MINIMUM BAYES RISK SIGNAL DETECTION FOR SPEECH ENHANCEMENT BASED ON A NARROWBAND DOA MODEL

Maja Taseska and Emanuël A.P. Habets

International Audio Laboratories Erlangen[*]
Am Wolfsmantel 33, 91058 Erlangen, Germany
{maja.taseska,emanuel.habets}@audiolabs-erlangen.de

ABSTRACT

A desired speech signal in hands-free communication systems is often degraded by background noise and interferers. Data-dependent spatial filters for desired speech extraction depend on the power spectral density (PSD) matrices of the desired and the undesired signals, which are commonly estimated recursively using a signal model-based speech presence probability (SPP). The SPP and the PSD matrix estimates are only accurate, if the statistics of the undesired signals vary more slowly compared to the desired signal. In practical situations with competing talkers, this assumption is violated. To estimate the PSD matrices of highly non-stationary signals, we propose a minimum Bayes risk detector based on a model for the narrowband direction-of-arrival estimates. The performance of the proposed detector and the objective quality of the extracted desired speech are evaluated using simulated and measured data.

Index Terms— Speech enhancement, PSD matrix estimation, signal detection

1. INTRODUCTION

A desired speech signal in practical communication systems, is often degraded by background noise and interfering speech signals. Particularly when the system is used in a hands-free mode, the undesired signals severely degrade the communication quality. Modern systems are often equipped with multiple microphones, which can be used to compute data-dependent spatial filters that reduce the undesired signals, while maintaining low distortion of the desired signal [1]. Many state-of-the-art spatial filters are implemented in the time-frequency (TF) domain [2] and require the PSD matrices of the desired and the undesired signals. Since the work of Ephraim and Malah and their SPP formulation in [3], different algorithms have been proposed that estimate the PSD matrices by recursive averaging controlled by the SPP [4–8]. The SPP-based PSD matrix estimation methods have been extended to multichannel signals, hence exploiting both spatial and spectral properties of the sound sources [9–14].

Voice activity detectors (VADs) and SPP estimation methods that are commonly used in single-channel [3, 4, 6, 15] and multichannel systems [9, 10, 12] are based on a particular signal model. For instance, in the multichannel case, the probability density functions (pdfs) of the observation vectors in the short-time Fourier transform (STFT) domain are modeled by complex Gaussian distributions. To obtain an SPP, the covariances of the Gaussian distributions are required, which correspond to the PSD matrices of the desired and the undesired signals. Due to the inherent problem that the SPP depends on the PSD matrices, and the PSD matrices are recursively updated using the SPP, such systems are only accurate if the statistics of the undesired signals change more slowly over time compared to the desired signal. Clearly, when multiple speech interferers are present, this assumption does not hold and the estimation accuracy rapidly deteriorates.

Numerous researchers have proposed methods to preserve the accuracy of the SPP in non-stationary conditions [7,8,12,14,16–18]. The authors in [14, 17] consider an undesired signal that consists of competing speech sources. To distinguish between desired and undesired talkers, they employ narrowband direction-of-arrival (DOA) estimates for the computation of an *a priori* SPP. However, in challenging multi-talk situations with possibly noisy DOA estimates, such computation of the *a priori* SPP is not sufficient to compensate for the estimation errors in the generalized likelihood ratio used to compute the SPP or a VAD. As a result, the estimated PSD matrices are inaccurate, leading to a low signal quality after spatial filtering.

To improve the performance in the presence of non-stationary interferers, we propose a minimum Bayes risk decision rule to detect TF bins where the desired speech is dominant. In contrast to conventional VADs, the proposed detector is not based on a signal model, but on a narrowband DOA model. The pdf of all DOA estimates is modeled as a mixture distribution, where the mixture components correspond to likelihoods under different hypotheses. A signal model-based paradigm is utilized only in estimating the mixture coefficients. To adapt to changing acoustic conditions, the model parameters are TF-dependent and computed using direct-to-diffuse ratio (DDR) estimates. The main contribution of the paper is presented in Section 3 and consists of (i) defining the mixture model and the mixture components, and (ii) estimating the model parameters. The detection scheme is evaluated in terms of receiver operating characteristics (ROC) and applied to extract a desired speech signal in different acoustic conditions.

