it depends on the SAD in the active channel and is able to detect more speech active bins that are not masked by noise. Figure 7 shows similar results for the different positions exemplarily evaluated for one speaker in the front ($m = 1$) and one in the back ($m = 3$). The results for the front position differ slightly from the ones for the backseat position especially regarding the segmental DCR. This is expected due to the room acoustics and the higher amount of cross-talk speech components in the front microphones caused by the backseat speakers.

Now, we evaluate the fullband SAD introduced in Section 6.1 and outlined further in Appendix 2. Error rates are computed based on the comparison of the binary SAD results after the processing compared with a reference fullband SAD mask. The reference assumes speech activity if the clean speech signal component level is larger than a certain threshold. This threshold is chosen 40 dB below the maximum level of the whole clean speech signal. The

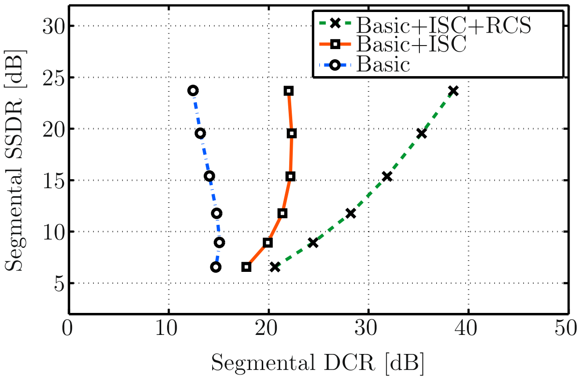

Figure 6 Performance evaluated in average over all positions. Performance of the different processing methods averaged over all positions. Increasing SNR represented by markers from bottom to top (-5, 0, 5, 10, 15, and 20 dB).

fullband reference mask $\mathrm{SAD}_{\mathrm{ref},m}(\ell)$ for each channel m is set to 1 if a minimum of 5% of all frequency subbands exceeds this threshold, otherwise it is zero. Error rates are computed for the basic SAD $\widetilde{\mathrm{SAD}}_m(\ell)$ and the enhanced one $\widehat{\mathrm{SAD}}_m(\ell)$, respectively. In case of the enhanced SAD, the fullband overall error in channel m for L signal time frames is computed by

$$E_m = \frac{1}{L}\sum_{\ell=1}^{L}\left|\mathrm{SAD}_{\mathrm{ref},m}(\ell) - \widehat{\mathrm{SAD}}_m(\ell)\right|, \qquad (57)$$

and accordingly for $\widetilde{\mathrm{SAD}}_m(\ell)$. In Figure 8, this overall error is depicted for the basic SAD (65) and the enhanced SAD (73) again for six different SNRs. The results are based on the mean SAD across all positions and conditions of the whole dataset. It is evident that the enhanced SAD yields a detection with a lower overall error. With higher SNRs, the overall error is decreasing, and the SAD seems to be more reliable.

Exemplarily formulated for the enhanced SAD, the false detections are covered by the false-positive rate, and the missed detections are measured by the false-negative rate:

$$r_{\mathrm{FP},m} = \frac{\sum_{\ell=1}^{L}\left(\widehat{\mathrm{SAD}}_m(\ell)\cdot\left(1-\mathrm{SAD}_{\mathrm{ref},m}(\ell)\right)\right)}{\sum_{\ell=1}^{L}\left(1-\mathrm{SAD}_{\mathrm{ref},m}(\ell)\right)} \quad \text{and}$$

$$r_{\mathrm{FN},m} = \frac{\sum_{\ell=1}^{L}\left(\left(1-\widehat{\mathrm{SAD}}_m(\ell)\right)\cdot\mathrm{SAD}_{\mathrm{ref},m}(\ell)\right)}{\sum_{\ell=1}^{L}\mathrm{SAD}_{\mathrm{ref},m}(\ell)}. \qquad (58)$$

Figure 9 shows the advantage of the enhanced SAD by lower false-positive rates. In contrast, the false-negative

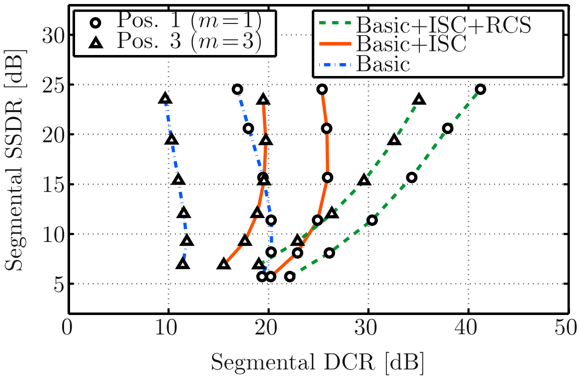

Figure 7 Performance evaluated for position $m = 1$ and $m = 3$. Performance of the different processing methods evaluated for the different positions $m = 1$ and $m = 3$. Increasing SNR represented by markers from bottom to top (-5, 0, 5, 10, 15, and 20 dB).

