Schmidt *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2020) 2020:11
https://doi.org/10.1186/s13636-020-00178-0

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Motor data-regularized nonnegative matrix factorization for ego-noise suppression

Alexander Schmidt*, Andreas Brendel, Thomas Haubner, and Walter Kellermann

Abstract

Ego-noise, i.e., the noise a robot causes by its own motions, significantly corrupts the microphone signal and severely impairs the robot's capability to interact seamlessly with its environment. Therefore, suitable ego-noise suppression techniques are required. For this, it is intuitive to use also motor data collected by proprioceptors mounted to the joints of the robot since it describes the physical state of the robot and provides additional information about the ego-noise sources. In this paper, we use a dictionary-based approach for ego-noise suppression in a semi-supervised manner: first, an ego-noise dictionary is learned and subsequently used to estimate the ego-noise components of a mixture by computing a weighted sum of dictionary entries. The estimation of the weights is very sensitive against other signals beside ego-noise contained in the mixture. For increased robustness, we therefore propose to incorporate knowledge about the physical state of the robot to the estimation of the weights. This is achieved by introducing a motor data-based regularization term to the estimation problem which promotes similar weights for similar physical states. The regularization is derived by representing the motor data as a graph and imprints the intrinsic structure of the motor data space onto the dictionary model. We analyze the proposed method and evaluate its ego-noise suppression performance for a large variety of different movements and demonstrate the superiority of the proposed method compared to an approach without using motor data.

Keywords: Ego-noise, Motor data, Robot audition, Humanoïd robot

1 Introduction

Microphone-equipped robots are exposed to various kinds of noise, specifically to self-created noise, which is referred to as *ego-noise* in the following. It is caused by the robot's electrical and mechanical components such as rotating motors and joints as well as the moving body parts. Ego-noise is a crucial problem in robot audition [1, 2] since it severely corrupts the recorded microphone signals and impairs the robot's capability to react to unanticipated acoustic events. For this reason, ego-noise suppression is a crucial preprocessing step in robot audition.

Ego-noise suppression is particularly challenging for several reasons. First, ego-noise is usually louder than other signals of interest, e.g., a desired speech signal ("target"), since the ego-noise sources are typically located in immediate proximity of the microphones. For example, for the humanoïd robot NAO™, which we will use as experimental platform in this paper, the microphones are mounted to the head of the robot. Thereby, they are only few centimeters away from the shoulder motors and joints, cf. Fig. 1. Another challenging aspect of ego-noise is that it cannot be modeled as a single static point interferer as the joints are located all over the body of the robot and the resulting structure-borne sound is transduced to air not just at isolated points. Furthermore, ego-noise is highly non-stationary since typically different move-

*Correspondence: alexander.as.schmidt@fau.de
Multimedia Communications and Signal Processing, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Cauerstrasse 7, 91058 Erlangen, Germany

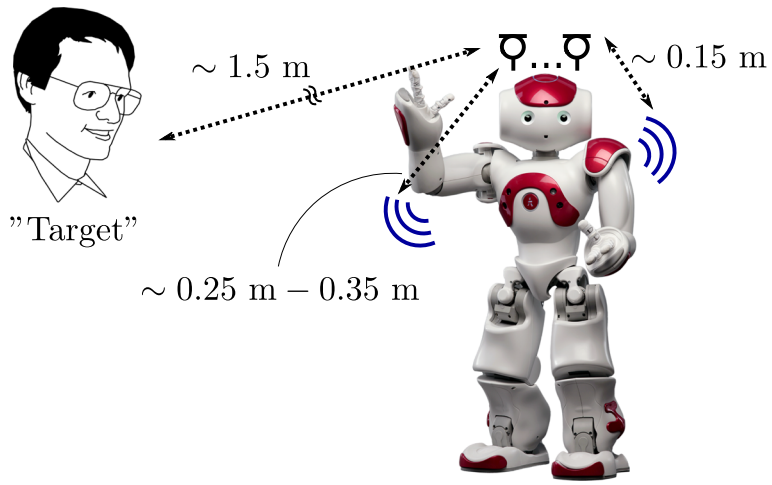

Fig. 1 Illustration of a typical human-robot interaction scenario. The desired source ("target") is shown on the left. Two exemplary, spatially distributed ego-noise sources are shown in blue. Besides, typical distances between the sources and the microphones are given. Obviously, the ego-noise sources are located close to the microphones while the distance between target and microphones is typically larger. Image of robot taken from [12]

ments are performed successively with varying speeds and accelerations.