2. PROBLEM FORMULATION

A microphone array consisting of M microphones captures the signal of a desired speaker, background noise and an unknown number of interferers. The m-th microphone signal in the STFT domain is given by

$$Y_m(k,n) = X_{m,\mathrm{d}}(k,n) + X_{m,\mathrm{i}}(k,n) + V_m(k,n), \qquad (1)$$

where $X_{m,\mathrm{d}}$, $X_{m,\mathrm{i}}$, and V_m denote the signals of the desired speaker, the interfering speakers and the noise, respectively, and k and n denote the frequency and time indices. The $M \times 1$ vectors $\boldsymbol{y}$, $\boldsymbol{x}_\mathrm{d}$, $\boldsymbol{x}_\mathrm{i}$, and $\boldsymbol{v}$ denote the corresponding signals received at all microphones. The different signals are modeled as zero-mean mu-

[*] A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS, Germany.

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

tually uncorrelated random processes, so that the PSD matrices for each TF bin are related as

$$\boldsymbol{\Phi}_{\boldsymbol{y}}(k,n) = \boldsymbol{\Phi}_{\boldsymbol{x},\mathrm{d}}(k,n) + \boldsymbol{\Phi}_{\boldsymbol{x},\mathrm{i}}(k,n) + \boldsymbol{\Phi}_{\boldsymbol{v}}(k,n), \tag{2}$$

where $\boldsymbol{\Phi}_{\boldsymbol{y}}(k,n) = \mathrm{E}\left[\boldsymbol{y}(k,n)\boldsymbol{y}^{\mathrm{H}}(k,n)\right]$ and $\mathrm{E}\left[\cdot\right]$ is the expectation operator. The remaining PSD matrices are defined accordingly. We assume that the noise signal statistics vary more slowly than the desired and the undesired speech signal statistics.

To estimate the PSD matrices, each TF bin needs to be associated with the desired or the undesired signal, which can be done using an SPP or a voice activity detector (VAD) [11, 13, 14]. Signal model-based approaches for SPP estimation are based on the following hypotheses and the associated likelihood pdfs

$$\mathcal{H}_{\mathrm{d}},\ f(\boldsymbol{y}|\mathcal{H}_{\mathrm{d}};\boldsymbol{\Phi}_{\boldsymbol{x},\mathrm{d}}) \quad \text{the desired signal is dominant,} \tag{3a}$$

$$\mathcal{H}_{\mathrm{u}},\ f(\boldsymbol{y}|\mathcal{H}_{\mathrm{u}};\boldsymbol{\Phi}_{\mathrm{u}}) \quad \text{the undesired signal is dominant,} \tag{3b}$$

where $\mathcal{H}_{\mathrm{u}} = \mathcal{H}_{\mathrm{i}} \cup \mathcal{H}_{v}$ and $\boldsymbol{\Phi}_{\mathrm{u}} = \boldsymbol{\Phi}_{\boldsymbol{x},\mathrm{i}} + \boldsymbol{\Phi}_{\boldsymbol{v}}$. If the PSD matrices are known, the generalized likelihood ratio can be computed and used to obtain an SPP [10]. However, as discussed in the introduction, this procedure is only accurate when the undesired signal is varying at the slower rate than the desired speech signal. The goal in this work is to design a detector which can be used for accurate PSD estimation in scenarios with non-stationary speech interferers.

3. PROPOSED DOA MODEL-BASED DETECTOR

For each TF bin, a DOA estimate $\hat{\theta}(k,n)$ is obtained from the signals $\boldsymbol{y}(k,n)$. The DOA can be estimated e.g. using normalized observation vectors of an arbitrary planar microphone array, as reported in [19]. An optimal detector which minimizes the Bayes risk for a false alarm cost $C_{\mathrm{du}} > 0$ and a misdetection cost $C_{\mathrm{ud}} > 0$, is given by the following decision rule [20]

$$\begin{aligned} &\text{decide } \ \mathcal{H}_{\mathrm{d}} = 1 \ \text{ if } \ \frac{f(\mathcal{H}_{\mathrm{d}}|\hat{\theta})}{f(\mathcal{H}_{\mathrm{u}}|\hat{\theta})} > \frac{C_{\mathrm{du}}}{C_{\mathrm{ud}}}, \\ &\text{decide } \ \mathcal{H}_{\mathrm{d}} = 0 \ \text{ otherwise.} \end{aligned} \tag{4}$$

To compute the posterior probabilities $f(\mathcal{H}_{\mathrm{d}}|\hat{\theta})$ and $f(\mathcal{H}_{\mathrm{u}}|\hat{\theta})$ we start from a mixture model for the narrowband DOA estimates. The processing steps to compute the detector in (4) are illustrated in Fig. 1, and detailed in the remaining part of this section.