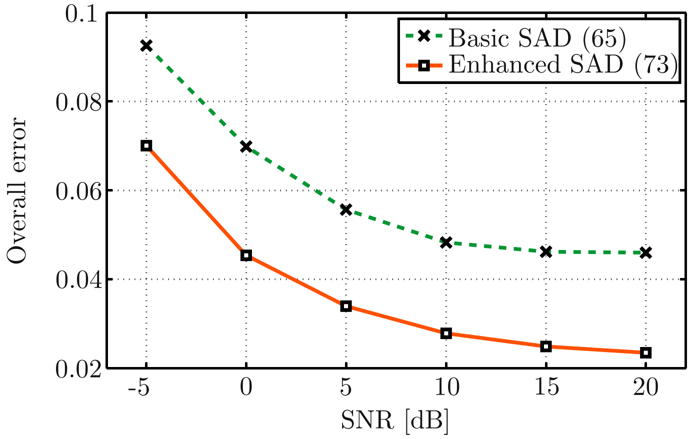

Figure 8 Overall error of the fullband SAD. Overall error of the fullband SAD (mean over all positions) for six different SNRs increasing from left to right.

rates are slightly higher. However, to avoid, e.g., the adaptation of adaptive filters to wrong events, it seems to be more important to obtain a lower false-positive rate.

8 Conclusions

In this contribution, a dynamic multi-channel system for speech signal enhancement in an automotive environment with distributed speaker-dedicated microphones has been presented. The proposed system supports multiple speakers in a car. It can be freely configured to obtain different mixed output signals that can be passed to various speech signal applications. Selected signals can be combined to different output signals by dynamic signal combining, whereas cross-talk components of signals not of interest are cancelled in these output signals within an interfering speaker cancellation approach and proper postprocessing. Furthermore, stationary noise is reduced.

The ability of the system to combine various input signals to several output signals has been shown. Furthermore, the suppression of interfering speech in each output signal has been evaluated by the computation of instrumental quality measures indicating speech distortion as well as cross-talk cancellation capability. Different configurations of the system have been investigated, showing the advantage of the complete speech enhancement system comprising interfering speaker cancellation, cross-talk suppression, and dynamic signal combination.

To control the whole system, a robust speaker activity detection based on signal power ratios has been proposed. Within the evaluation, it can be shown that an enhancement of the introduced basic fullband approach yields further improvements regarding detection rates.

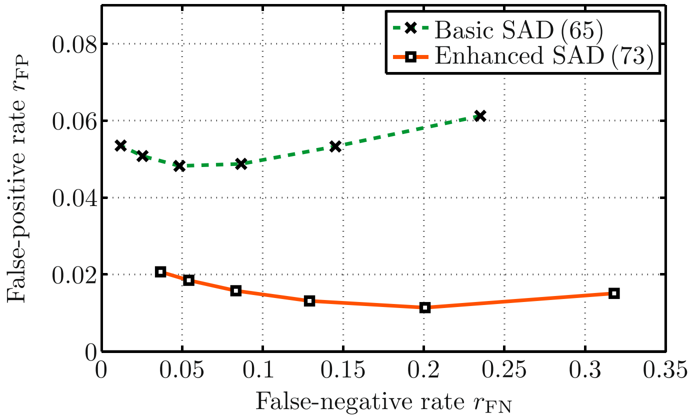

Figure 9 False-positive and false-negative error rates for the fullband SAD (mean over all positions). The SNR is decreasing from left to right (20, 15, 10, 5, 0, and -5 dB).

Instead of using only one microphone for each speaker, the proposed methods can also be applied to the output signals of multiple processed microphone subgroups. It may be advantageous to use a beamformer for each of the positions in the car to further improve the characteristics of the whole processing and to exploit the room acoustics furthermore by spatial filtering.

Appendices

Appendix 1: basic fullband speaker activity detection

The basic fullband SAD is based on the logarithmic quantity of the SPR estimate from (42), thus we write

$$\widehat{\mathrm{SPR}}'_m(\ell,k) = 10\log_{10}\left(\widehat{\mathrm{SPR}}_m(\ell,k)\right). \tag{59}$$

In order to consider only SPR values during periods showing a certain SNR (43) with $\hat{\xi}_m(\ell,k) > \Theta_{\mathrm{SNR1}}$, a modified quantity is defined by

$$\widetilde{\mathrm{SPR}}_m(\ell,k) = \begin{cases} \widehat{\mathrm{SPR}}'_m(\ell,k), \text{ if } \hat{\xi}_m(\ell,k) \geq \Theta_{\mathrm{SNR1}}, \\ 0, \qquad\qquad \text{else.} \end{cases} \tag{60}$$

To evaluate the SPR for each channel, it is observed how many positive (+) or negative (-) values for $\widetilde{\mathrm{SPR}}_m(\ell,k)$ are observed in each frame. Thus, a resulting positive counter follows

$$c_m^+(\ell) = \sum_{k=0}^{K/2} c_m^+(\ell,k), \qquad \text{with}$$

$$c_m^+(\ell,k) = \begin{matrix} 1, \text{ if } \widetilde{\mathrm{SPR}}_m(\ell,k) \geq 0, \\ 0, \text{ else.} \end{matrix} \tag{61}$$

