The 6th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems, Sep. 29-Oct. 2, 2013, Seoul, Korea

Speaker Activity Detection for Distributed Microphone Systems in Cars

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

[1]: Nuance Communications Deutschland GmbH, Niederlassung Ulm, Acoustic Speech Enhancement Research, Germany
[2]: Technische Universität Braunschweig, Institute for Communications Technology, Germany
E-mail: {timo.matheja, markus.buck}@nuance.com; t.fingscheidt@tu-bs.de

Abstract In this contribution a new framework for energy-based acoustic speaker activity detection for distributed microphones in automotive environments is presented. The method relies on the evaluation of signal power ratios between the available multiple speaker-dedicated microphone signals. To obtain a robust fullband as well as a robust frequency-selective speaker activity detection, the overall framework comprises three main parts. It is shown that the proposed detection mechanism yields an improvement in error rates compared to applying a basic speaker activity detection only.

Keywords Distributed microphones, signal power ratios, speaker activity detection, control of speech enhancement systems, automotive speech application

1. INTRODUCTION

In speech communication systems in automotive environments the interest in more comfort within hands-free telephony or speech dialog systems is increasing. Distributed and speaker dedicated microphones mounted close to each passenger in the car enable all speakers, e.g., to participate in hands-free conference phone calls at the same time. To control the necessary speech signal processing, such as adaptive filter or signal combination algorithms within distributed microphone setups, it has to be known which speaker is speaking at which time instance. A possible application scenario may be the activation of a speech dialog system by an utterance of a specific speaker. Due to the special arrangement of the microphones close to the particular speakers it is possible to exploit the different and characteristic signal power ratios occurring between the available microphone channel signals. Based on this information an energy-based speaker activity detection (SAD) can be performed as addressed in [1, 2, 3].

Regarding beamforming approaches, the signal power ratio between desired and interfering signal also is meaningful to be evaluated, e.g., as in [4]. However, in case of speaker-dedicated microphones, the particular microphone signals may be used directly to represent the desired and the interfering speech signal component [1, 2, 5, 6]. A control mechanism for a further signal processing unit based on signal powers only may suffer from special room acoustics, undesired noisy transient signals or, e.g., interfering backseat passengers who do not have a dedicated microphone. This problem is often not further investigated in energy-based methods as in [1] or signal evaluations linked to spatial cues in general [7, 8]. Alternatively to an energy-based approach in [9] phase differences are exploited to distinguish between different sectors where either the driver or the front passenger shows speech activity in a two-channel system in a car.

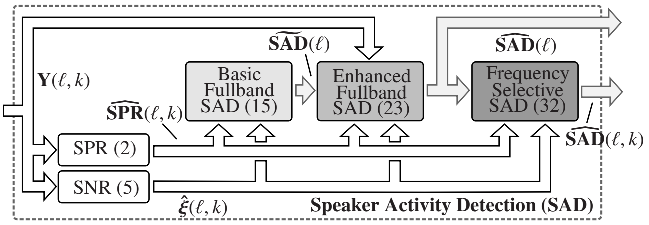
$\mathbf{Y}(\ell,k)$
$\widehat{\mathbf{SAD}}(\ell)$
Basic Fullband SAD (15)
Enhanced Fullband SAD (23)
$\widehat{\mathbf{SAD}}(\ell)$
Frequency Selective SAD (32)
$\widehat{\mathbf{SPR}}(\ell,k)$
$\widehat{\mathbf{SAD}}(\ell,k)$
SPR (2)
SNR (5)
$\hat{\boldsymbol{\xi}}(\ell,k)$
Speaker Activity Detection (SAD)

Fig. 1. Speaker activity detection framework overview (references to following equations are added in round brackets).

In this contribution a SAD framework for the arrangement of speaker-dedicated distributed cardioid microphones is proposed. The paper is organized as follows: In Section 2 an overview of the processing framework is given. The speaker activity detection is presented in Section 3 including the determination of the signal power ratios, the basic as well as the enhanced fullband SAD mechanism and the frequency-selective SAD. The subsequent Section 4 considers the evaluation of the overall system. Finally some conclusions are drawn.

2. SYSTEM OVERVIEW

The concept of the proposed energy-based SAD framework is based on the evaluation of the resulting signal power ratio (SPR) in each of $M \geq 2$ microphone channels. The processing is performed in the discrete Fourier transform domain with the frame index ℓ and the frequency subband index k at a sampling rate of $f_s = 16\,\text{kHz}$. The time domain signal is segmented by a Hann window with a frame length of $K = 512$ samples and a frame shift of 25 %. An overview of the whole framework is depicted in Fig. 1 where the overall speaker activity detection framework mainly consists of three stages (gray blocks in Fig. 1). Bold arrows and symbols indicate the multi-channel signals stacked in vectors. Using the microphone signal spectra $\mathbf{Y}(\ell, k)$, the power ratio $\widehat{\mathbf{SPR}}(\ell, k)$ and the signal-to-noise ratio (SNR) $\hat{\boldsymbol{\xi}}(\ell, k)$ are computed. Applying these quantities the basic fullband speaker activity detection $\widetilde{\mathbf{SAD}}(\ell)$, the enhanced fulllband detector $\widehat{\mathbf{SAD}}(\ell)$ and the frequency-selective speaker activity detector $\widehat{\mathbf{SAD}}(\ell, k)$ are determined.