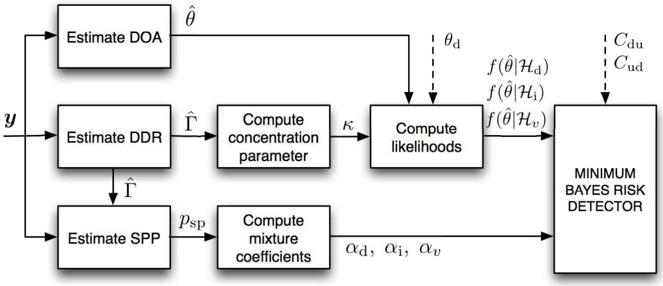

Estimate DOA
$\hat{\theta}$
θ_{d}
C_{du}
C_{ud}
$\boldsymbol{y}$
Estimate DDR
$\hat{\Gamma}$
Compute concentration parameter
κ
Compute likelihoods
$f(\hat{\theta}|\mathcal{H}_{\mathrm{d}})$
$f(\hat{\theta}|\mathcal{H}_{\mathrm{i}})$
$f(\hat{\theta}|\mathcal{H}_{v})$
MINIMUM BAYES RISK DETECTOR
$\hat{\Gamma}$
Estimate SPP
p_{sp}
Compute mixture coefficients
$\alpha_{\mathrm{d}},\ \alpha_{\mathrm{i}},\ \alpha_{v}$

Fig. 1. Block Diagram

3.1. Narrowband DOA model

We propose to model the pdf of the observed DOA estimates by the following mixture distribution

$$f(\hat{\theta}) = \alpha_{\mathrm{d}}\, f(\hat{\theta}|\mathcal{H}_{\mathrm{d}}) + \alpha_{\mathrm{i}}\, f(\hat{\theta}|\mathcal{H}_{\mathrm{i}}) + \alpha_{v}\, f(\hat{\theta}|\mathcal{H}_{v}), \tag{5}$$

where α_{d}, α_{i} and α_{v} denote the mixture coefficients, satisfying $\alpha_{\mathrm{d}} + \alpha_{\mathrm{i}} + \alpha_{v} = 1$. Given the mixture components, the posterior probabilities required in (4) can be computed as

$$f(\mathcal{H}_{\mathrm{d}}|\hat{\theta}) = \frac{\alpha_{\mathrm{d}}\, f(\hat{\theta}|\mathcal{H}_{\mathrm{d}})}{\alpha_{\mathrm{d}}\, f(\hat{\theta}|\mathcal{H}_{\mathrm{d}}) + \alpha_{\mathrm{i}}\, f(\hat{\theta}|\mathcal{H}_{\mathrm{i}}) + \alpha_{v}\, f(\hat{\theta}|\mathcal{H}_{v})}, \tag{6}$$

$$f(\mathcal{H}_{\mathrm{u}}|\hat{\theta}) = \frac{\alpha_{\mathrm{i}}\, f(\hat{\theta}|\mathcal{H}_{\mathrm{i}}) + \alpha_{v}\, f(\hat{\theta}|\mathcal{H}_{v})}{\alpha_{\mathrm{d}}\, f(\hat{\theta}|\mathcal{H}_{\mathrm{d}}) + \alpha_{\mathrm{i}}\, f(\hat{\theta}|\mathcal{H}_{\mathrm{i}}) + \alpha_{v}\, f(\hat{\theta}|\mathcal{H}_{v})}. \tag{7}$$

