ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3315-6701-9/26/$31.00 ©2026 IEEE | DOI: 10.1109/ICASSP55912.2026.11464958

DNN-BASED ONLINE SOURCE COUNTING BASED ON SPATIAL GENERALIZED MAGNITUDE SQUARED COHERENCE

Henri Gode, Simon Doclo

Department of Medical Physics and Acoustics and Cluster of Excellence Hearing4all,
Carl von Ossietzky Universität Oldenburg, Germany, e-mail: henri.gode@uni-oldenburg.de

ABSTRACT

The number of active sound sources is a key parameter in many acoustic signal processing tasks, such as source localization, source separation, and multi-microphone speech enhancement. This paper proposes a novel method for online source counting by detecting changes in the number of active sources based on spatial coherence. The proposed method exploits the fact that a single coherent source in spatially white background noise yields high spatial coherence, whereas only noise results in low spatial coherence. By applying a spatial whitening operation, the source counting problem is reformulated as a change detection task, aiming to identify the time frames when the number of active sources changes. The method leverages the generalized magnitude-squared coherence as a measure to quantify spatial coherence, providing features for a compact neural network trained to detect source count changes framewise. Simulation results with binaural hearing aids in reverberant acoustic scenes with up to 4 speakers and background noise demonstrate the effectiveness of the proposed method for online source counting.

Index Terms— source counting, spatial coherence, GRU, TCN, binaural hearing aids

1. INTRODUCTION

Enhancing speech quality and intelligibility in complex acoustic environments with multiple competing sound sources is a primary goal of assistive listening devices, such as binaural hearing aids. To achieve this, advanced multi-microphone algorithms are employed for tasks such as source localization, source separation, and speech enhancement [1–4]. A fundamental prerequisite for many of these algorithms is knowledge about the number of active sound sources. In dynamic scenarios where sources may activate and deactivate over time, it is crucial to perform online source counting, i.e., relying only on past information.

Several methods have been proposed for counting active sound sources. Single-microphone methods typically rely on deep neural networks (DNNs) to estimate the source count directly from the mixture [5–7]. In contrast, multi-microphone methods are able to leverage spatial information. Some methods perform clustering on spatial features such as direction-of-arrival estimates, often requiring a-priori knowledge of the microphone array geometry and hence limiting their applicability [8–11]. Geometry-agnostic methods have been proposed based on the eigenvalue distribution of spatial correlation or coherence matrices [12, 13]. While effective, these aforementioned methods and also recent DNN models that jointly perform source counting and separation [14, 15] typically estimate the source count over long signal segments ($1-20$s), which is too slow for low-latency speech communication applications. Other methods tackling online speaker counting [16–18] operate with latencies of at least 200ms or rely on Ambisonics features [16], which is not directly applicable to binaural hearing aids. As a more suitable baseline for online processing, we will consider the framewise source counting method proposed in [19], using the generalized magnitude-squared coherence (GMSC) [20] on spatially whitened signals. However, this method is limited to detecting only source activations and relies on manually-tuned thresholds.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence strategy – Project ID 390895286 – EXC 2177/1.

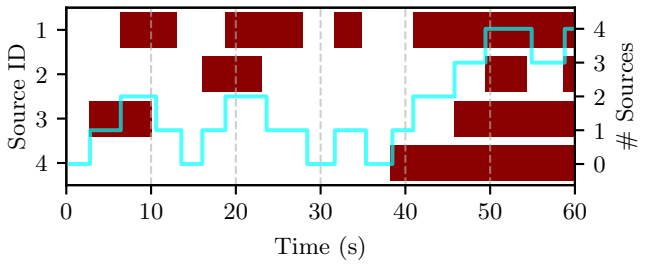

Fig. 1. Exemplary source activity timeline with 4 sources, showing activity intervals and the total number of active sources.

Inspired by the GMSC-based method in [19], in this paper we propose a more robust DNN-based source counting method. First, we introduce a novel set of features designed to detect source deactivations, based on a time-reversed whitening approach. Second, to overcome the limitations of manually tuned thresholds, we train compact neural networks to provide a framewise estimate of the number of active sources based on the spatial coherence features in the whitened domain. We consider two causal architectures, a temporal convolutional network (TCN) and a gated recurrent unit (GRU)-based recurrent neural network, The proposed systems are evaluated for a binaural hearing aid setup on recorded noisy and reverberant acoustic scenarios with a time-varying number of speakers (see exemplary scenario in Figure 1). Results show that the proposed DNN-based source count estimators significantly outperform the conventional threshold-based method in terms of accuracy and mean absolute error, and especially benefit from incorporating the proposed deactivation features. The GRU-based source count estimator consistently outperforms the TCN-based source count estimator, achieving a frame-wise accuracy of $91.9\,\%$.