One of the first approaches for ego-noise suppression goes back to the SIG humanoïd robot [1] which was equipped with microphones mounted inside the robot's housing near the motors in order to record ego-noise reference signals. These signals were subsequently used as reference for adaptive filtering-based ego-noise cancelation. Interestingly, these reference signals were interpreted as additional auditory perception channels of the robot. Internal microphones were also used in [3] for speech enhancement for a human-robot dialog system. In this approach, the recorded reference signals are incorporated into a frequency-domain semi-blind source separation algorithm with subsequent multichannel Wiener filtering.

In many robot designs, it is not possible to mount additional reference microphones inside the robot due to space and hardware constraints. Furthermore, a potentially large number of internal microphones are required to obtain reference signals for each ego-noise source. This drawback motivates approaches which operate on the external microphone signals only. Here, it can be exploited that ego-noise exhibits a characteristic structure in the Short-Time Fourier Transform (STFT) domain. Due to the limited number of degrees of freedom for the movements of the robot, those spectral patterns cannot be arbitrarily diverse. These two properties motivate the use of learning-based dictionary methods where the ego-noise signals are approximated by prototype signals, so-called atoms, which are collected in a dictionary. Then, for each time frame, a linear combination of atoms has to be found which optimally fits the current ego-noise signal with respect to the chosen criterion. An example for such a dictionary learning algorithm is K-SVD [4], which has been applied for multichannel ego-noise suppression in [5]. Another widely used approach to train a dictionary is *nonnegative matrix factorization* (NMF) [6–8]. For NMF, the dictionary is restricted to nonnegative elements only which is well-suited to model power spectral densities (PSDs) of acoustic sources. An according approach for ego-noise suppression has been investigated, e.g., in [9]. The concept of NMF has been extended for multichannel recordings [10] by extending the (nonnegative) source model by an additional spatial model. This has been applied for ego-noise suppression in [11].

Besides methods using the audio modality only (referred to as *audio only-based methods* in the following), other ego-noise suppression approaches use knowledge about motor information given by, e.g., motor commands or motor data such as engine rotation frequency, joints' angle, or angular velocities collected by proprioceptors. The advantage of using motor data compared to motor commands is that the emitted ego-noise is directly related to the instantaneous internal state of the robot, measured by motor data. Since a robot is not a fully deterministic system, this measured state may be significantly different from the target state defined by the motor command.

Typically, a sufficiently accurate analytical model of the dependency between motor data and emitted ego-noise can usually not be obtained since the mechanical dependencies and interactions between structure and airborne sounds are highly complex. Therefore, current ego-noise suppression approaches model these dependencies entirely or partly by learning-based strategies. For example in [13], a neural network-based approach is used

to predict the PSDs of ego-noise caused by the Aibo^{TM} robot. The feedforward neural network, consisting of two hidden layers containing thirty nodes each, is fed with angular position and velocity data of current and past time frames. The PSD estimates are subsequently used for spectral subtraction which was shown to result in a significant improvement of speech recognition rates. In [14], it is demonstrated that the harmonic structure of ego-noise can be estimated using motor data. This prior knowledge is included to a single-channel NMF-based ego-noise modeling. It is proposed to approximate the currently observed ego-noise spectrogram by combining elements from a dictionary $\boldsymbol{D}_{\mathrm{H}}$ which models the harmonic structure and another dictionary $\boldsymbol{D}_{\mathrm{R}}$ which captures the residual part of the ego-noise. The benefit of this approach is that only $\boldsymbol{D}_{\mathrm{R}}$ requires a prior learning step while $\boldsymbol{D}_{\mathrm{H}}$ is completely motor data-driven. It is shown that the proposed approach significantly outperforms an audio only-based method for the suppression of ego-noise that is not well represented in the training data. Although this approach is close to the proposed method from a methodical point of view, it aims at a different direction since it explicitly addresses the suppression of ego-noise if training and test data are unbalanced. This is not the case in the scenario considered in this paper.

Other popular methods for ego-noise suppression combining audio and motor information are template-based approaches. Here, the key idea is to save the characteristic spectral shape of the ego-noise as PSD templates in a data base. In [15], each template is associated with a motor command which triggered the current movement. Based on this, during application, matching templates are identified and temporally aligned to the recorded signal. An alternative template-based approach was presented in [16, 17], where motor data instead of motor commands are used to identify the templates in the data base. For a current motor data sample, the nearest neighbor in the motor data space is searched and the associated template used as ego-noise estimate.