Within the first stage the basic fullband SAD is determined based on the SPR and the SNR values. As an elementary idea it can be distinguished between different active speakers by analyzing how many positive and negative values occur for the logarithmic SPR in each frame for each channel m. To further enhance the robustness, the room acoustics are exploited in the second stage. Due to *multipath-induced fading* the signal power in a distant microphone can be larger than in the speaker-dedicated one while a speaker is active close to his dedicated microphone. The location of these special subbands is caused by the room acoustics. Thus, they may be assumed to be characteristic for the position of the speaker in the car cabin and can be observed as a distinguishable feature. While the effect of the *multipath-induced fading* subbands can be exploited to determine an enhanced fullband SAD a simple frequency-selective SAD within the third framework stage would suffer from this fading effect. It is proposed to model the SPR by a Gaussian distribution during the processing. An active speaker is identified based on this model. In the end the three SAD stages are combined to form a fullband as well as a frequency-selective SAD result. For realization of the speaker activity detection system no training period is necessary.

3. ROBUST SPEAKER ACTIVITY DETECTION

The three stages of the SAD framework are based on the evaluation of the characteristic signal power ratio (SPR) occurring in each microphone channel. Subsequently the overall signal processing is described in detail starting with the determination of the signal power ratios followed by the elaboration of the three SAD stages.

3.1. Signal Power Ratios

Before considering the SAD methods it has to be focused on the determination of the SPRs. Assuming that speech and noise components are uncorrelated and that the microphone signal spectra are a superposition of speech and noise components, the speech signal power spectral density (PSD) estimate $\hat{\Phi}_{\Sigma\Sigma,m}(\ell, k)$ in channel m can be determined by

$$\hat{\Phi}_{\Sigma\Sigma,m}(\ell, k) = \max\left\{\hat{\Phi}_{\mathrm{YY},m}(\ell, k) - \hat{\Phi}_{\mathrm{NN},m}(\ell, k), 0\right\}, \quad (1)$$

where $\hat{\Phi}_{\mathrm{YY,m}}(\ell, k)$ may be estimated by temporal smoothing of the squared magnitude of the microphone signal spectra $Y_m(\ell, k)$. The noise PSD estimate $\hat{\Phi}_{\mathrm{NN},m}(\ell, k)$ can be determined, e.g., by the improved minimum controlled recursive averaging approach [10]. Note that within the measure in (1) direct speech components originating from the speaker related to the considered microphone are included as well as cross-talk components from other sources and speakers. The SPR in each channel m is expressed similar to [3] for a system with $M \geq 2$ microphones as

$$\widehat{\mathrm{SPR}}_m(\ell, k) = \frac{\max\left\{\hat{\Phi}_{\Sigma\Sigma,m}(\ell, k), \epsilon\right\}}{\max\left\{\max\limits_{\substack{m' \in \{1,\ldots,M\} \\ m' \neq m}}\left\{\hat{\Phi}_{\Sigma\Sigma,m'}(\ell, k)\right\}, \epsilon\right\}}, \quad (2)$$

with the small value ϵ. It is assumed that one microphone always captures the speech best because each speaker has a dedicated microphone close to his position. Thus, the active speaker can be identified by evaluating the SPR values among the available microphones. Furthermore, the logarithmic SPR quantity enhances differences for lower values and results in

$$\widehat{\mathrm{SPR}}'_m(\ell, k) = 10 \log_{10}\left(\widehat{\mathrm{SPR}}_m(\ell, k)\right). \quad (3)$$

3.2. Basic Fullband Speaker Activity Detection

Basically, speech activity in the m-th speaker related microphone channel can be detected by evaluating if the occurring logarithmic SPR is larger than 0 dB. To avoid considering the SPR during periods where the SNR $\xi_m(\ell, k)$ shows only small values lower than a threshold Θ_{SNR1}, a modified quantity for the logarithmic power ratio in (3) is defined by

$$\widetilde{\mathrm{SPR}}_m(\ell, k) = \begin{cases} \widehat{\mathrm{SPR}}'_m(\ell, k), & \text{if } \hat{\xi}_m(\ell, k) \geq \Theta_{\mathrm{SNR1}}, \\ 0, & \text{else.} \end{cases} \quad (4)$$

With a noise estimate $\hat{\Phi}'_{\mathrm{NN},m}(\ell, k)$ for determination of a reliable SNR quantity the SNR is determined similar to [11]:

$$\hat{\xi}_m(\ell, k) = \frac{\min\left\{\hat{\Phi}_{\mathrm{YY,m}}(\ell, k), |Y_m(\ell, k)|^2\right\} - \hat{\Phi}'_{\mathrm{NN},m}(\ell, k)}{\hat{\Phi}'_{\mathrm{NN},m}(\ell, k)}. \quad (5)$$

Using the overestimation factor γ_{SNR} the considered noise PSD results in

$$\hat{\Phi}'_{\mathrm{NN},m}(\ell, k) = \gamma_{\mathrm{SNR}} \cdot \hat{\Phi}_{\mathrm{NN},m}(\ell, k). \quad (6)$$

Based on (4) the power ratios are evaluated by observing how many positive (+) or negative (-) values occur in each frame. Hence, for the positive counter follows:

$$c_m^+(\ell) = \sum_{k=0}^{K/2} c_m^+(\ell, k), \quad (7)$$

with

$$c_m^{+}(\ell,k) = \begin{cases} 1, & \text{if } \widetilde{\mathrm{SPR}}_m(\ell,k) > 0 \ , \\ 0, & \text{else.} \end{cases} \tag{8}$$

Equivalently the negative counter can be determined by

$$c_m^{-}(\ell) = \sum_{k=0}^{K/2} c_m^{-}(\ell,k) \ , \tag{9}$$

considering

$$c_m^{-}(\ell,k) = \begin{cases} 1, & \text{if } \widetilde{\mathrm{SPR}}_m(\ell,k) < 0 \ , \\ 0, & \text{else.} \end{cases} \tag{10}$$