2. SIGNAL MODEL

We consider an acoustic scenario with a time-varying number of spatially-stationary sound sources in a noisy and reverberant environment, recorded by an array of M microphones. In the short-time Fourier transform (STFT) domain, the m-th microphone signal is denoted by $y_{m,t,f}$, with time frame t and frequency f. The microphone signals are stacked in the vector

$$\mathbf{y}_{t,f} = \begin{bmatrix} y_{1,t,f} & \cdots & y_{M,t,f} \end{bmatrix}^{\mathrm{T}} \in \mathbb{C}^{M}, \tag{1}$$

where $\{\cdot\}^{\mathrm{T}}$ denotes the transpose. For notational simplicity, the frequency index f will be omitted, except when it is explicitly required. The

microphone signal vector $\mathbf{y}_t$ is modeled as the sum of all active source contributions and additive noise, i.e.,

$$\mathbf{y}_t = \sum_{k=1}^{K_{\max}} \mathcal{I}_{k,t}\mathbf{x}_{k,t} + \mathbf{n}_t, \tag{2}$$

where $K_{\max}$ denotes the total number of potential sources, $\mathbf{x}_{k,t} \in \mathbb{C}^M$ denotes the k-th source vector, and $\mathbf{n}_t \in \mathbb{C}^M$ denotes the noise vector. The frequency-independent binary indicator $\mathcal{I}_{k,t} \in \{0,1\}$ denotes the (broadband) activity of the k-th source at time frame t, modeling the time-varying number of active sources. The method considered in this paper aims at estimating the number of active sources

$$K_t = \sum_{k=1}^{K_{\max}} \mathcal{I}_{k,t}, \tag{3}$$

per frame using only $\mathbf{y}_t$ from current and past frames.

Assuming that the length of the STFT analysis window is sufficiently large compared to the length of the room impulse response, each source component $\mathbf{x}_{k,t}$ can be written as [21]

$$\mathbf{x}_{k,t} = \mathbf{a}_k s_{k,t}, \tag{4}$$

where $s_{k,t}$ denotes the STFT coefficient of the k-th source signal and $\mathbf{a}_k \in \mathbb{C}^M$ denotes the time-invariant acoustic transfer function (ATF) vector from the k-th source to the microphone array. The ATF vectors of all sources are assumed to be linearly independent. Assuming statistical independence between all sources and the noise, the covariance matrix of the microphone signals $\mathbf{R}_{y,t} = \mathbb{E}\{\mathbf{y}_t\mathbf{y}_t^{\mathrm{H}}\}$ can be expressed as

$$\mathbf{R}_{y,t} = \sum_{k=1}^{K_{\max}} \mathcal{I}_{k,t}\phi_{k,t}\mathbf{a}_k\mathbf{a}_k^{\mathrm{H}} + \mathbf{R}_n \quad \in \mathbb{C}^{M\times M}, \tag{5}$$

where $\{\cdot\}^{\mathrm{H}}$ denotes the conjugate transpose, $\phi_{k,t} = \mathbb{E}\{|s_{k,t}|^2\}$ denotes the power spectral density (PSD) of the k-th source, and $\mathbf{R}_n = \mathbb{E}\{\mathbf{n}_t\mathbf{n}_t^{\mathrm{H}}\}$ denotes the noise covariance matrix, assumed to be time-invariant.

Considering the activation or deactivation of a single source k' between two consecutive time frames $t-1$ and t, the number of active sources in (3) changes by one, i.e., $K_t = K_{t-1} \pm 1$. Assuming the PSDs of the other active sources do not change significantly between consecutive time frames, the matrix $\mathbf{R}_{y,t}$ in (5) can be expressed as

$$\boxed{\mathbf{R}_{y,t} \approx \mathbf{R}_{y,t-1} \pm \phi_{k'}\mathbf{a}_{k'}\mathbf{a}_{k'}^{\mathrm{H}}} \tag{6}$$

where $\phi_{k'}$ denotes the PSD of the k'-th source (in frame t for activation and frame $t-1$ for deactivation). It can be observed in (6) that a source activation corresponds to the addition of a rank-1 matrix, whereas a source deactivation corresponds to the subtraction of a rank-1 matrix. The core idea of the spatial-coherence based source counting methods, in this paper is to utilize these rank-1 updates of the noisy covariance matrix to track the number of active sources K_t.