The concept to associate motor data with ego-noise templates was adopted in [18]. However, there, motor data samples are linked to a set of atoms from a learned dictionary-based ego-noise model. Nonlinear classifiers in the motor data space are used to associate a motor data sample to a set of atoms, whose elements are subsequently combined to approximate the current ego-noise recording. Thereby, the classifiers replace the expensive iterative search for atoms in the dictionary.

In this paper, the idea of choosing atoms depending on motor data is adopted from [18] and we propose to expand the conventional, audio only-based NMF model by a motor data-dependent regularization term, which promotes similar atom activations in those time frames in which similar motor data is measured. The proposed regularization term is derived from a graph structure which encodes the similarity between the motor data samples. While the main benefit of the method in [18] was a reduction of computational complexity, the presented approach in this paper results in a significant performance improvement. The proposed method is inspired by graph-regularized NMF [19, 20], which was proposed in the context of clustering and classification of text documents. There, the NMF model and the regularization are operating in the same data space. In this work, however, we learn an NMF model on acoustic data while the regularization encodes the geometry of the motor data space. Thus, we combine an acoustic model with non-acoustic reference information.

This paper is structured as follows. In Section 2.1, we describe the used motor data. After succinctly introducing NMF in Section 2.2, we present the novel motor data-regularized NMF in Section 2.3 Thereby, we first describe the construction of the motor data graph structure in Subsection 2.3.1 and derive the proposed regularization term in Subsection. 2.3.2. Then, the modified NMF optimization problem is formulated and according update rules are presented in Subsection. 2.3.3. The resulting novel ego-noise suppression algorithm is summarized in Section 2.4 and its efficacy is demonstrated in Section 3.

2 Motor data-regularized NMF for ego-noise suppression

In the following, we consider the bin-wise squared magnitude of a single-channel microphone signal in the STFT domain, represented in spectrograms denoted as $\boldsymbol{Y} = \left[\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L\right] \in \mathbb{R}_+^{F \times L}$, where F is the number of frequency bins and L is the number of considered time frames.

2.1 Motor data descriptions and definitions

The physical state of a robot can be described by motor data, collected by proprioceptors providing angular position information of the joints driven by the motors. In the following, we consider a robot which is equipped with $m = 1, \ldots, M$ proprioceptors each capturing one angle of a joint. We denote the s-th observed angular position in STFT frame ℓ for proprioceptor m by $\alpha_{\ell,m}^{(s)} \in \mathbb{R}$. Within frame ℓ, a total number of S_ℓ motor data samples is observed, i.e., $s = 1, \ldots, S_\ell$. In this paper, we account for the fact that the motor data is not necessarily synchronized with the audio data recording so that for a fixed observation interval for the audio data, the number of motor data may vary, i.e., S_ℓ may change with ℓ. This is specifically the case for the NAO robot used for the experiments in this paper.

Depending on the kind of ego-noise, only a subset of proprioceptors is relevant for ego-noise suppression. For example, if only ego-noise caused by arm movements is present, only motor data of the arm joints are required.