Equivalently, it can be written for the negative counter:

$$c_m^-(\ell) = \sum_{k=0}^{K/2} c_m^-(\ell,k), \qquad \text{with}$$

$$c_m^-(\ell,k) = \begin{cases} 1, \text{ if } \widetilde{\mathrm{SPR}}_m(\ell,k) < 0, \\ 0, \text{ else.} \end{cases} \tag{62}$$

Based on these quantities and with an SNR-dependent soft weighting function $G_m^{\mathrm{c}}(\ell)$, a soft frame-based speaker activity detection measure can be formulated by

$$\chi_m^{\mathrm{SAD}}(\ell) = G_m^{\mathrm{c}}(\ell) \cdot \frac{c_m^+(\ell) - c_m^-(\ell)}{c_m^+(\ell) + c_m^-(\ell)}. \tag{63}$$

We compute the soft weighting function in (63) using *subgroup* SNRs as

$$G_m^{\mathrm{c}}(\ell) = \min\left\{\hat{\xi}^{\mathrm{G}}_{\max,m}(\ell)/10, 1\right\}. \tag{64}$$

For the calculation of the subgroup SNRs and the maximum SNR, see (83) and (84) in Appendix 4. Finally, the basic fullband SAD can be achieved by thresholding

$$\widetilde{\mathrm{SAD}}_m(\ell) = \begin{cases} 1, \text{ if } \chi_m^{\mathrm{SAD}}(\ell) > \Theta_{\mathrm{SAD1}}, \\ 0, \text{ else.} \end{cases} \tag{65}$$

Double-talk is detected based on a measure that evaluates whether the positive counter $c_m^+(\ell)$ exceeds a certain limit Θ_{DTM} during fullband detected speech activity in several channels. This result is held in each channel for some frames in order to detect continuous regions of double-talk. If the measure is true for more than one channel, general double-talk $\widehat{\mathrm{DTD}}(\ell) = 1$ is assumed. Preferred parameter settings for this section can be found in Table 3.

Appendix 2: enhanced fullband speaker activity detection based on multipath-induced fading patterns

An overview of the enhanced fullband SAD (dark-gray block in Figure 3) is depicted in Figure 10, where the dark-shaded area includes the SAD decision as well as the power ratio pattern determination. The bright-shaded area comprises the update of the reference pattern set. Parameter settings used in the following are represented in Table 4.

Power ratio patterns

As pointed out in Section 6.1, we want to exploit the characteristics of the SPR over frequency. Initially, we want to define a measure to highlight the multipath-induced fading subbands. Therefore, we aim at obtaining high values for the characteristic small power ratios and small values for inconspicuous and not relevant high power ratios. We propose a mapping yielding the following quantity [29]:

$$\chi_m^{\mathrm{PAT}}(\ell,k) = \max\left\{1 - \gamma_{\mathrm{PAT}} \cdot \widehat{\mathrm{SPR}}_m(\ell,k), \Theta_{\mathrm{PAT1}}\right\}. \tag{66}$$

Table 3 Preferred parameter settings for the implementation of the basic fullband speaker activity detection method

Parameters	Value
Θ_{SNR1}	0.25
Θ_{SAD1}	0.0025
Θ_{DTM}	30

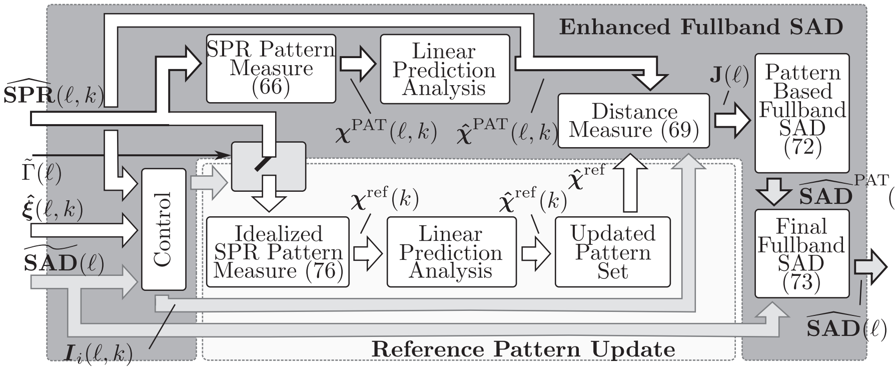
Enhanced Fullband SAD
SPR Pattern Measure (66)
Linear Prediction Analysis
$\widehat{\mathbf{SPR}}(\ell,k)$
$\chi^{\mathrm{PAT}}(\ell,k)$
$\hat{\chi}^{\mathrm{PAT}}(\ell,k)$
Distance Measure (69)
$\mathbf{J}(\ell)$
Pattern Based Fullband SAD (72)
$\tilde{\Gamma}(\ell)$
$\hat{\boldsymbol{\xi}}(\ell,k)$
$\widetilde{\mathbf{SAD}}(\ell)$
Control
$\chi^{\mathrm{ref}}(k)$
$\hat{\chi}^{\mathrm{ref}}(k)$
$\hat{\chi}^{\mathrm{ref}}$
Idealized SPR Pattern Measure (76)
Linear Prediction Analysis
Updated Pattern Set
$\widehat{\mathbf{SAD}}^{\mathrm{PAT}}(\ell)$
Final Fullband SAD (73)
$\widehat{\mathbf{SAD}}(\ell)$
$\boldsymbol{I}_i(\ell,k)$
Reference Pattern Update

Figure 10 Enhanced fullband SAD stage. Overview of the enhanced fullband SAD based on power ratio patterns (dark-gray block in Figure 3).