3. CONVENTIONAL SOURCE ACTIVATION DETECTION

The conventional method for source activation detection [19] directly exploits the rank-1 update property in (6). The core idea is to reformulate the problem of estimating the number of sources per time frame into the problem of detecting the activation of a single coherent source in spatially white noise. This is achieved by a whitening process (taking into account already active sources and noise), so that the spatial coherence in this whitened domain, is expected to increase significantly only when a new source activates.

3.1. Whitened GMSC Features for Source Activation Detection

To detect the addition of the rank-1 matrix corresponding to a new source activation, the conventional method [19] whitens the covariance matrix $\mathbf{R}_{y,t}$ at time frame t with a reference covariance matrix denoted as $\mathbf{R}_{v,t}$. Using a matrix square-root decomposition of $\mathbf{R}_{v,t}$, i.e., $\mathbf{R}_{v,t} = \mathbf{R}_{v,t}^{\mathrm{H}/2}\mathbf{R}_{v,t}^{1/2}$ (e.g., Cholesky decomposition), the whitened covariance matrix $\mathbf{R}_{w,t}^{\mathrm{a}}$ is computed as

$$\boxed{\mathbf{R}_{w,t}^{\mathrm{a}} = \mathbf{R}_{v,t}^{-\mathrm{H}/2}\mathbf{R}_{y,t}\mathbf{R}_{v,t}^{-1/2}} \tag{7}$$

The reference covariance matrix is used as an estimate of the covariance matrix before the activation, containing the $K_t - 1$ already active sources and the noise. In [19] it was proposed to use the covariance matrix $\mathbf{R}_{y,t}$ from t_v frames in the past as the reference covariance matrix, i.e.,

$$\mathbf{R}_{v,t} = \mathbf{R}_{y,t-t_v}, \tag{8}$$

where t_v corresponds to the assumed minimum time between two source activations. As long as no new source activates, $\mathbf{R}_{y,t} \approx \mathbf{R}_{v,t}$, such that the whitened covariance matrix $\mathbf{R}_{w,t}^{\mathrm{a}}$ is approximately equal to the identity matrix, resulting in low spatial coherence. When a new source becomes active, $\mathbf{R}_{w,t}^{\mathrm{a}}$ contains an additional rank-1 component, leading to a sharp increase in spatial coherence. To quantify this, it was proposed in [19] to use the GMSC [20], which extends the concept of spatial coherence between two microphones to multiple microphones. The GMSC is defined as

$$\gamma_t^{\mathrm{a}} = \frac{\lambda_{\max}\{\mathbf{\Gamma}_t^{\mathrm{a}}\} - 1}{M-1}, \tag{9}$$

yielding a value between 0 and 1, where $\lambda_{\max}\{\cdot\}$ denotes the principal eigenvalue and $\mathbf{\Gamma}_t^{\mathrm{a}}$ denotes the coherence matrix of the whitened signals, i.e.,

$$\mathbf{\Gamma}_t^{\mathrm{a}} = \mathbf{D}_{w,t}^{-1/2}\mathbf{R}_{w,t}^{\mathrm{a}}\mathbf{D}_{w,t}^{-1/2}. \tag{10}$$

where $\mathbf{D}_{w,t}$ denotes a diagonal matrix containing the diagonal entries of $\mathbf{R}_{w,t}^{\mathrm{a}}$. Although presented without a frequency index for notational simplicity, the whitening in (7) and the GMSC calculation in (9) and (10) are performed for each frequency f. The resulting frequency-dependent GMSC values $\gamma_{t,f}^{\mathrm{a}}$ are then stacked into the GMSC feature vector

$$\boldsymbol{\gamma}_t^{\mathrm{a}} = \begin{bmatrix}\gamma_{t,1}^{\mathrm{a}} & \cdots & \gamma_{t,F}^{\mathrm{a}}\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^F, \tag{11}$$

where F denotes the number of frequency bins. This feature vector, whose dimension does not depend on the number of microphones M, serves as the input for the subsequent detection stages. For an exemplary acoustic scenario, the top subplot of Figure 2 illustrates the GMSC-based activation features. It can be observed that high GMSC values occur at time frames when new sources activate.