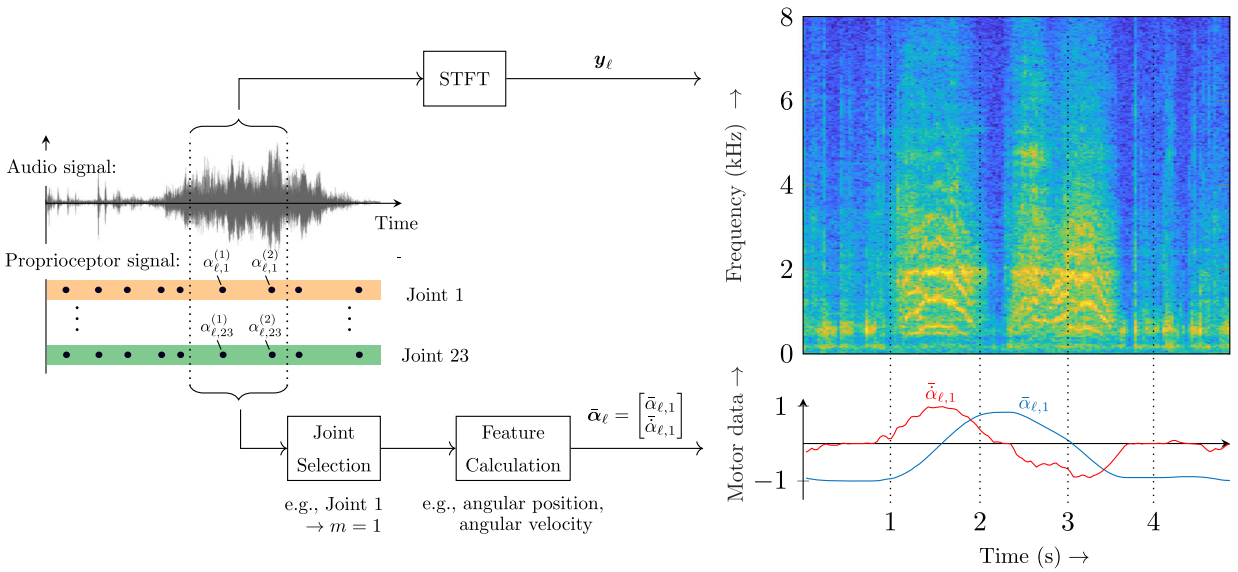

Fig. 2 Left: Illustration of data collection and processing. We consider a robot for which motor data of 23 joints and a single-channel microphone signal are recorded. The ℓth STFT frame is processed. Note that the total number of motor data samples can be varying from frame to frame, c.f., Section 2.1. **Right:** Example spectrogram for right arm shoulder ego-noise and corresponding normalized motor data samples (angular position and angular velocity) of this joint

In the following, we denote the index set of relevant proprioceptors for these joints by $\mathcal{M}$.

From proprioceptor data collected for proprioceptor m, the instantaneous angular velocity can be estimated by

$$\dot{\alpha}_{\ell,m}^{(s)} = \frac{\alpha_{\ell,m}^{(s)} - \alpha_{\ell,m}^{(s-1)}}{\Delta T_\ell^{(s)}}, \quad m \in \mathcal{M}, \tag{1}$$

where $\Delta T_\ell^{(s)}$ denotes the time difference between adjacent observations $\alpha_{\ell,m}^{(s)}$ and $\alpha_{\ell,m}^{(s-1)}$. Note that for $s = 1$, $\alpha_{\ell,m}^{(s-1)}$ is chosen to be the last angular sample of previous the frame $\ell - 1$. Analogously, angular acceleration $\ddot{\alpha}_{\ell,m}^{(s)}$ can be computed from successive angular velocity estimates $\dot{\alpha}_{\ell,m}^{(s)}$ and $\dot{\alpha}_{\ell,m}^{(s-1)}$.

To associate each spectrogram frame $\boldsymbol{y}_\ell$ with a single motor data sample, we propose first to compute the arithmetic average of all S_ℓ angular positions in STFT frame ℓ

$$\bar{\alpha}_{\ell,m} = \frac{1}{S_\ell} \sum_{s=1}^{S_\ell} \alpha_{\ell,m}^{(s)}, \quad m \in \mathcal{M}. \tag{2}$$

We proceed analogously for angular velocity and acceleration and obtain $\dot{\bar{\alpha}}_{\ell,m}, \ddot{\bar{\alpha}}_{\ell,m}$, respectively. We then concatenate the averaged angular data for all considered proprioceptors in a feature vector

$$\bar{\boldsymbol{\alpha}}_\ell = \left[\bar{\alpha}_{\ell,1}, \ldots, \bar{\alpha}_{\ell,m}, \dot{\bar{\alpha}}_{\ell,m}, \ddot{\bar{\alpha}}_{\ell,m}, \ldots, \ddot{\bar{\alpha}}_{\ell,M}\right]^{\mathrm{T}}, \tag{3}$$

which we will refer to as motor data vector for frame ℓ in the following. The left part of Fig. 2 exemplarily illustrates the described preprocessing of the data.

2.2 NMF for ego-noise suppression

In the following, we briefly summarize NMF. We introduce succinctly how semi-supervised NMF can be used for ego-noise suppression and explain the main drawback of the known approach before we introduce the proposed motor data-based regularization.