Large power ratios are mapped to the lower bound Θ_{PAT1}. γ_{PAT} allows the scalability of the behavior of the mapping function. Using $\gamma_{\mathrm{PAT}} < 1$ forces an underestimation of the power ratio $\widehat{\mathrm{SPR}}_m(\ell,k)$. Hence, the limit for highlighting subbands as multipath-induced fading ones can be controlled. A strong underestimation is appropriate to highlight the subbands that are anomalously highly attenuated by the room acoustics in the considered channel m. Even positive but small power ratios are evaluated in this case. In order to obtain a smoothed spectrum indicating the position of the multipath-induced fading subbands, a linear prediction analysis is performed. The autocorrelation coefficients $\varphi_{p,m}(\ell)$ are computed by the inverse discrete Fourier transform of the magnitude squares of the quantity $\chi_m^{\mathrm{PAT}}(\ell,k)$. Thus, the Yule-Walker auto-regressive equations for solving the prediction problem with order N_{p} and the filter coefficients $a_{i,m}(\ell)$ are

$$\varphi_{p,m}(\ell) = \sum_{i=1}^{N_{\mathrm{p}}} a_{i,m}(\ell)\cdot\varphi_{p-i,m}(\ell),\; p = 1,2,\ldots,N_{\mathrm{p}}. \tag{67}$$

Table 4 Preferred parameter settings for the implementation of the proposed enhanced fullband speaker activity detection method

Parameter	Value
Θ_{PAT1}	0.05
γ_{PAT}	0.25
N_{p}	100
N_{PAT}	80
Θ_{SNR2}	0.25
L_{PAT}	150
Θ_{SAD2}	40
Θ_{COH}	0.02
Θ_{PAT2}	Θ_{PAT1}
Θ_{SNR3}	2
Θ_{PAT3}	-6

After applying the Levinson-Durbin algorithm and using the frequency response of the filter coefficients $a_{i,m}(\ell)$ represented by $A_m(\ell,k)$, the logarithmic estimate of $\chi_m^{\mathrm{PAT}}(\ell,k)$ is recovered by

$$\hat{\chi}_m^{\mathrm{PAT}}(\ell,k) = 10\log_{10}\left(\left|\frac{E_m(\ell,k)}{1-A_m(\ell,k)}\right|\right), \tag{68}$$

with the prediction error signal $E_m(\ell,k)$ used for normalization. Based on these patterns, an enhanced speaker activity detection can be performed by comparing the currently observed pattern $\hat{\chi}_m^{\mathrm{PAT}}(\ell,k)$ with a reference pattern set that is characteristic for the active speaker's location. The reference pattern set consists of N_{PAT} different patterns which shall represent the characteristics of the specific speaker positions including some variations. A Euclidean distance measure $\tilde{J}_{i,m}(\ell,k)$ between each reference pattern $\hat{\chi}_{i,m}^{\mathrm{ref}}(k)$ with $i = 1,\ldots,N_{\mathrm{PAT}}$ and the currently estimated pattern is determined:

$$\tilde{J}_{i,m}(\ell,k) = \left(\hat{\chi}_{i,m}^{\mathrm{ref}}(k) - \hat{\chi}_m^{\mathrm{PAT}}(\ell,k)\right)^2. \tag{69}$$

The mean value of this distance measure $\tilde{J}_{i,m}(\ell,k)$ over the relevant subbands is a quantity for the detection of the activity of the mth speaker:

$$\bar{J}_{i,m}(\ell) = \frac{1}{N_{i,m}}\sum_{k=0}^{K/2}\tilde{J}_{i,m}(\ell,k)\cdot\mathrm{I}_{i,m}(\ell,k), \tag{70}$$

with $N_{i,m} \in \{1,\ldots,K/2+1\}$ subbands to evaluate for each pattern. The function $\mathrm{I}_{i,m}(\ell,k)$ indicates the subbands where an evaluation of the patterns seems to be reasonable. For small values of $N_{i,m}$, the previous distance measure is used. To draw reliable conclusions from the distance measure during fullband detected single-talk speech periods, an SNR of Θ_{SNR2} has to be exceeded.