3.1.1. Modelling the pdfs $f(\hat{\theta}|\mathcal{H}_{\mathrm{d}})$ and $f(\hat{\theta}|\mathcal{H}_{v})$.

We propose to model $f(\hat{\theta}|\mathcal{H}_{\mathrm{d}})$ by a von Mises distribution, which is a distribution on the circle, closely approximating a wrapped Gaussian distribution [21]. A von Mises pdf is characterized by a mean μ and a concentration parameter κ, and is given by

$$f(\hat{\theta}|\mathcal{H}_{\mathrm{d}};\mu,\kappa) = c_{\mathcal{M}}(\kappa)\,\mathrm{e}^{\kappa\cos(\theta-\mu)}. \tag{8}$$

The normalization constant $c_{\mathcal{M}}(\kappa) = [2\pi I_0(\kappa)]^{-1}$ is derived in [22], and I_0 is the modified Bessel function of the first kind. Provided that the DOA estimator is unbiased, the mean μ is equal to the DOA of the desired source, which is assumed to be known. The computation of κ is discussed in Section 3.2. For $\kappa = 0$ the distribution is uniform on the circle. For spatially isotropic noise, the pdf $f(\hat{\theta}|\mathcal{H}_{v})$ can be modeled as a uniform distribution, i.e., $f(\hat{\theta}|\mathcal{H}_{v}) = (2\pi)^{-1}$.

3.1.2. Modelling the pdf $f(\hat{\theta}|\mathcal{H}_{\mathrm{i}})$

If for each TF bin the number and the DOAs of the interferers were known, a multimodal distribution on the circle would be a good model for $f(\hat{\theta}|\mathcal{H}_{\mathrm{i}})$. In practice, the number of interferers and their DOAs are often unknown and may vary quickly. Therefore, we propose to use a distribution that is nearly uniform at directions which are sufficiently far from the desired source (i.e., interferers can be anywhere with approximately equal probability), and has an anti-mode centered at the DOA of the desired source. We construct such distribution as follows: consider a function $g(\theta,\mu,\kappa)$ given by

$$g(\theta,\mu,\kappa) = -\mathrm{e}^{\kappa\cos(\theta-\mu)} + \mathrm{e}^{\kappa}, \tag{9}$$

which attains a minimum value $g(\theta,\mu,\kappa) = 0$ for $\theta = \mu$. As θ deviates from μ, $g(\theta,\mu,\kappa)$ approaches a uniform function. To make it a valid pdf, g needs to be normalized by a constant $c_{\mathcal{A}}$ such that

$$\int c_{\mathcal{A}}\, g(\theta,\mu,\kappa)\,\mathrm{d}\theta = \mathrm{c}_{\mathcal{A}} \int -\mathrm{e}^{\kappa\cos(\theta-\mu)} + \mathrm{e}^{\kappa}\,\mathrm{d}\theta = 1. \tag{10}$$

The integral of the first term is equal to the normalization constant $c_{\mathcal{M}}$ of the von Mises distribution, and the second term is an integral of a constant, evaluated on the circle. Therefore, the normalzation constant and the resulting pdf are given by

$$c_{\mathcal{A}}(\kappa) = [-2\pi(I_0(\kappa) + \mathrm{e}^{\kappa})]^{-1}, \tag{11}$$

$$f(\hat{\theta}|\mathcal{H}_{\mathrm{i}};\mu,\kappa) = c_{\mathcal{A}}(\kappa)\,(-\mathrm{e}^{\kappa\cos(\theta-\mu)} + \mathrm{e}^{\kappa}). \tag{12}$$

3.2. Parameter estimation

To evaluate the detector in (4), the mixture coefficients α_{d}, α_{i} and α_{v}, and the concentration κ need to be estimated. We propose to compute these parameters for each TF using the DDR and a signal model-based SPP. In this manner, the DOA model $f(\theta)$ is time-varying and adapted at each TF bin, while the SPP provides prior information to detect TF-bins where only noise is present. Note that time-varying models have also been used for a signal model-based VAD in [23], where several candidate models are used to characterize the distribution of the spectral coefficients at a given time.