Regarding these quantities a soft frame-based SAD measure may be written by

$$\chi_m^{\mathrm{SAD}}(\ell) = G_m^{\mathrm{c}}(\ell) \cdot \frac{c_m^{+}(\ell) - c_m^{-}(\ell)}{c_m^{+}(\ell) + c_m^{-}(\ell)} \ , \tag{11}$$

where $G_m^{\mathrm{c}}(\ell)$ is an SNR-dependent soft weighting function to pay more attention to high SNR periods. In order to consider the SNR within certain frequency regions the weighting function is computed by applying maximum *subgroup* SNRs:

$$G_m^{\mathrm{c}}(\ell) = \min\left\{\hat{\xi}_{\max,m}^{\mathrm{G}}(\ell)/10, 1\right\} \ . \tag{12}$$

The maximum SNR across K' different frequency subgroup SNRs $\hat{\xi}_m^{\mathrm{G}}(\ell,\text{\ae})$ is given by

$$\hat{\xi}_{\max,m}^{\mathrm{G}}(\ell) = \max_{\text{\ae}\in\{1,\ldots,K'\}} \left\{\hat{\xi}_m^{\mathrm{G}}(\ell,\text{\ae})\right\} \ . \tag{13}$$

The grouped SNR values can each be computed in the range between certain DFT bins $k_{\text{\ae}}$ and $k_{\text{\ae}+1}$ with $\text{\ae} = 1, 2, ..., K'$ and $\{k_{\text{\ae}}\} = \{4, 28, 53, 78, 103, 128, 153, 178, 203, 228, 253\}$. We write for the mean SNR in the æ-th subgroup:

$$\hat{\xi}_m^{\mathrm{G}}(\ell,\text{\ae}) = \frac{1}{k_{\text{\ae}+1} - k_{\text{\ae}}} \sum_{k=k_{\text{\ae}}+1}^{k_{\text{\ae}+1}} \hat{\xi}_m(\ell,k) \ . \tag{14}$$

Finally, the basic fullband SAD is obtained by thresholding using Θ_{SAD1}:

$$\widetilde{\mathrm{SAD}}_m(\ell) = \begin{cases} 1, & \text{if } \chi_m^{\mathrm{SAD}}(\ell) > \Theta_{\mathrm{SAD1}} \ , \\ 0, & \text{else.} \end{cases} \tag{15}$$

During double-talk situations the evaluation of the signal power ratios is not reliable anymore. Thus, regions of double-talk have to be detected in order to be careful within the SAD. Considering the positive and negative counters, a double-talk measure can be determined by evaluating if $c_m^{+}(\ell)$ exceeds a limit Θ_{DTM} during periods of detected fullband speech activity in several channels. To detect regions of double-talk this result is held for some frames in each channel. General double-talk $\widehat{\mathrm{DTD}}(\ell) = 1$ is detected if the measure is true for more than one channel. Preferred parameter settings for the realization of the basic fullband SAD can be found in Tab. 1.

Table 1. Preferred parameter settings for the implementation of the basic fullband SAD algorithm (for $M = 4$).

$\Theta_{\mathrm{SNR1}} = 0.25$	$\gamma_{\mathrm{SNR}} = 4$	$K' = 10$
$\Theta_{\mathrm{SAD1}} = 0.0025$	$\Theta_{\mathrm{DTM}} = 30$	

3.3. Enhanced Fullband Speaker Activity Detection

By simply evaluating the power ratios like presented in the basic approach, the risk still exists to mistake transient interferers like blinker noise, outside passing cars or speech from interfering speakers for speech components of a certain speaker. In a two-channel system where only the two front passengers have dedicated microphones, incorrect detections may occur during speech activity of some backseat passengers. As was first introduced by the authors in [3] the robustness for these and for other similar situations can be increased by applying an enhanced fullband SAD method based on the evaluation of SPR patterns. This additional algorithm follows subsequently to the basic SAD as depicted in Fig. 1. The room acoustics of the special geometry can be exploited by this further step. A lower amount of energy may occur in the speaker's dedicated m-th microphone signal compared to the energy in a distant microphone channel. This results in a negative logarithmic SPR. Due to multipath propagation effects a sharp decline of energy may occur in some special so-called *multipath-induced fading* subbands in the speaker dedicated microphone channel signals. The number and location of these subbands is characteristic for the position of a sound source in the car and can be identified within the SPR. Appropriate SPR patterns representing this effect along the frequency may indicate the activity of a considered speaker if the observed pattern matches a reference pattern out of a determined reference pattern set.

Primarily a measure for highlighting the *multipath-induced fading* subbands is defined. Therefore, high values have to be obtained for the characteristic small power ratios within these distinguishing subbands. Inconspicuous and not relevant high SPRs should be mapped to a lower bound Θ_{PAT1}. According to [3] we propose a mapping yielding the following quantity:

$$\chi_m^{\mathrm{PAT}}(\ell,k) = \max\left\{1 - \gamma_{\mathrm{PAT}} \cdot \widehat{\mathrm{SPR}}_m(\ell,k), \Theta_{\mathrm{PAT1}}\right\} , \tag{16}$$

where γ_{PAT} enables a scalability of the characteristics of the mapping. The limit for highlighting subbands as *multipath-induced fading* ones can be modified, e.g., by underestimating the linear power ratio $\widehat{\mathrm{SPR}}_m(\ell,k)$ using $\gamma_{\mathrm{PAT}} < 1$. By applying a strong underestimation also subbands are highlighted that are anomalously heavily attenuated by the room acoustics but where the speaker dedicated microphone signal still shows the highest energy.