Furthermore, only those subbands should be evaluated, where multipath-induced fading subbands occur either in $\hat{\chi}^{\text{ref}}_{i,m}(k)$ or in $\hat{\chi}^{\text{PAT}}_{m}(\ell,k)$ indicated by some peaks. For SAD, the best matching pattern $\hat{\chi}^{\text{ref}}_{j_m,nm}(k)$ is further analyzed with

$$j_m = \underset{i\in\{1,\ldots,N_{\text{PAT}}\}}{\operatorname{argmin}} \left\{\bar{J}_{i,m}(\ell)\right\}. \tag{71}$$

In order to detect speech regions rather than single-speech active frames, a minimum for $\bar{J}_{j_m,m}(\ell)$ over L_{PAT} past frames is determined during basic fullband SAD. This minimum is denoted by $J_m(\ell)$. The resulting pattern-based SAD indicator function $\widehat{\text{SAD}}^{\text{PAT}}_{m}(\ell)$ is obtained by comparing $J_m(\ell)$ with a threshold based on its tracked global minimum $\Theta_{m,\min}(\ell)$ including an additional offset Θ_{SAD2}:

$$\widehat{\text{SAD}}^{\text{PAT}}_{m}(\ell) = \begin{cases} 1, & \text{if } J_m(\ell) < (\Theta_{m,\min}(\ell) + \Theta_{\text{SAD2}}), \\ 0, & \text{else.} \end{cases} \tag{72}$$

If $J_m(\ell)$ is close to zero, it should force $\widehat{\text{SAD}}^{\text{PAT}}_{m}(\ell) = 0$ due to the challengeable reliability. In combination with the basic fullband SAD in (65), the final enhanced fullband SAD is obtained:

$$\widehat{\text{SAD}}_m(\ell) = \widetilde{\text{SAD}}_m(\ell) \cdot \widehat{\text{SAD}}^{\text{PAT}}_{m}(\ell). \tag{73}$$

Reference pattern set

Due to the room acoustics in a car, the occurring patterns may change over time if the speaker slightly moves. We propose to update the reference pattern set $\hat{\boldsymbol{\chi}}^{\text{ref}}_{m}(k)$ during the processing within a first in-first out system of length N_{PAT} by including new patterns $\hat{\chi}^{\text{ref}}_{i,m}(k)$ as was proposed similarly by the authors in [29]. Only if speaker activity can be assumed quite likely, the occurring pattern shall be included into the reference pattern set. Beside the basic SAD, a fullband coherence measure is used for accepting new reference patterns in order to further reduce misdetections. The magnitude squared coherence (MSC) can be computed between two channels m and m' with the cross PSD $\hat{\Phi}_{\text{YY},m,m'}(\ell,k)$ and the two auto PSDs $\hat{\Phi}_{\text{YY},m}(\ell,k)$ and $\hat{\Phi}_{\text{YY},m'}(\ell,k)$ [35]. With the appropriate SNR threshold $\Theta_{\text{SNR3}} = 2$, it follows for a modified coherence

$$\tilde{\Gamma}_{m,m'}(\ell,k) = \begin{cases} \dfrac{\left|\hat{\Phi}_{\text{YY},m,m'}(\ell,k)\right|^2}{\hat{\Phi}_{\text{YY},m}(\ell,k)\cdot\hat{\Phi}_{\text{YY},m'}(\ell,k)}, & \text{if } \hat{\xi}_m(\ell,k) > \Theta_{\text{SNR3}}, \\ 0, & \text{else.} \end{cases} \tag{74}$$

To obtain a channel-independent *fullband* coherence quantity $\tilde{\Gamma}(\ell)$, we determine the mean MSC measure over all subbands and search for the maximum of these quantities over all channel combinations:

$$\tilde{\Gamma}(\ell) = \max_{\substack{m,m'\in\{1,\ldots,M\}\\ m\neq m'}} \left\{ \frac{1}{K/2+1} \sum_{k=0}^{K/2} \tilde{\Gamma}_{m,m'}(\ell,k) \right\}. \tag{75}$$

Because, furthermore, only the characteristic subbands should occur as peaks in the reference pattern set, using a modified measure for highlighting the multipath-induced fading subbands is proposed. New patterns are included if three fullband conditions are fulfilled: The basic fullband SAD in (65) with a stricter threshold $\Theta_{\text{SAD1}} = 0.5$ has to indicate speech, whereas double-talk must not occur and a certain threshold Θ_{COH} has to be exceeded by the coherence measure $\tilde{\Gamma}(\ell)$. Instead of simply including the currently appearing spectrum from (68) into the reference pattern set, the calculation from (66) is modified to

$$\chi^{\text{ref}}_{i,m}(\ell,k) = \begin{cases} \chi^{\text{PAT}}_{m}(\ell,k), & \text{if } \mathrm{I}^{\text{ref}}_{i,m}(\ell,k) > 0, \\ \Theta_{\text{PAT2}}, & \text{else,} \end{cases} \tag{76}$$

for obtaining the reference patterns. The reference indicator function $\mathrm{I}^{\text{ref}}_{i,m}(\ell,k)$ includes a new characteristic frequency subband into the current reference pattern if the frequency-selective SNR quantity $\hat{\xi}_m(\ell,k)$ is larger than a threshold Θ_{SNR3}. Furthermore, the currently occurring pattern has to show a peak at some frequencies and therewith has to exceed a threshold of Θ_{PAT3} dB there. Otherwise, the constant Θ_{PAT2} is set. Based on this modified quantity, the linear prediction is performed, and the reference pattern set can be updated by the new entry $\hat{\chi}^{\text{ref}}_{i,m}(k)$.