In practice, the covariance matrix $\mathbf{R}_{y,t}$ is estimated from the observed microphone signals by recursive smoothing, i.e.,

$$\widehat{\mathbf{R}}_{y,t} = \alpha\widehat{\mathbf{R}}_{y,t-1} + (1-\alpha)\mathbf{y}_t\mathbf{y}_t^{\mathrm{H}}, \tag{12}$$

where the forgetting factor $\alpha = e^{-t_{\mathrm{fs}}/t_\alpha}$ depends on the STFT frame shift t_{fs} and a smoothing time constant t_α.

3.2. Threshold-Based Detection

For the conventional threshold-based source activation detection [19], the feature vector $\boldsymbol{\gamma}_t^{\mathrm{a}}$ is first condensed into a single broadband value $\tilde{\gamma}_t^{\mathrm{a}}$ by computing a weighted average, i.e.,

$$\tilde{\gamma}_t^{\mathrm{a}} = \mathbf{w}_t^{\mathrm{T}}\boldsymbol{\gamma}_t^{\mathrm{a}}, \tag{13}$$

where the weight vector $\mathbf{w}_t = \begin{bmatrix}w_{t,1} & \cdots & w_{t,F}\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^F$ contains the normalized power of the whitened signal in each frequency bin, i.e.,

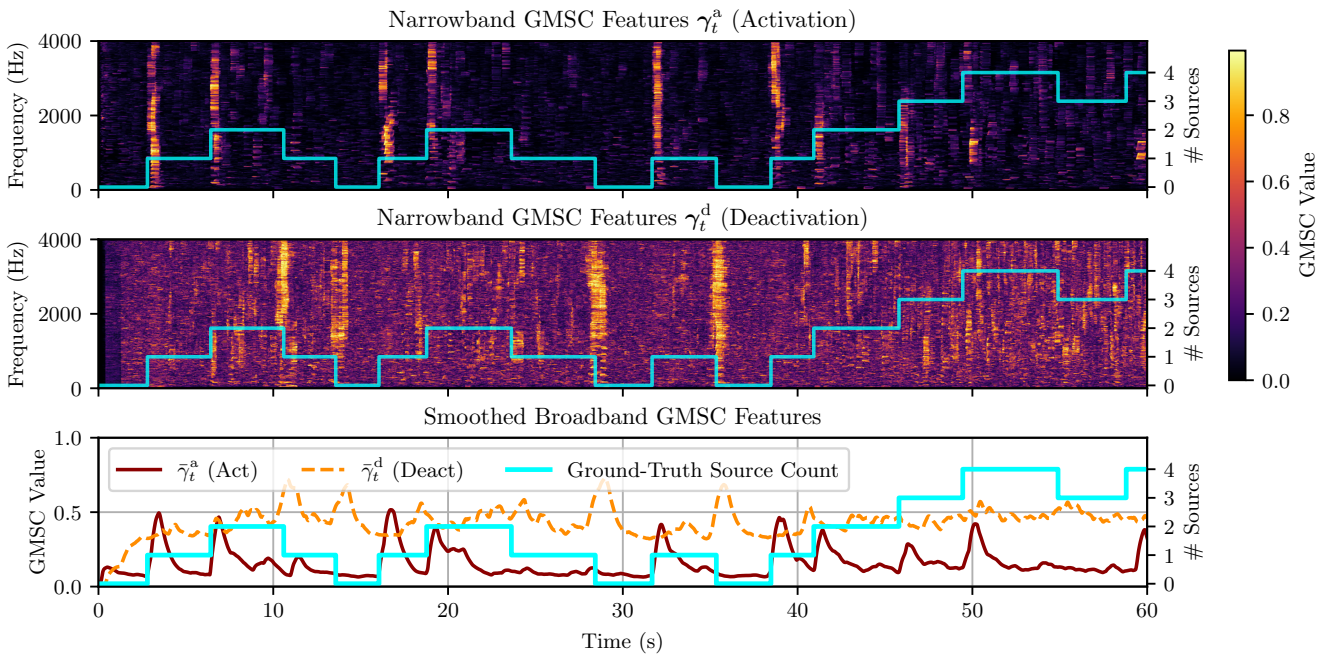

Fig. 2. Illustration of the GMSC-based features (SNR=9.5dB). Top: GMSC features for activation. Middle: GMSC features for deactivation. Bottom: Corresponding broadband features after recursive smoothing.