For indication of the position of the *multipath-induced fading* subbands a smoothed spectrum is obtained by performing a linear prediction analysis. Therefore, autocorrelation coefficients $\varphi_{p,m}(\ell)$ are determined by the inverse discrete Fourier transform of the magnitude squares of $\chi_m^{\mathrm{PAT}}(\ell,k)$. For solving the prediction problem the Yule-Walker autoregressive equations with order N_p and the filter coefficients $a_{i,m}(\ell)$ are

$$\varphi_{p,m}(\ell) = \sum_{i=1}^{N_{\mathrm{p}}} a_{i,m}(\ell) \cdot \varphi_{p-i,m}(\ell),\ p = 1,2,...,N_{\mathrm{p}}\ . \quad (17)$$

The logarithmic estimate of $\chi_m^{\mathrm{PAT}}(\ell,k)$ in (16) can be recovered after applying the Levinson-Durbin algorithm and using the frequency response of the filter coefficients $a_{i,m}(\ell)$ represented by $A_m(\ell,k)$:

$$\hat{\chi}_m^{\mathrm{PAT}}(\ell,k) = 10\log_{10}\left(\left|\frac{E_m(\ell,k)}{1-A_m(\ell,k)}\right|\right)\ , \quad (18)$$

where $E_m(\ell,k)$ is the prediction error signal used for normalization. Based on this observed pattern $\hat{\chi}_m^{\mathrm{PAT}}(\ell,k)$ an enhanced SAD can be realized by comparison with a reference pattern that is characteristic for the speaker's location and orientation. The reference pattern is part of a reference pattern set with N_{PAT} entries representing the state of the currently considered speaker's position including some variations. Between each reference pattern $\hat{\chi}_{i,m}^{\mathrm{ref}}(k)$ with $i = 1,...,N_{\mathrm{PAT}}$ and the currently observed pattern a Euclidean distance measure is determined as [3]

$$\tilde{J}_{i,m}(\ell,k) = \left(\hat{\chi}_{i,m}^{\mathrm{ref}}(k) - \hat{\chi}_m^{\mathrm{PAT}}(\ell,k)\right)^2\ . \quad (19)$$

A quantity for the detection of the speaker in the m-th channel may be obtained by computing a weighted mean value of this distance measure over frequency with $N_{i,m} \in \{1,...,K/2{+}1\}$ subbands that have to be evaluated for each pattern:

$$\overline{J}_{i,m}(\ell) = \frac{1}{N_{i,m}} \sum_{k=0}^{K/2} \tilde{J}_{i,m}(\ell,k) \cdot \mathrm{I}_{i,m}(\ell,k)\ . \quad (20)$$

The subbands where an evaluation of the patterns seems to be reasonable are indicated by the indicator function $\mathrm{I}_{i,m}(\ell,k)$. Therefore, an SNR of Θ_{SNR2} has to be exceeded during basically detected fullband single-talk periods. Furthermore, only those subbands should be evaluated where some peaks occur either in $\hat{\chi}_{i,m}^{\mathrm{ref}}(k)$ or in $\hat{\chi}_m^{\mathrm{PAT}}(\ell,k)$ meaning that some *multipath-induced fading* subbands exist in the current frame. Please note that the previous distance measure is used for small values of $N_{i,m}$ due to the purpose of reliability of the resulting patterns. For the SAD the best matching pattern $\hat{\chi}_{j_m,m}^{\mathrm{ref}}(k)$ is analyzed with

$$j_m = \underset{i\in\{1,...,N_{\mathrm{PAT}}\}}{\mathrm{argmin}} \left\{\overline{J}_{i,m}(\ell)\right\}\ . \quad (21)$$

It is aimed at detecting speech regions rather than single speech active frames. Thus, during basic fullband SAD the minimum distance measure $J_m(\ell)$ is obtained by searching the minimum of $\overline{J}_{j_m,m}(\ell)$ over the past L_{PAT} frames. Finally the pattern-based SAD measure $\widehat{\mathrm{SAD}}_m^{\mathrm{PAT}}(\ell) \in \{0,1\}$ is determined by comparing $J_m(\ell)$ with a threshold based on its tracked global minimum $\Theta_{m,\min}(\ell)$ including an additional offset Θ_{SAD2} as

$$\widehat{\mathrm{SAD}}_m^{\mathrm{PAT}}(\ell) = \begin{cases} 1, & \text{if } J_m(\ell) < (\Theta_{m,\min}(\ell) + \Theta_{\mathrm{SAD2}})\ , \\ 0, & \text{else.} \end{cases} \quad (22)$$

The enhanced SAD is defined in combination with the basic fullband SAD in (15) as

$$\widehat{\mathrm{SAD}}_m(\ell) = \widetilde{\mathrm{SAD}}_m(\ell) \cdot \widehat{\mathrm{SAD}}_m^{\mathrm{PAT}}(\ell)\ . \quad (23)$$

In order to realize the computation of the distance measure in (19) the reference patterns have to be determined. Due to the changing room acoustics in a car during movements of a speaker we propose to update the reference pattern set by including new patterns $\hat{\chi}_{i,m}^{\mathrm{ref}}(k)$ within a first-in first-out system of length N_{PAT}. New patterns should only be considered if speech activity can be assumed quite likely. Therefore, the basic SAD (15) and a fullband coherence measure may be used to decide whether a new pattern is accepted for the reference pattern set. The coherence quantity is used in order to focus on sound sources that cause coherent microphone signals. Within the considered setup this is the case for the desired speech signals. As presented in [12] the well-known magnitude squared coherence (MSC) can be computed between two channels m and m' regarding the cross PSD $\hat{\Phi}_{\mathrm{YY,m,m'}}(\ell,k)$ and the two auto PSDs $\hat{\Phi}_{\mathrm{YY},m}(\ell,k)$ and $\hat{\Phi}_{\mathrm{YY},m'}(\ell,k)$. Applying an appropriate SNR threshold Θ_{SNR3}, a modified coherence is written as

$$\tilde{\Gamma}_{m,m'}(\ell,k) = \begin{cases} \frac{\left|\hat{\Phi}_{\mathrm{YY,m,m'}}(\ell,k)\right|^2}{\hat{\Phi}_{\mathrm{YY},m}(\ell,k)\cdot\hat{\Phi}_{\mathrm{YY},m'}(\ell,k)}, & \text{if } \hat{\xi}_m(\ell,k) > \Theta_{\mathrm{SNR3}}\ , \\ 0, & \text{else.} \end{cases} \quad (24)$$