Appendix 3: frequency-selective speaker activity detection

An overview of the frequency-selective SAD (light-gray block in Figure 3) is shown in Figure 11, where the first block describes the SPR model and its adaptation, and the second block shows the model-based SAD as already presented similarly by the authors in [31]. It is supposed that the SPR in the mth channel can be represented by the random variable $\mathcal{Y}$, where one realization can take the value $\widehat{\text{SPR}}'_m(\ell,k) = 10\log_{10}\left(\widehat{\text{SPR}}_m(\ell,k)\right)$. We assume that this SPR in the mth channel is normally distributed in each subband with $(\mathcal{Y}|H_{1,m}) \sim \mathcal{N}(\mu_m,\sigma_m^2)$ during voice activity of the mth speaker indicated by the hypothesis $H_{1,m}$. Hence, the conditional probability density function of $\mathcal{Y}$ may be modeled by a single Gaussian

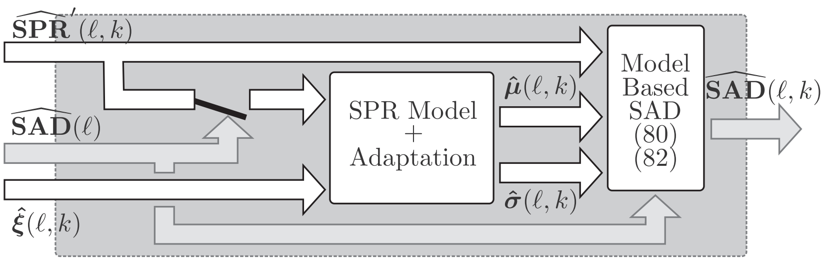
$\widehat{\mathbf{SPR}}'(\ell,k)$
$\widehat{\mathbf{SAD}}(\ell)$
$\hat{\boldsymbol{\xi}}(\ell,k)$
SPR Model + Adaptation
$\hat{\boldsymbol{\mu}}(\ell,k)$
$\hat{\boldsymbol{\sigma}}(\ell,k)$
Model Based SAD (80) (82)
$\widehat{\mathbf{SAD}}(\ell,k)$

Figure 11 Frequency-selective SAD. Overview of the frequency-selective SAD stage (light-gray block in Figure 3).

distribution [36] with mean $\mu_m(\ell,k)$ and variance $\sigma_m n^2(\ell,k)$:

$$p_{\mathcal{Y}|H_{1,m}}\Big(\widehat{\mathrm{SPR}}'_m(\ell,k)\Big) = \frac{1}{\sqrt{2\pi}\,\sigma_m(\ell,k)} \cdot \exp\left\{-\frac{\Big(\widehat{\mathrm{SPR}}'_m(\ell,k)-\mu_m(\ell,k)\Big)^2}{2\sigma_m^2(\ell,k)}\right\}. \quad (77)$$

For modeling this distribution in one channel, the mean value and the variance have to be estimated during single-talk periods of the related speaker, where an SNR of at least Θ_{SNR4} has to be exceeded. Otherwise, the previous result from the last frame is used. The mean value $\mu_m(\ell,k)$ can be estimated by smoothing the SPR over time with the constant γ_μ [31]:

$$\hat{\mu}_m(\ell,k) = \gamma_\mu \cdot \hat{\mu}_m(\ell-1,k) + (1-\gamma_\mu)\cdot \widehat{\mathrm{SPR}}'_m(\ell,k). \quad (78)$$

Simultaneously, an estimate for the variance $\sigma_m^2(\ell,k)$ can be calculated with the smoothing constant γ_σ:

$$\hat{\sigma}_m^2(\ell,k) = \gamma_\sigma \cdot \hat{\sigma}_m^2(\ell-1,k) + (1-\gamma_\sigma)\cdot\Big(\widehat{\mathrm{SPR}}'_m(\ell,k) - \hat{\mu}_m(\ell,k)\Big)^2. \quad (79)$$

Hence, the SAD may be determined based on the model parameters without considering the sign of the SPR value itself. The decision whether speech is detected for an observed $\widehat{\mathrm{SPR}}'_m(\ell,k)$ is made based on the model in (77) in combination with the estimated parameters. For a positive decision, the probability density function has to reach a certain threshold Θ_{p}, and full-band speaker activity and no double-talk have to be detected. Therewith, for the frequency-selective SAD, it follows

$$\widehat{\mathrm{SAD}}'_m(\ell,k) = \begin{cases} 1, & \text{if} p_{\mathcal{Y}|H_{1,m}}\Big(\widehat{\mathrm{SPR}}'_m(\ell,k)\Big) > \Theta_{\mathrm{p}} \\ & \wedge\ \widehat{\mathrm{SAD}}_m(\ell) = 1 \wedge \widehat{\mathrm{DTD}}(\ell) = 0, \\ \delta_{m,m_{\mathrm{pmax}}}, & \text{if } \widehat{\mathrm{DTD}}(\ell) = 1, \\ 0, & \text{else.} \end{cases} \quad (80)$$

During double-talk, the channel related to the maximum resulting modified SPR is determined by the Kronecker delta. For the second index, we have

$$m_{\mathrm{pmax}} = \underset{m\in\{1,\ldots,M\}}{\operatorname{argmax}} \left\{\widetilde{\mathrm{SPR}}_m(\ell,k)\right\}. \quad (81)$$

The final frequency-selective SAD results after comparing the SNR estimate with the limit Θ_{SNR4}

$$\widehat{\mathrm{SAD}}_m(\ell,k) = \begin{cases} \widehat{\mathrm{SAD}}'_m(\ell,k), & \text{if } \hat{\xi}_m(\ell,k) \geq \Theta_{\mathrm{SNR4}}, \\ 0, & \text{else.} \end{cases} \quad (82)$$

During activity of more than one speaker, it can be still distinguished between the different speakers in a frequency-selective manner due to the assumption of the sparseness of speech across the subbands. Preferred parameter settings can be found in Table 5.