3.2.1. *Estimating mixture coefficients α_{d}, α_{i} and α_v*

To detect noise-only frames, an SPP $p_{\mathrm{sp}}(k,n)$ which indicates the presence of any speech signal (desired or undesired) is required. Hence, $p_{\mathrm{sp}}(k,n)$ should be high when a speech signal is present, and low when only background noise is present. As discussed in Sections 1 and 2, the Gaussian signal model proposed in [10] provides a reliable SPP estimate if the noise statistics vary more slowly than the speech signal statistics. Therefore, we use the two hypothesis model in (3), where the PSD matrices of the Gaussian distributions in (3a) and (3b) are $\boldsymbol{\Phi}_{\boldsymbol{x},\mathrm{d}} + \boldsymbol{\Phi}_{\boldsymbol{x},\mathrm{i}}$ and $\boldsymbol{\Phi}_{\boldsymbol{v}}$, respectively. In addition, to improve the accuracy of the SPP, different methods to compute the *a-priori* SPP exist [8, 12, 16]. In this work we use the DDR-based approach from [12], which relies on the assumption that the speech signals are more coherent across the microphone array than the background noise.

Once an SPP as described above is available, the mixture coefficients can be computed as follows

$$\alpha_v(k,n) = 1 - p_{\mathrm{sp}}(k,n) \tag{13}$$

$$\alpha_{\mathrm{d}}(k,n) = \alpha_{\mathrm{i}}(k,n) = 0.5\ p_{\mathrm{sp}}(k,n). \tag{14}$$

In this manner, in the TF bins where the SPP is low, the noise mixture coefficient α_v attains a large value, whereas when the SPP is high, the speech mixture coefficients α_{d} and α_{i} attain large values. The coefficients α_{d} and α_{i} are set to be equal, as we have no prior information on whether the detected speech corresponds to a desired or an interfering source.

3.2.2. *Estimating concentration parameter κ*

If the DDR is high at a given TF bin, the estimated DOA accurately indicates the true DOA of the direct sound. In such cases, $f(\hat{\theta}|z_{\mathrm{d}})$ and $f(\hat{\theta}|z_{\mathrm{i}})$ should have their mode and anti-mode with high concentration. If the DDR is low, the mode and the anti-mode should spread due to the possible errors in the DOA estimates. Based on this observation, we propose compute the concentration parameter κ using a DDR estimate $\hat{\Gamma}(k,n)$ per TF bin. In particular, we used the following sigmoid-like function

$$\kappa(k,n) = l_{\min} + (l_{\max} - l_{\min})\,\frac{10^{c\rho/10}}{10^{c\rho/10} + \hat{\Gamma}(k,n)^{\rho}}, \tag{15}$$

where $l_{\min}$ and $l_{\max}$ determine the minimum and maximum values of the function, c (in dB) controls the offset along the Γ axis, and ρ controls the steepness of transition region.

The parameters of the mapping function were empirically determined by analyzing the distribution of the DOA estimates for the given DOA estimator under different acoustic conditions. Based on this data, we computed the maximum likelihood (ML) estimates of the concentration parameter κ [21]. The mapping function parameters were chosen such that the mapping function closely follows the behavior of the ML estimates of κ for different DDR values. The investigation resulted in the following parameter values: $l_{\min} = 0$, $l_{\max} = 8$, $c = 15$, and $\rho = -1.2$.

4. PERFORMANCE EVALUATION

4.1. Experimental setup

We evaluated the proposed detector with simulated and measured data, focusing on the following aspects: (i) detection performance in terms of ROC, and (ii) objective quality of an extracted desired signal when the detector is applied for spatial filtering. The scenarios used for the simulations and the measurements are as illustrated in Figure 3. The objective at each array is to detect the TF bins where the signal of the nearest talker is dominant and subsequently enhance that signal. In the scenario in Fig. 3(a), the distance between each source and the nearest array is 0.7 m, whereas in Fig. 3(b), the desired source and the interferer are located at 0.75 m and 2 m from the array, respectively. Each array consists of three omnidirectional microphones forming a uniform circular array with diameter 3 cm.

In the simulated scenarios, a room impulse response (RIR) for each source-microphone pair was computed by the image source model [24] and convolved with clean speech. Diffuse noise was simulated as described in [25]. The microphone signals were obtained by summing the convolved speech signals, the diffuse noise signal and an uncorrelated noise signal. Measurements were done for scenario as shown in Fig. 3(a), using omnidirectional DPA miniature microphones. To generate approximately diffuse sound, ten loudspeakers were placed facing the walls of the room. For the evaluation, clean speech signals were convolved with the measured RIRs for the three sources. To add background noise, different babble speech signals were convolved with the measured RIRs of the ten loudspeakers. The processing was done at a sampling rate of 16 kHz, with an STFT frame length of 64 ms with 50% overlap.