For the purpose of obtaining a channel-independent and fullband coherence measure $\tilde{\Gamma}(\ell)$ the mean of the modified coherence over all subbands is determined, and it is searched for the maximum of the resulting values over all channel combinations afterwards:

$$\tilde{\Gamma}(\ell) = \max_{\substack{m,m'\in\{1,...,M\} \\ m\neq m'}} \left\{ \frac{1}{K/2{+}1} \sum_{k=0}^{K/2} \tilde{\Gamma}_{m,m'}(\ell,k) \right\}\ . \quad (25)$$

Because only the characteristic subbands should be highlighted and should occur as peaks in the reference patterns a

Table 2. Preferred parameter settings for the implementation of the enhanced fullband SAD algorithm (for $M = 4$).

$\Theta_{\text{PAT1}} = 0.05$	$\gamma_{\text{PAT}} = 0.25$	$N_{\text{p}} = 100$
$N_{\text{PAT}} = 80$	$\Theta_{\text{SNR2}} = 0.25$	$L_{\text{PAT}} = 150$
$\Theta_{\text{SAD2}} = 40$	$\Theta_{\text{SNR3}} = 2$	$\Theta_{\text{SAD3}} = 0.5$
$\Theta_{\text{COH}} = 0.02$	$\Theta_{\text{PAT2}} = \Theta_{\text{PAT1}}$	$\Theta_{\text{PAT3}} = -6$

modified measure may be used for indicating the *multipath-induced fading* subbands. Three conditions have to be fulfilled for including new patterns in the reference pattern set: Speech has to be indicated by the basic fullband SAD in (15) with a stricter threshold $\Theta_{\text{SAD1}} = \Theta_{\text{SAD3}}$ in the absence of double-talk, whereas the coherence measure $\tilde{\Gamma}(\ell)$ has to reach a certain threshold Θ_{COH}. For obtaining the reference pattern to be included in the reference pattern set, the expression in (16) is modified previously to the linear prediction analysis:

$$\chi_{i,m}^{\text{ref}}(\ell,k) = \begin{cases} \chi_m^{\text{PAT}}(\ell,k), & \text{if } \mathrm{I}_{i,m}^{\text{ref}}(\ell,k) > 0 \ , \\ \Theta_{\text{PAT2}}, & \text{else.} \end{cases} \tag{26}$$

The reference pattern indicator function $\mathrm{I}_{i,m}^{\text{ref}}(\ell,k)$ indicates inclusion of a new characteristic frequency subband in the currently observed reference pattern if the frequency-selective SNR $\hat{\xi}_m(\ell,k)$ exceeds a threshold Θ_{SNR3}. Furthermore, the current pattern has to be larger than Θ_{PAT3} dB at some frequency subbands to verify the existence of a *multipath-induced fading* subband.

In this way the reference pattern set is updated and the enhanced SAD is obtained by evaluation of the proposed distance measure. Preferred parameter settings for the implementation of the algorithm can be found in Tab. 2.

3.4. Frequency-Selective Speaker Activity Detection

While the effect of the *multipath-induced fading* subbands can be exploited to determine an enhanced fullband SAD, a simple frequency-selective SAD within the third framework stage (cf. Fig. 1) would suffer from this fading effect. It can only be roughly distinguished between the speakers in a subband-selective manner by simply evaluating if a positive or negative logarithmic SPR occurs. Hence, we propose to model the SPR by a Gaussian distribution [2]. The model parameters are estimated during detected fullband speaker activity, and the resulting probability density function is evaluated by comparing with a threshold.

It is supposed that the SPR is described by a random variable $\mathcal{Y}$ in the m-th channel where one realization is represented by $\widehat{\text{SPR}}'_m(\ell,k) = 10\log_{10}\left(\widehat{\text{SPR}}_m(\ell,k)\right)$ as denoted in (3). During speech activity of speaker m indicated by the hypothesis $H_{1,m}$ a normal distribution of the SPR is assumed with $(\mathcal{Y}|H_{1,m}) \sim \mathcal{N}(\mu_m, \sigma_m^2)$. Thus, the conditional probability density function of $\mathcal{Y}$ can be modeled by a single Gaussian distribution regarding mean $\mu_m(\ell,k)$ and variance $\sigma_m^2(\ell,k)$:

$$p_{\mathcal{Y}|H_{1,m}}\left(\widehat{\text{SPR}}'_m(\ell,k)\right) = \frac{e^{-\frac{\left(\widehat{\text{SPR}}'_m(\ell,k)-\mu_m(\ell,k)\right)^2}{2\sigma_m^2(\ell,k)}}}{\sqrt{2\pi}\sigma_m(\ell,k)} \ . \tag{27}$$

To define this distribution more closely the values for the mean and the variance have to be estimated. This is done during single-talk periods of the related m-th speaker. The mean $\mu_m(\ell,k)$ is determined by smoothing the SPR over time with the smoothing constant γ_μ [2]:

$$\hat{\mu}_m(\ell,k) = \quad \gamma_\mu \cdot \hat{\mu}_m(\ell-1,k) + (1-\gamma_\mu) \cdot \widehat{\text{SPR}}'_m(\ell,k) \ . \tag{28}$$