$w_{t,f} = \mathrm{tr}\{\mathbf{R}^{\mathrm{a}}_{w,t,f}\}/\sum_{f=1}^{F}\mathrm{tr}\{\mathbf{R}^{\mathrm{a}}_{w,t,f}\}$, where $\mathrm{tr}\{\cdot\}$ denotes the trace. The broadband GMSC $\tilde{\gamma}_t^{\mathrm{a}}$ serves as a direct indicator of a new source activation. To improve robustness, $\tilde{\gamma}_t^{\mathrm{a}}$ is recursively smoothed, i.e.,

$$\bar{\gamma}_t^{\mathrm{a}} = \beta\bar{\gamma}_{t-1} + (1-\beta)\tilde{\gamma}_t^{\mathrm{a}}, \tag{14}$$

where the forgetting factor $\beta = e^{-t_{\mathrm{fs}}/t_\gamma}$ depends on the STFT frame shift t_{fs} and a smoothing time constant t_γ. An example of the smoothed broadband GMSC $\bar{\gamma}_t^{\mathrm{a}}$ is shown in the bottom subplot of Figure 2 (red curve). A source activation is then detected if the smoothed feature $\bar{\gamma}_t^{\mathrm{a}}$ exceeds a predefined threshold γ_τ^{a}. Upon detection, the source count K_t is incremented by one. It should be noted that determining a single threshold that performs robustly across diverse acoustic scenarios (e.g., varying SNRs or source positions) is non-trivial.

4. SOURCE DEACTIVATION DETECTION FEATURES

A disadvantage of the conventional source activity estimation method [19] is that it can only detect source activations, but no source deactivations, which is obviously unrealistic. To extend the conventional method to also detect source deactivations, we propose a deactivation feature based on interpreting a deactivation as a time-reversed activation. Following this intuition, a similar whitening-based approach as in Section 3.1 can be used by swapping the roles of the current and reference covariance matrices. Instead of whitening the covariance matrix $\mathbf{R}_{y,t}$ at frame t with the covariance matrix $\mathbf{R}_{y,t-t_v}$ from the past, we whiten the covariance matrix $\mathbf{R}_{y,t-t_v}$ from the past with the current covariance matrix $\mathbf{R}_{y,t}$, i.e.,

$$\boxed{\mathbf{R}^{\mathrm{d}}_{w,t} = \mathbf{R}_{y,t}^{-\mathrm{H}/2}\mathbf{R}_{y,t-t_v}\mathbf{R}_{y,t}^{-1/2}} \tag{15}$$

As long as no source deactivates, $\mathbf{R}_{y,t} \approx \mathbf{R}_{y,t-t_v}$, such that the whitened covariance matrix $\mathbf{R}^{\mathrm{d}}_{w,t}$ is approximately equal to the identity matrix, resulting in a low GMSC value. If a source deactivates at time $t-t_v$, then $\mathbf{R}^{\mathrm{d}}_{w,t}$ becomes approximately equal to an identity matrix plus a rank-1 component, leading to a high GMSC value. This effectively turns deactivation detection into an activation detection problem, with an inherent processing delay of t_v frames.

Due to this inherent processing delay, we decided not to use recursive smoothing for estimating the covariance matrix as in (12), but to use a more instantaneous moving-average estimate, i.e.,

$$\widehat{\mathbf{R}}_{y,t} = \frac{1}{L}\sum_{l=0}^{L-1}\mathbf{y}_{t-l}\mathbf{y}_{t-l}^{\mathrm{H}}. \tag{16}$$

where the sliding window length L is chosen to be smaller than t_v.