4.2. Evaluation of detection performance

The performance of the proposed detector is first evaluated using ROC curves. As a comparison, we evaluated the signal model-based SPP used in [17], where narrowband DOAs are used in the *a priori* SPP. The SPP as computed in [17] is used to obtain a minimum Bayes risk detector in (4). We denote our proposed DOA model-based detector as $\mathcal{D}_{\mathrm{dm}}$, and the signal model-based detector as $\mathcal{D}_{\mathrm{sm}}$. For the application to spatial filtering, it is important that the detector can operate at very low false alarm rates. The reason for this requirement is that false alarms quickly lead to distortion of the desired signal as the look direction of the spatial filter is erroneously modified. Therefore, we focus on the performance only at the operating region with sufficiently small false alarm rates.

We first investigated the effect of reverberation in a simulated scenario illustrated in 3(a). Values of 0.2 s, 0.35 s, 0.5 s, and 0.65 s were used for the T_{60}, resulting in four ROC curves for each of the two detectors [Fig. 2(a)]. The ROC curves shift upwards as the reverberation time increases. Each array is used to detect the nearest source, while considering the remaining two sources as interferers. The detection performance is averaged over the three sources. Uncorrelated sensor noise and diffuse babble noise were added to the signals, such that the signal-to-noise ratio (SNR) at the reference microphones was 35 dB for the sensor noise and 7 dB for the babble noise. To evaluate the detector for different DOAs, we simulated the scenario in Fig. 3(b), where the interferer initially had the same DOA as the desired source. The simulation was repeated as the interferer was moved clockwise with steps of 10 degrees. The reverberation time was $T_{60} = 0.35$ s and the SNR was as in the previous experiment. The ROC curves for several DOAs are shown in Fig. 2(b), and shift upwards as the difference between the DOA of the desired and the undesired source decreases. Finally, we applied the detector to a measured data in a room with $T_{60} = 0.16$ s, for the scenario in Fig. 3(a). To investigate the effect of background noise, we repeated the experiment with SNRs in the range [-14,12] dB. The ROC curves shift upwards as the SNR decreases, as shown in Fig. 2(c). As shown in Fig. 2, the proposed detector $\mathcal{D}_{\mathrm{dm}}$ provides an advantage over the signal model-based detector $\mathcal{D}_{\mathrm{sm}}$ for all acoustic conditions that were investigated. Although slight performance degradation for increasing reverberation and noise is observed in Fig. 2(a) and Fig. 2(c), the performance is similar for different DOA even when the interferer and the desired source have the same DOA [Fig. 2(b)]. This is due to the fact that the DDR information helps in distinguishing a distant interferer from desired speech source, even if they have the same DOAs.

4.3. Evaluation in a speech enhancement task

To demonstrate the applicability of the proposed detector for spatial filtering, we used the detector to extract a desired signal in several scenarios. Standard recursive PSD estimation was performed [11, 12], where if $\mathcal{H}_{\mathrm{d}} = 1$, the desired signal PSD matrix is updated with an averaging constant of 0.8, and if $\mathcal{H}_{\mathrm{d}} = 0$, the undesired signal PSD matrix is updated. We used the costs $C_{\mathrm{du}} = 2$ and $C_{\mathrm{ud}} = 1$. The spatial filter performance was evaluated for an ideal detector $\mathcal{D}_{\mathrm{id}}$, the signal model-based detector $\mathcal{D}_{\mathrm{sm}}$, and the proposed detector $\mathcal{D}_{\mathrm{dm}}$, in terms of segmental background-and-sensor-noise reduction (segNR), segmental interference reduction (segIR), and speech distortion index ν_{sd}. The performance measures are defined in [13]. The desired signals were extracted using a minimum variance distortionless response (MVDR) filter. The ideal bin-wise detector is set to one if the instantaneous power of the desired speech at a given TF bin is larger than the instantaneous noise power, and if it constitutes at least 90% of the total instantaneous speech power (desired and undesired) at that TF bin. Otherwise, the ideal detector is set to zero, and the PSD matrix of the desired signal is not updated.