Similarly the variance $\sigma_m^2(\ell,k)$ can be estimated by applying the constant γ_σ:

$$\begin{aligned} \hat{\sigma}_m^2(\ell,k) = \gamma_\sigma \cdot \hat{\sigma}_m^2(\ell-1,k) + \\ (1-\gamma_\sigma) \cdot \left(\widehat{\text{SPR}}'_m(\ell,k) - \hat{\mu}_m(\ell,k)\right)^2 \ . \end{aligned} \tag{29}$$

Whether speech is detected for an observed $\widehat{\text{SPR}}'_m(\ell,k)$ or not is determined based on the SPR model in (27) regarding the estimated parameters. A certain threshold Θ_p has to be reached by the probability density function, and fullband speaker activity in the absence of double-talk has to be detected for a positive decision of the frequency-selective SAD:

$$\begin{aligned} &\widehat{\text{SAD}}'_m(\ell,k) = \\ &\begin{cases} 1, & \text{if } p_{\mathcal{Y}|H_{1,m}}\left(\widehat{\text{SPR}}'_m(\ell,k)\right) > \Theta_p \\ & \wedge\, \widehat{\text{SAD}}_m(\ell){=}1 \wedge \widehat{\text{DTD}}(\ell){=}0, \\ \delta_{m,m_{\text{pmax}}}, & \text{if } \widehat{\text{DTD}}(\ell) = 1, \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{30}$$

During double-talk periods the active channel is selected by the Kronecker delta that indicates the channel related to the maximum resulting modified SPR as defined in (4). The second index results in

$$m_{\text{pmax}} = \underset{m\in\{1,\ldots,M\}}{\text{argmax}} \left\{\widetilde{\text{SPR}}_m(\ell,k)\right\} . \tag{31}$$

After comparing the SNR estimate with a limit Θ_{SNR4} we have for the final frequency-selective SAD:

$$\widehat{\text{SAD}}_m(\ell,k) = \begin{cases} \widehat{\text{SAD}}'_m(\ell,k), & \text{if } \hat{\xi}_m(\ell,k) \geq \Theta_{\text{SNR4}} \ , \\ 0, & \text{else.} \end{cases} \tag{32}$$

Due to the sparseness of speech it could still be distinguished between different active sources in a frequency-selective manner. For the implementation of this algorithm preferred parameter settings are depicted in Tab. 3.

Table 3. Preferred parameter settings for the implementation of the frequency-selective model-based SAD algorithm.

$\gamma_\mu = 0.83$	$\gamma_\sigma = 0.8$	$\Theta_p = 0.01$	$\Theta_{\mathrm{SNR4}} = 0.25$

4. EVALUATION

For evaluation of the SAD performance a measurement database recorded in an Audi A6 with four distributed speaker dedicated microphones is applied. Both driver and front passenger each have a dedicated microphone close to their heads in the A-pillar. The microphones of the two backseat passengers are located in the ceiling in front of each seat. Speech and noise signal components have been recorded independently and are combined to noisy microphone signals afterwards. The database comprises clean speech signal components of four female and four male speakers speaking four different Harvard test utterances obtained from [13] for all four available seating positions in the car. During the recording, car noise with an average sound pressure level of around 65 dB(A) has been played back via headphones to cause a Lombard effect. Hence, the database comprises 128 test sentences (8 speakers x 4 positions x 4 utterances). Noise signals have been recorded for six different speeds (50 km/h, 80 km/h, 100 km/h, 130 km/h, 160 km/h, 180 km/h) with all windows closed. Additionally for the first five scenarios, signals with a slightly opened front right window were recorded.

Realistic noisy microphone signals are obtained by superposition of the signal components according to ITU-T Recommendation P.56 as presented by the authors in [14]. Each noisy signal simulation for the following evaluation comprises four different speakers assigned to the four various seats and speaking four different test utterances one after another at different noise scenarios and SNRs, whereas SNR $\in \{-5, 0, 5, 10, 15, 20\}$ dB. For the considered evaluation four such arrangements are simulated. The speakers are randomly selected out of the whole measurement dataset to obtain a certain signal subset. We observe two different setups depicted in Fig. 2: A four-channel system in setup 1 as well as a two-channel system in setup 2. In both cases the same signals are simulated but for the latter the speaker activity detection is processed based on two microphone signals only (for $M = 2$ with $\Theta_{\mathrm{DTM}} = 60$ and $\Theta_{\mathrm{SAD2}} = 50$).

It is focused on evaluating the SAD by computing error rates based on the comparison of the binary SAD results with a reference SAD mask. To obtain initially a frequency-selective SAD reference mask $\mathrm{SAD}_{\mathrm{ref,m}}(\ell, k)$, it is analyzed where the level in the spectrum of the clean speech signal component of the considered speaker exceeds a certain threshold. If this threshold is reached in some subbands the fullband SAD reference mask $\mathrm{SAD}_{\mathrm{ref,m}}(\ell)$ for each channel m is set to one, otherwise zero. Error rates are determined for the basic as well as for the enhanced SAD. In case of the enhanced speaker activity detector $\widehat{\mathrm{SAD}}_m(\ell)$, the fullband overall error related to channel m is computed for L signal frames by:

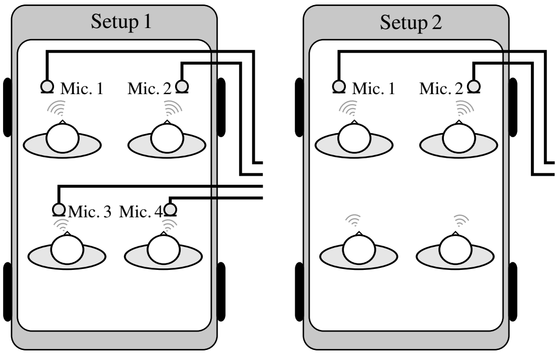

Fig. 2. Different simulated setups for evaluation. Left: Four-channel setup (setup 1). Right: Two-channel setup (setup 2).