From the whitened matrix $\mathbf{R}^{\mathrm{d}}_{w,t}$, the feature vector $\boldsymbol{\gamma}_t^{\mathrm{d}}$ for deactivation is computed similarly to (9) - (11). For the deactivation detection, a similar threshold-based method as described in Section 3.2 can be applied to the averaged and smoothed broadband value of this feature vector $\bar{\gamma}_t^{\mathrm{d}}$ (computed similarly to (13) and (14)), where a detection now leads to decrementing the source count K_t by one. An example of the proposed GMSC deactivation feature vector $\boldsymbol{\gamma}_t^{\mathrm{d}}$ and the smoothed broadband GMSC $\bar{\gamma}_t^{\mathrm{d}}$ is illustrated in the middle and bottom subplots (orange curve) of Figure 2.

5. PROPOSED DNN-BASED SOURCE COUNT ESTIMATION

To overcome the limitations of the threshold-based method, which requires careful manual tuning and may not generalize well, we propose to train a DNN to estimate the number of active sources K_t at each time frame from spatial coherence features. As input, we use the GMSC-based feature vectors $\boldsymbol{\gamma}_t^{\mathrm{a}}$ and $\boldsymbol{\gamma}_t^{\mathrm{d}}$ for activation and deactivation, which are concatenated to form a combined feature vector $\boldsymbol{\zeta}_t \in \mathbb{R}^{2F}$. We investigate two causal network architectures: a TCN [22], which uses dilated convolutions to create a large receptive field, and a GRU-based recurrent neural network (RNN) [23], which uses its internal state to capture temporal dependencies. For both architectures, a final softmax layer outputs probabilities $p_{t,k}$ for each possible source count $k \in \{0,\ldots,K_{\max}\}$, and the final estimate is taken as the most likely class, i.e., $\widehat{K}_t = \underset{k}{\mathrm{argmax}}\, p_{t,k}$.

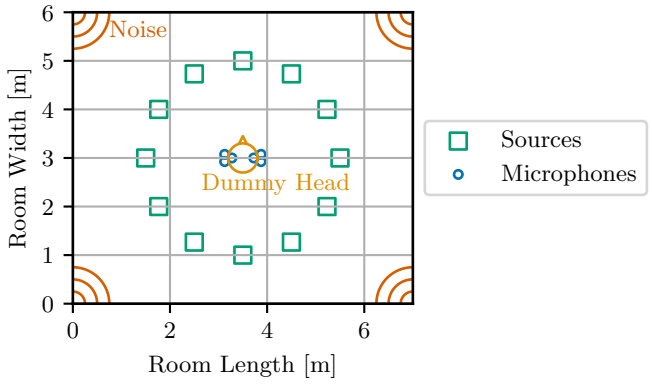

Fig. 3. Acoustic setup of the BRUDEX database [24].

6. EVALUATION

In this section, the performance of the conventional threshold-based method and the proposed DNN-based source count estimators is compared for several acoustic scenarios with up to 4 speakers and background noise. The experimental setup, including dataset generation, algorithmic implementation, and training procedure, is detailed below.

6.1. Dataset

Acoustic scenarios were generated by convolving clean speech from Librispeech [25] with binaural room impulse responses (RIRs) from the BRUDEX database [24] for the low-reverberation condition ($T_{60} \approx 310$ms) at a sampling frequency of 8kHz. As depicted in Figure 3, the RIRs cover 12 source positions on a circle of radius 1m around a dummy head with an angular spacing of 30°. The dummy head was equipped with a behind-the-ear (BTE) hearing aid on each side, with each BTE having a front and a rear microphone spaced 1.5cm apart. Additionally, an in-ear microphone was present on each side of the head. This setup provides various microphone array configurations using 2 to 6 of these microphones. Quasi-diffuse babble, cafeteria or white noise from BRUDEX was added at a random SNR between 5dB and 15dB. To create dynamic scenes, the number of simultaneously active sources varies up to $K_{\max} = 4$, with activation/deactivation events occurring at random intervals between 2s and 5s. Based on this, two distinct datasets were generated: Dataset A (20s scenarios) contains only source activations, while Dataset B (60s scenarios) contains both activations and deactivations. Each dataset is split into 6000 training, 600 validation, and 600 test scenarios, where for each scenario the source positions, clean speech signals, microphone array configuration, SNR, noise type, and activity pattern were chosen randomly.