The results for the scenario in Fig. 3(a) using measured data are given in Table 1, for desired signal-to-background noise ratios of 7.5 dB and 0.5 dB. When using the proposed detector $\mathcal{D}_{\mathrm{dm}}$, all talkers are extracted with very low distortion and significant segIR and segNR in several scenarios, approaching the performance of the ideal detector $\mathcal{D}_{\mathrm{id}}$. Although the speech distortion when using the signal model-based detector $\mathcal{D}_{\mathrm{sm}}$ remains low as for $\mathcal{D}_{\mathrm{dm}}$, using $\mathcal{D}_{\mathrm{sm}}$ results in much lower and insufficient interference reduction. An MVDR filter to extract the desired signal was also applied to the scenario in Fig. 3(b), when the interferer and the desired source have the same DOA. The results shown in Table 2, demonstrate that also in such scenario, the proposed detector outperforms the signal model-based, as indicated by the ROC curves as well. Nevertheless, due to the identical DOA of the sources, the interference reduction deteriorates for all detectors in general. Note that the performance loss is also related to the inherent subspace overlap between the PSD matrices when the sources have same DOAs. A recent study on this was published by the present authors in [26].

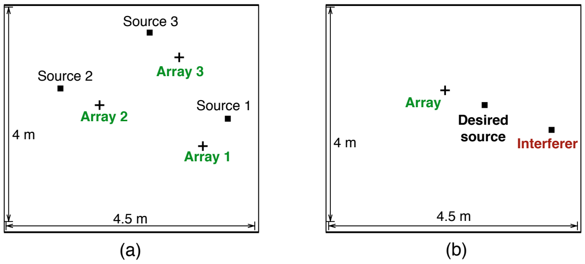


Fig. 3. Scenarios used in the evaluation.

	SNR 7.5 dB			SNR 0.5 dB		
	$\mathcal{D}_{\mathrm{id}}$	$\mathcal{D}_{\mathrm{sm}}$	$\mathcal{D}_{\mathrm{dm}}$	$\mathcal{D}_{\mathrm{id}}$	$\mathcal{D}_{\mathrm{sm}}$	$\mathcal{D}_{\mathrm{dm}}$
segNR [dB]	6.4	3.8	6.5	7.8	2.6	8.1
segIR [dB]	12.8	1.7	11.3	11.5	1.4	10.5
ν_{sd}	0.01	0.04	0.03	0.01	0.04	0.04
segNR [dB]	6.8	2.3	3.6	9.5	1.4	3.8
segIR [dB]	13.0	1.8	9.7	11.6	1.8	8.1
ν_{sd}	0.01	0.03	0.01	0.01	0.03	0.02
segNR [dB]	2.2	3.8	5.9	4.0	2.3	7.5
segIR [dB]	11.3	2.3	9.5	10.0	1.8	8.6
ν_{sd}	0.01	0.03	0.03	0.01	0.02	0.04

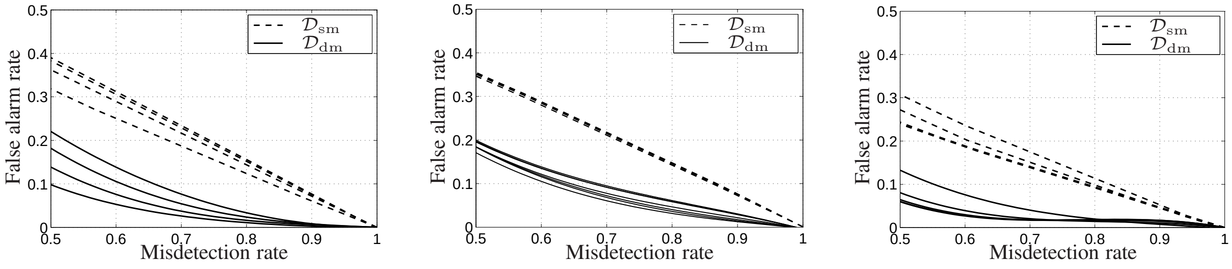
Table 1. Results for `Talker1` (top), `Talker2` (middle), and `Talker3` (bottom). The signal-to-interference ratio at the reference microphone of each talker is 3.5 dB, 2.9 dB, and 3.6 dB, respectively.