$$E_m = \sum_{\ell=1}^{L} \left| \mathrm{SAD}_{\mathrm{ref,m}}(\ell) - \widehat{\mathrm{SAD}}_m(\ell) \right|, \qquad (33)$$

and accordingly for the basic SAD $\widetilde{\mathrm{SAD}}_m(\ell)$. In Fig. 3 this overall error is depicted for both fullband SAD methods for the six different SNRs regarding the four-channel setup (setup 1). The values are based on the mean SAD results across all positions, conditions and arrangements regarding the selected data subset. The advantage of the enhanced SAD that yields a detection with a lower overall error is visible. With higher SNRs the overall error is decreasing and the SAD seems to be more reliable. Similar results are shown for the two-channel setup (setup 2) in Fig. 4. Here the advantage of the enhanced SAD and the difference between the basic and enhanced method is even more obvious because the basic SAD suffers from the two interfering backseat passengers who do not have a dedicated microphone. The overall error is not decreasing with higher SNRs due to the fact that the interfering backseat speakers are not masked by noise anymore, and the risk of wrong detections increases.

The advantages of the fullband enhanced SAD also have influence on the frequency-selective SAD because the frequency-selective SAD model is adapted regarding a previously determined fullband SAD. Thus, a lower error rate of the frequency-selective SAD in (30) is expected if the enhanced fullband SAD is used instead of the basic version. According to (33) with the additional sum across all subbands a frequency-selective overall error can be computed by applying the introduced reference signal spectrum and the frequency-selective detector $\widehat{\mathrm{SAD}}'_m(\ell, k)$. Fig. 5 depicts the overall error of the frequency-selective SAD for each of the two underlying fullband SAD methods for the four-channel

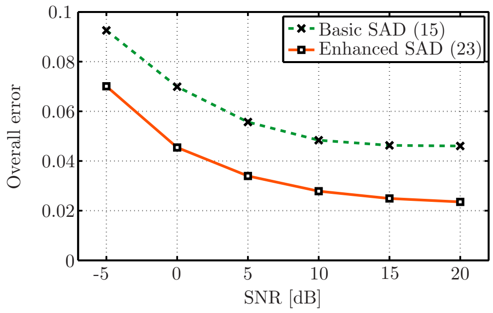

Fig. 3. Setup 1: Overall error of the basic and the enhanced fullband SAD (mean over all four positions) for six different SNRs.

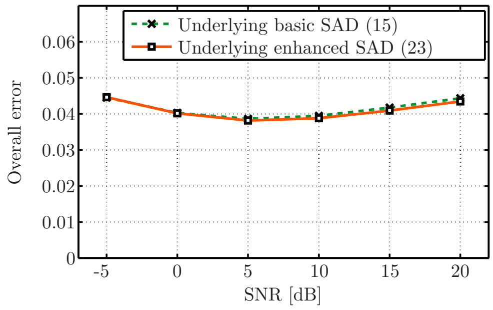

Fig. 5. Setup 1: Overall error of the frequency-selective SAD (mean over all four positions) for the underlying basic and the enhanced fullband SAD. Six different SNRs are considered.

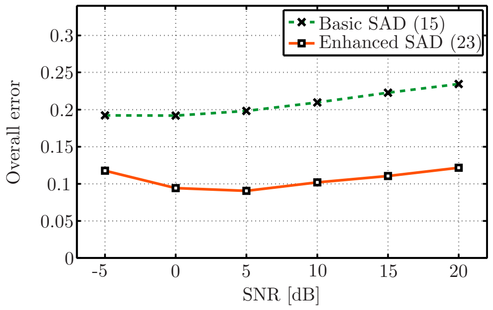

Fig. 4. Setup 2: Overall error of the basic and the enhanced fullband SAD (mean over all two positions) for six different SNRs.

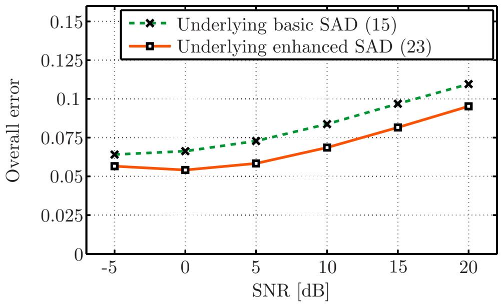

Fig. 6. Setup 2: Overall error of the frequency-selective SAD (mean over all two positions) for the underlying basic and the enhanced fullband SAD. Six different SNRs are considered.

setup. Referring to the different SNR values the differences in the error rate are really small. The results for the two-channel setup are shown in Fig. 6 where the differences are more clearly visible at an overall higher error rate.

To get more differentiating error curves for the fullband SAD false-positive and false-negative rates are determined to illustrate the false as well as the missed detections. The false-positive rate is computed again exemplarily for the enhanced SAD by

$$r_{\mathrm{FP},m} = \frac{\sum\limits_{\ell=1}^{L}\left(\widehat{\mathrm{SAD}}_m(\ell)\cdot(1-\mathrm{SAD}_{\mathrm{ref,m}}(\ell))\right)}{\sum\limits_{\ell=1}^{L}(1-\mathrm{SAD}_{\mathrm{ref,m}}(\ell))}\,, \qquad (34)$$

and the false-negative rate by

$$r_{\mathrm{FN},m} = \frac{\sum\limits_{\ell=1}^{L}\left((1-\widehat{\mathrm{SAD}}_m(\ell))\cdot\mathrm{SAD}_{\mathrm{ref,m}}(\ell)\right)}{\sum\limits_{\ell=1}^{L}\mathrm{SAD}_{\mathrm{ref,m}}(\ell)}\,. \qquad (35)$$

The results are depicted in Fig. 7 for the first setup. The false-positive rates are lower for the enhanced SAD while in contrast the false-negative rates are slightly higher. To make a system more robust and to avoid, e.g., the adaptation of adaptive filters to wrongly detected events, it seems to be more important to obtain a lower false-positive rate. Similar results can be observed in Fig. 8 for the two-channel setup. Here the error rates in general are higher due to the interfering backseat passengers but again the difference between the underlying basic and the enhanced fullband SAD is more obvious than in the four-channel scenario.