Table 5 Preferred parameter settings for the implementation of the frequency-selective model-based speaker activity detection method

Parameter	Value
γ_μ	0.83
γ_σ	0.8
Θ_{p}	0.01
Θ_{SNR4}	0.25

Appendix 4: signal-to-noise ratio subgroups

Further processing (regarding a DFT length of $K = 512$ and a sampling frequency of $f_s = 16$ kHz) grouped SNR values can be computed for $K' = 10$ different frequency subgroups, each covering DFT bin $k_{\ae}, \ldots, k_{\ae+1} - 1$, with $\ae =, 2, \ldots, K'$ and $\{k_{\ae}\} = \{4, 28, 53, 78, 103, 128, 153, 178, 203, 228, 253\}$. For the mean SNR computed for the æth subgroup follows

$$\hat{\xi}_m^{\mathrm{G}}(\ell, \ae) = \frac{1}{k_{\ae+1} - k_{\ae}} \sum_{k=k_{\ae}}^{k_{\ae+1}-1} \hat{\xi}_m(\ell, k+1). \tag{83}$$

Then, the maximum SNR across the SNRs of the frequency subgroups is given by

$$\hat{\xi}_{\max,m}^{\mathrm{G}}(\ell) = \max_{\ae \in \{1,\ldots,K'\}} \left\{ \hat{\xi}_m^{\mathrm{G}}(\ell, \ae) \right\}. \tag{84}$$

Abbreviations

AGC: Automatic gain control; DCR: Direct-to-cross-talk ratio; DFT: Discrete Fourier transform; DSC: Dynamic signal combination; ENR: Extended noise reduction; ISC: Interfering speaker cancellation; MSC: Magnitude squared coherence; PSD: Power spectral density; RCS: Residual cross-talk suppression; SAD: Speaker activity detection; SE: Signal enhancement; SNR: Signal-to-noise ratio; SPR: Signal power ratio; SSDR: Speech-to-speech distortion ratio.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Nuance Communications Deutschland GmbH, Acoustic Speech Enhancement Research, Ulm D-89077, Germany. [2]Technische Universität Braunschweig, Institute for Communications Technology, Braunschweig D-38106, Germany.