$\mathcal{D}_{\mathrm{id}}$	$\mathcal{D}_{\mathrm{sm}}$	$\mathcal{D}_{\mathrm{dm}}$	$\mathcal{D}_{\mathrm{id}}$	$\mathcal{D}_{\mathrm{sm}}$	$\mathcal{D}_{\mathrm{dm}}$	$\mathcal{D}_{\mathrm{id}}$	$\mathcal{D}_{\mathrm{sm}}$	$\mathcal{D}_{\mathrm{dm}}$
3.9	3.2	5.6	6.9	1.5	3.1	0.05	0.07	0.07

Table 2. Simulated scenario where the desired and interfering talker have the same DOA: segNR (left), segIR (middle) and ν_{sd} (right).

5. CONCLUSIONS

We proposed a minimum Bayes risk detector based on a narrowband DOA model which can be applied for the estimation of PSD matrices in spatial filtering. In contrast to classical signal model-based VADs and SPP estimation frameworks, the proposed detector can operate with very low false alarm rates in non-stationary scenarios with multiple speech interferers and high levels of background noise. The robustness of the detector was demonstrated by ROC curves, and by objective evaluation of extracted desired signals in different scenarios and various acoustic conditions. Audio samples with extracted source signals can be found at http://www.audiolabs-erlangen.de/resources/2015-ICASSP-DOAdet.



(a) Effect of reverberation

(b) Effect of interferer's DOA, $T_{60} = 0.35$ s

(c) Effect of background noise, $T_{60} = 0.15$ s

Fig. 2. ROC curves for the two detectors with simulated [Figures (a) and (b)] and measured data [Fig. (c)]. The different lines in each plot correspond to different T_{60} values in Fig.(a), different DOAs of the interferer in Fig.(b), and different background noise levels in Fig.(c).

6. REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, Springer-Verlag, Berlin, Germany, 2008.

[2] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag, 2011.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[4] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[5] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001.

[6] R.C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2010, pp. 4266–4269.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[8] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.

[9] I. Potamitis, "Estimation of speech presence probability in the field of microphone array," *IEEE Signal Processing Letters*, vol. 11, no. 12, pp. 956–959, Dec. 2004.

[10] M. Souden, J. Chen, J. Benesty, and S. Affes, "Gaussian model-based multichannel speech presence probability," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 5, pp. 1072–1077, Jul. 2010.

[11] M. Souden, J. Chen, J. Benesty, and S. Affes, "An integrated solution for online multichannel noise tracking and reduction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2159–2169, Sep. 2011.

[12] M. Taseska and E. A. P. Habets, "MMSE-based blind source extraction in diffuse noise fields using a complex coherence-based a priori SAP estimator," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sep. 2012.

[13] M. Taseska and E. A. P. Habets, "Informed spatial filtering with distributed arrays," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 7, pp. 1195–1207, Jul. 2014.

[14] D. P. Jarrett, M. Taseska, E. A. P. Habets, and P. Naylor, "Noise reduction in the spherical harmonic domain using a tradeoff beamformer and narrowband DOA estimates," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 5, pp. 967–977, May 2014.

[15] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detector," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[16] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[17] D. P. Jarrett, E. A. P. Habets, and P. A. Naylor, "Spherical harmonic domain noise reduction using an MVDR beamformer and DOA-based second-order statistics estimation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013.

[18] T. Nakatani, S. Araki, T. Yoshioka, M. Delcroix, and M. Fujumoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 12, pp. 2516–2531, Dec. 2013.

[19] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. Intl. Workshop Acoust. Signal Enhancement (IWAENC)*, Sep. 2005.

[20] S. Kay, *Fundamentals of statistical signal processing, Volume II: Detection theory*, Prentice Hall, 1998.

[21] K. V. Mardia and P. E. Jupp, *Directional Statistics*, Wiley-Blackwell, New York, NY, USA, 1999.

[22] M. Abramowitz and I. A. Stegun, Eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York, USA, 1972.

[23] J. Chang, N. S. Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 1965–1976, Jun. 2006.

[24] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2006.

[25] E. A. P. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *J. Acoust. Soc. Am.*, vol. 122, no. 6, pp. 3464–3470, Dec. 2007.

[26] M. Taseska and E. A. P. Habets, "A subspace-based perspective on spatial filtering performance with distributed and co-located microphone arrays," in *ITG Fachtagung Sprachkommunikation*, Erlangen, Germany, Sep. 2014.