5. CONCLUSIONS

In this contribution a speaker activity detection framework for speaker-dedicated and distributed cardioid microphones has been presented. The proposed system comprises three stages including a basic as well as an enhanced fullband and a frequency-selective speaker activity detector. The particular three algorithms for each part of the overall framework focus on the exploitation of signal power ratios between the available microphone signals. The methods have been described in detail, and both a binary decision for a fullband and for a frequency-selective speaker activity detection are obtained. It can be shown that the error rates of the detection algorithms are improved by applying an enhanced fullband speaker activity detection method exploiting the characteristic room acoustics instead of using an introduced basic approach.

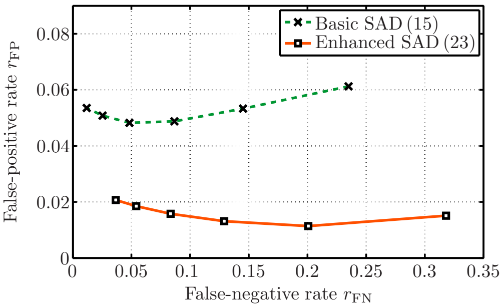

Fig. 7. Setup 1: False-positive and false-negative rates of the basic and the enhanced fullband SAD (mean over all four positions). The SNR is decreasing from left to right (20, 15, 10, 5, 0, -5 dB).

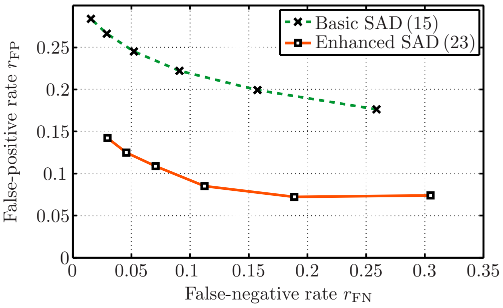

Fig. 8. Setup 2: False-positive and false-negative rates of the basic and the enhanced fullband SAD (mean over all two positions). The SNR is decreasing from left to right (20, 15, 10, 5, 0, -5 dB).

6. REFERENCES

[1] A. Lombard, W. Kellermann, “Multichannel Cross-Talk Cancellation in a Call-Center Scenario Using Frequency-Domain Adaptive Filtering,” in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, Seattle, Washington, USA, Sep. 2008.

[2] T. Matheja, M. Buck, “Robust Voice Activity Detection for Distributed Microphones by Modeling of Power Ratios,” in *Proc. ITG-Fachtagung Sprachkommunikation*, Bochum, Germany, Oct. 2010.

[3] T. Matheja, M. Buck, T. Wolff, “Enhanced Speaker Activity Detection for Distributed Microphones by Exploitation of Signal Power Ratio Patterns,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2501-2504, Kyoto, Japan, Mar. 2012.

[4] W. Herbordt, T. Trini, W. Kellermann, “Robust Spatial Estimation of the Signal-to-Interference Ratio (SIR) for Non-Stationary Mixtures,” in *Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pp. 247-250, Kyoto, Japan, Sep. 2003.

[5] T. Matheja, M. Buck, T. Wolff, “Robust Adaptive Cancellation of Interfering Speakers for Distributed Microphone Systems in Cars,” in *Proc. Deutsche Jahrestagung für Akustik (DAGA)*, pp. 255-256, Berlin, Germany, Mar. 2010.

[6] J. Bourgeois, W. Minker, *Time-Domain Beamforming and Blind Source Separation*, Springer, Heidelberg, Germany, 2009.

[7] M. Takimoto, T. Nishino, H. Hoshino, K. Takeda, “Estimation of Speaker and Listener Positions in a Car Using Binaural Signals,” *Acoustical Science and Technology*, vol. 29, no. 8, pp. 110-112, Jan. 2008.

[8] E. Cheng, J. Lukasiak, I. S. Burnett, D. Stirling, “Using Spatial Cues for Meeting Speech Segmentation,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pp. 350-353, Amsterdam, The Netherlands, Jul. 2005

[9] G. Lathoud, J. Bourgeois, J. Freudenberger, “Sector-Based Detection for Hands-Free Speech Enhancement in Cars,” *EURASIP Journal on Applied Signal Processing*, Apr. 2006.

[10] I. Cohen, “Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466-475, Sep. 2003.

[11] R. Martin, “An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals,” in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, Berlin, Germany, pp. 1093-1096, Sep. 1993.

[12] G. C. Carter, “Coherence and Time Delay Estimation,” in *Proc. of the IEEE*, vol. 75, no. 2, pp. 236-255, Feb. 1987.

[13] IEEE Standards Publication No. 297, “IEEE Recommended Practice for Speech Quality Measurements,” in *IEEE Transactions on Audio and Electroacoustics*, vol. 17, no. 3, pp. 225-246, Sep. 1969.

[14] T. Matheja, M. Buck, T. Fingscheidt, “A Multi-Channel Quality Assessment Setup Applied to a Distributed Microphone Speech Enhancement System with Spectral Boosting,” in *Proc. ITG-Fachtagung Sprachkommunikation*, Braunschweig, Germany, pp. 119-122, Sep. 2012.