6.2. Algorithmic Implementation and Training

The microphone signals are processed using an STFT framework with a frame length of 800 samples (corresponding to 100ms), resulting in $F = 401$, a frame shift of 25%, i.e., $t_{\text{fs}} = 25$ms, and a square-root Hann analysis window. The GMSC-based activation features are computed with a recursive smoothing of $t_\alpha = 500$ms in (12), while the deactivation features are computed with a sliding window of $L = 14$ (corresponding to 350ms) in (16). For both features, the delay for the reference covariance matrix in (8) and (15) is equal to $t_v = 20$ frames (corresponding to 500ms).

We compare three source count estimators using either only activation features for Dataset A and Dataset B, or using concatenated activation and deactivation features for Dataset B:

- The conventional **threshold-based** method using broadband feature smoothing of $t_\gamma = 500$ms in (14) and fixed thresholds of $\gamma_\tau^{\text{a}} = 0.24$ (activation) and $\gamma_\tau^{\text{d}} = 0.62$ (deactivation), which were determined using a grid search on the test set of Dataset A and Dataset B, respectively.

Table 1. Accuracy and MAE results on Dataset A and Dataset B for different feature sets ($\boldsymbol{\gamma}_t^{\text{a}}$ or $\boldsymbol{\zeta}_t$).

Features	Detector Method	Accuracy [%]	MAE
Dataset A (Activations only)			
Features: $\boldsymbol{\gamma}_t^a$	Threshold-based	83.1	0.18
	TCN-based	94.2	0.06
	RNN-based	**95.6**	**0.04**
Dataset B (Activations & Deactivations)			
Features: $\boldsymbol{\gamma}_t^a$	TCN-based	83.8	0.17
	RNN-based	**89.0**	**0.12**
Features: $\boldsymbol{\zeta}_t$	Threshold-based	33.6	1.21
	TCN-based	86.4	0.14
	RNN-based	**91.9**	**0.09**

- An **TCN-based** estimator using 3 stacks of 3 layers with a kernel size of 3, a bottleneck dimension of 128, and a hidden dimension of 256, resulting in a receptive field of 43 frames (=1.075s).
- An **RNN-based** estimator using a 3-layer GRU with a hidden size of half the input dimension.

The DNN models were trained for 100 epochs using the AdamW optimizer with a learning rate of 10^{-4} and a batch size of 30 to minimize the cross-entropy loss between the estimated and the ground-truth source count. This ground-truth, K_t was computed using (3), where the ground-truth activity indicator $\mathcal{I}_{k,t}$ was determined by applying a power-based voice activity detector to the source components $\mathbf{x}_{k,t}\, \forall k \in \{1,\ldots,K_{\max}\}$.

6.3. Results

Performance is evaluated on the test sets using two framewise metrics: the classification accuracy (percentage of frames where $\widehat{K}_t = K_t$) and the mean absolute error (MAE), i.e., $\text{MAE} = {}^1\!/_T \sum_{t=1}^{T} |\widehat{K}_t - K_t|$, where T is the total number of frames in a test set. The results are presented in Table 1. On the activation-only Dataset A, both DNN-based methods significantly outperform the threshold-based method, with the RNN achieving the best accuracy of 95.6 %. On the more challenging Dataset B, which includes both activations and deactivations, the performance of the DNN models using only activation features $\boldsymbol{\gamma}_t^{\text{a}}$ is, as expected, lower than on Dataset A. The benefit of the proposed deactivation features $\boldsymbol{\gamma}_t^{\text{d}}$ is clearly demonstrated: while the threshold-based method fails completely on Dataset B (accuracy of 33.6 %), both DNN-based methods show a notable performance improvement compared to only using the activation features $\boldsymbol{\gamma}_t^{\text{a}}$. The RNN-based estimator consistently performs best, reaching an accuracy of 91.9% when using both activation and deactivation features. Finally, the proposed method is real-time feasible, with an average real-time factor of 0.007 (dominated by feature extraction) measured on an NVIDIA RTX A5000 GPU.

7. CONCLUSION

This paper extends a spatial coherence-based method for online source activation detection. We propose GMSC-based features for detecting source deactivations using a time-reversed whitening approach, complementing existing activation features. Furthermore, we propose to replace the conventional threshold-based source count estimator with more robust DNN-based classifiers (TCN and GRU) that directly estimate the source count from the combined feature set. The evaluation on real-world acoustic scenarios shows that the DNN-based methods significantly outperform the threshold-based method. Crucially, incorporating the proposed deactivation features substantially improves performance in scenarios involving both activations and deactivations. The RNN-based source count estimator proves most effective, achieving a frame-wise accuracy of 91.9 %.