Received: 11 July 2013 Accepted: 11 December 2013
Published: 27 December 2013

References

1. M Brandstein, D Ward, (eds.), *Microphone Arrays: Signal Processing Techniques and Applications*, 1st edn. (Springer, Berlin, 2001)
2. BD Van Veen, KM Buckley, Beamforming: A versatile approach to spatial filtering. IEEE ASSP Mag. **5**(2), 4–24 (1988)
3. J Freudenberger, S Stenzel, B Venditti, Microphone diversity combining for in-car applications. EURASIP J. Adv. Signal Process. **2010**, 1–13 (2010)
4. T Gerkmann, R Martin. Soft decision combining for dual channel noise reduction, in *Proceedings of the International Conference on Spoken Language Processing (INTERSPEECH)* (Pittsburgh, Pennsylvania, USA, 17–21 Sept 2006), pp. 2134–2137
5. H Banno, T Shinde, K Takeda, F Itakura. In-car speech recognition using distributed microphones: adapting to automatically detected driving conditions, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Hong Kong, China, 6–10 April 2003), pp. I-324–I-327
6. W Li, K Takeda, F Itakura. Optimizing regression for in-car speech recognition using multiple distributed microphones, in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)* (Jeju, Korea, 4–8 Oct 2004), pp. 2689–2692
7. Y Shimizu, S Kajita, K Takeda, F Itakura. Speech recognition based on space diversity using distributed multi-microphone, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Istanbul, Turkey, 5–9 June 2000), pp. III-1747–III-1750
8. F Hummes, J Qi, T Fingscheidt. Robust Acoustic Speaker Localization with Distributed Microphones, in *Proceedings of the European Signal Processing Conference (EUSIPCO)* (Barcelona, Spain, 29 Aug–2 Sept 2011), pp. 240–244
9. B Widrow, JR Glover, JM McCool, J Kaunitz, CS Williams, RH Hearn, JR Zeidler, E Dong, RC Goodlin, Adaptive noise cancelling: principles and applications. Proc. IEEE **63**(12), 1692–1716 (1975)
10. A Hirano, K Nakayama, S Arai, M Deguchi, A low-distortion noise canceller and its learning algorithm in presence of crosstalk. IEICE Trans. Fundamentals Electron. Commun. Comput. Sci. **E84-A**(2), 414–421 (2001)
11. A Lombard, W Kellermann. Multichannel cross-talk cancellation in a call-center scenario using frequency-domain adaptive filtering, in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)* (Seattle, Washington, USA, 14–17 Sept 2008)
12. E Robledo-Arnuncio, BH Juang. Blind source separation of acoustic mixtures with distributed microphones, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Honolulu, Hawai, USA, 15–20 April 2007), pp. III-949–III-952
13. JP Dmochowski, Z Liu, PA Chou. Blind source separation in a distributed microphone meeting environment for improved teleconferencing, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Las Vegas, Nevada, USA, 30 March–4 April 2008), pp. 88–92
14. R Aichner, M Zourub, H Buchner, W Kellermann. Residual cross-talk and noise suppression for convolutive blind source separation, in *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)* (Braunschweig, Germany, 20–23 March 2006), pp. 41–42
15. S Han, J Cui, P Li. Post-processing for frequency-domain blind source separation in hearing aids, in *Proceedings of the International Conference on Information, Communications and Signal Processing (ICICS)* (Macau, China, 8–10 Dec 2009), pp. 356–360
16. M Jeub, C Herglotz, CM Nelke, C Beaugeant, P Vary. Noise reduction for dual-microphone mobile phones exploiting power level differences, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Kyoto, Japan, 25–30 March 2012), pp. 1693–1696
17. MM Sondhi, DR Morgan, JL Hall, Stereophonic acoustic echo cancellation–an overview of the fundamental problem. IEEE Signal Process. Lett. **2**(8), 148–151 (1995)
18. H Buchner. Acoustic echo cancellation for multiple reproduction channels: from first principles to real-time solutions, in *Proceedings of the ITG-Fachtagung Sprachkommunikation* (Aachen, Germany, 8–10 Oct 2008), pp. 1–4
19. J Bourgeois, W Minker, *Time-Domain Beamforming and, Blind Source Separation* (Springer, Heidelberg, 2009)
20. A Sugiyama, ed. by E Hänsler, G Schmidt. Low-distortion noise cancellers—Revival of a classical technique, in *Speech and Audio Processing in Adverse Environments* (Springer Berlin, 2008), pp. 229–264
21. S Haykin, *Adaptive Filter Theory*, 4th edn. (Prentice Hall, Upper Saddle River, 2002)
22. T Matheja, M Buck, T Wolff. Robust adaptive cancellation of interfering speakers for distributed microphone systems in cars, in *Proceedings of the Deutsche Jahrestagung für Akustik (DAGA)* (Berlin, Germany, 15–18 March 2010), pp. 255–256
23. E Hänsler, G Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, vol. 1 (Wiley, Hoboken, 2004)
24. T Matheja, M Buck, A Eichentopf. Dynamic signal combining for distributed microphone systems in car environments, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Prague, Czech Republic, 22–27 May 2011), pp. 5092–5095
25. K Linhard, T Haulick. Noise subtraction with parametric recursive gain curves, in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)* (Budapest, Hungary, 5–9 Sept 1999), pp. 2611–2614
26. I Cohen, Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Process. **11**(5), 466–475 (2003)
27. T Matheja, M Buck, T Fingscheidt. A multi-channel quality assessment setup applied to a distributed microphone speech enhancement system with spectral boosting, in *Proceedings of the ITG-Fachtagung Sprachkommunikation* (Braunschweig, Germany, 26–28 Sept 2012), pp. 119–122
28. T Matheja, M Buck, T Fingscheidt. Speaker activity detection for distributed microphone systems in cars, in *Proceedings of the 6th Biennial*

Workshop on Digital Signal Processing for In-Vehicle Systems (Seoul, Korea, 29 Sept–2 Oct 2013)

29. T Matheja, M Buck, T Wolff. Enhanced speaker activity detection for distributed microphones by exploitation of signal power ratio patterns, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Kyoto, Japan, 25–30 March 2012), pp. 2501–2504
30. R Martin. An efficient algorithm to estimate the instantaneous SNR of speech signals, in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)* (Berlin, Germany, 22–25 Sept 1993), pp. 1093–1096
31. T Matheja, M Buck. Robust voice activity detection for distributed microphones by modeling of power ratios, in *Proceedings of the ITG-Fachtagung Sprachkommunikation* (Bochum, Germany, 6–8 Oct 2010)
32. International Telecommunication Union, *ITU-T Recommendation P56, Objective Measurement of Active Speech Level* (International Telecommunication Union, Geneva, 1993)
33. T Fingscheidt, S Suhadi. Quality assessment of speech enhancement systems by separation of enhanced speech, noise, and echo, in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)* (Antwerp, Belgium, 27–31 Aug 2007), pp. 818–821
34. PC Loizou, *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, 2013)
35. GC Carter, Coherence and time delay estimation. Proc. IEEE **75**(2), 236–255 (1987)
36. E Hänsler, *Statistische Signale - Grundlagen und Anwendungen*, 3rd edn. (Springer, Berlin, 2001)

doi:10.1186/1687-6180-2013-191
Cite this article as: Matheja *et al.*: A dynamic multi-channel speech enhancement system for distributed microphones in a car environment. *EURASIP Journal on Advances in Signal Processing* 2013 **2013**:191.