8. REFERENCES

[1] S. Doclo, W. Kellermann, S. Makino, and S. E. Nordholm, "Multichannel Signal Enhancement Algorithms for Assisted Listening Devices: Exploiting spatial diversity using multiple microphones," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 18–30, 2015.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[3] P. Pertilä, A. Brutti, P. Svaizer, and M. Omologo, "Multichannel source activity detection, localization, and tracking," *Audio Source Separation and Speech Enhancement*, pp. 47–64, 2018.

[4] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *The Journal of the Acoustical Society of America*, vol. 152, no. 1, pp. 107–151, 2022.

[5] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, 2018, pp. 5064–5068.

[6] F.-R. Stöter, S. Chakrabarty, B. Edler, and E. A. Habets, "Countnet: Estimating the number of concurrent speakers using supervised learning," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 2, pp. 268–282, 2019.

[7] S. R. Chetupalli and E. A. P. Habets, "Speaker counting and separation from single-channel noisy mixtures," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 1681–1692, 2023.

[8] D. Pavlidi, A. Griffin, M. Puigt, and A. Mouchtaris, "Real-time multiple sound source localization and counting using a circular microphone array," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2193–2206, 2013.

[9] L. Wang, T.-K. Hon, J. D. Reiss, and A. Cavallaro, "An iterative approach to source counting and localization using two distant microphones," *IEEE/ACM Trans. on Audio, Speech, and Language processing*, vol. 24, no. 6, pp. 1079–1093, 2016.

[10] J. Azcarreta, N. Ito, S. Araki, and T. Nakatani, "Permutation-free CGMM: Complex Gaussian mixture model with inverse Wishart mixture model based spatial prior for permutation-free source separation and source counting," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, 2018, pp. 51–55.

[11] S. Hafezi, A. H. Moore, and P. A. Naylor, "Spatial consistency for multiple source direction-of-arrival estimation and source counting," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4592–4603, 2019.

[12] B. Laufer-Goldshtein, R. Talmon, and S. Gannot, "Source counting and separation based on simplex analysis," *IEEE Trans. on Signal Processing*, vol. 66, no. 24, pp. 6458–6473, 2018.

[13] Y. Hsu and M. R. Bai, "Learning-based robust speaker counting and separation with the aid of spatial coherence," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, pp. 36, 2023.

[14] Z.-Q. Wang and D. Wang, "Count and separate: Incorporating speaker counting for continuous speaker separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, 2021, pp. 11–15.

[15] K. Saijo, W. Zhang, Z.-Q. Wang, S. Watanabe, T. Kobayashi, and T. Ogawa, "A single speech enhancement model unifying dereverberation, denoising, speaker counting, separation, and extraction," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Taipei, Taiwan, 2023, pp. 1–6.

[16] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "High-resolution speaker counting in reverberant rooms using CRNN with ambisonics features," in *Proc. European Signal Processing Conference*, Amsterdam, Netherlands, 2020, pp. 71–75.

[17] M. Yousefi and J. H. Hansen, "Real-time speaker counting in a cocktail party scenario using attention-guided convolutional neural network," *arXiv preprint arXiv:2111.00316*, 2021.

[18] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Overlapped speech detection and speaker counting using distant microphone arrays," *Computer Speech & Language*, vol. 72, pp. 101306, 2022.

[19] H. Gode and S. Doclo, "Closed-form successive relative transfer function vector estimation based on blind oblique projection incorporating noise whitening," *arXiv preprint arXiv:2508.04887*, 2025.

[20] D. Ramírez, J. Via, and I. Santamaria, "A generalization of the magnitude squared coherence spectrum for more than two signals: definition, properties and estimation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, USA, 2008, pp. 3769–3772.

[21] Y. Avargel and I. Cohen, "On Multiplicative Transfer Function Approximation in the Short-Time Fourier Transform Domain," *IEEE Signal Processing Letters*, vol. 14, no. 5, pp. 337–340, 2007.

[22] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[23] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[24] D. Fejgin, W. Middelberg, and S. Doclo, "Brudex database: Binaural room impulse responses with uniformly distributed external microphones," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, 2023, [Online]. Available: `https://doi.org/10.5281/zenodo.7986446`.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, South Brisbane, Australia, 2015, pp. 5206–5210.