The objective of NMF is to approximate the nonnegative matrix $\boldsymbol{Y}$, i.e., a matrix whose elements are all larger or equal than zero, by a product of two nonnegative matrices $\boldsymbol{D}$ and $\boldsymbol{H}$

$$\boldsymbol{Y} \approx \hat{\boldsymbol{Y}} = \boldsymbol{D}\boldsymbol{H} = [\boldsymbol{D}\boldsymbol{h}_1, \ldots, \boldsymbol{D}\boldsymbol{h}_L], \tag{4}$$

where $\boldsymbol{D} \in \mathbb{R}_+^{F \times K}$ is the so-called dictionary of size $F \times K$ and $\boldsymbol{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_L] \in \mathbb{R}_+^{K \times L}$ is referred to as activation matrix [8, 21]. This approach can be interpreted as approximating each column of $\boldsymbol{Y}$ by a weighted sum of columns of $\boldsymbol{D}$ (the so-called *atoms* or *bases*), where the weights are given by the corresponding column entries of $\boldsymbol{H}$. K is referred to as *size* of the dictionary and describes the number of atoms in $\boldsymbol{D}$. Typically, $K \ll F, L$ holds, i.e., NMF can be considered as a compact representation of data.

The factorization is achieved by minimizing a cost function which penalizes the dissimilarity between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ defined by the model parameters $\boldsymbol{D}, \boldsymbol{H}$. Typically, the cost function is applied element-wise on the elements of the

matrices $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$. In this paper, we consider the Euclidean distance between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ as cost function yielding the optimization problem

$$\begin{aligned} &\min_{\boldsymbol{D},\boldsymbol{H}} \ \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{H}\|_{\mathrm{F}}^2 \\ &\text{s.t.} \quad \boldsymbol{D},\boldsymbol{H} \succeq 0, \end{aligned} \tag{5}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm and $\boldsymbol{D},\boldsymbol{H} \succeq 0$ means that all elements of $\boldsymbol{D},\boldsymbol{H}$ are larger or equal to zero, ensuring nonnegativity. The optimization problem in Eq. 5 is typically solved using iterative updates alternating between $\boldsymbol{D},\boldsymbol{H}$ such that the nonnegativity of $\boldsymbol{D},\boldsymbol{H}$ is implicitly guaranteed if they are initialized with positive values. The update rules can be derived based on, e.g., the Majorization-Minimization principle or heuristic approaches [7, 8].

For ego-noise suppression, we apply a semi-supervised, two-stage strategy [21], c.f. Section 2.4: first, we use audio data containing ego-noise only and train an ego-noise dictionary. Then, given a mixture of ego-noise and speech, these dictionary elements remain constant and only its activations are estimated. For this, again, the same iterative update rules are used, which have shown to be sensitive to the additional speech signal. As a consequence, the atom activations are no longer estimated correctly. For improved robustness, we therefore propose to extend this audio only-based estimation of the activations by taking also the physical state of the robot, measured in terms of motor data, into account. Thus, the estimation of the activations is additionally guided by reference information which is completely unaffected by the speech signal.

2.3 Motor data-regularized NMF

The basic idea of our approach is that activations should be similar if the physical state of the robot is similar. For this, we measure the similarity between robot states in frames ℓ and j by comparing motor data vectors $\bar{\boldsymbol{\alpha}}_\ell$ and $\bar{\boldsymbol{\alpha}}_j$ and enforce similar activations $\boldsymbol{h}_\ell$ and $\boldsymbol{h}_j$ if $\bar{\boldsymbol{\alpha}}_\ell$ and $\bar{\boldsymbol{\alpha}}_j$ are close. This will be achieved by imprinting the intrinsic geometry of the motor data space to the NMF cost function. Results from spectral graph theory [22, 23] and manifold learning theory [24] have shown that local geometric structure of given data points can be modeled using an undirected graph. Based on these results, we first introduce a motor data-based graph structure and summarize subsequently how a regularization term, enforcing similar activations for similar motor data, can be derived. We then reformulate the NMF optimization problem Eq. 5 and present according update rules for its minimization.

2.3.1 Motor data graph structure

In the following, we define a graph where the motor data vectors $\bar{\boldsymbol{\alpha}}_1,\ldots,\bar{\boldsymbol{\alpha}}_L$ constitute the nodes. The edges connecting the nodes are assumed to be bidirectional, i.e., we obtain an undirected graph. A part of an exemplary graph is illustrated in Fig. 3. The edge which connects nodes $\bar{\boldsymbol{\alpha}}_\ell$ and $\bar{\boldsymbol{\alpha}}_j$ has weight $W_{\ell j} = W_{j\ell}$ and should reflect the affinity between the two motor data points. Dependent on the considered scenario, numerous measures have been proposed to quantify the affinity between $\bar{\boldsymbol{\alpha}}_\ell$ and $\bar{\boldsymbol{\alpha}}_j$ [22], e.g., a nearest-neighbor or dot-product weighting. In this paper, we determine the weight $W_{\ell j}$ using a Gaussian kernel

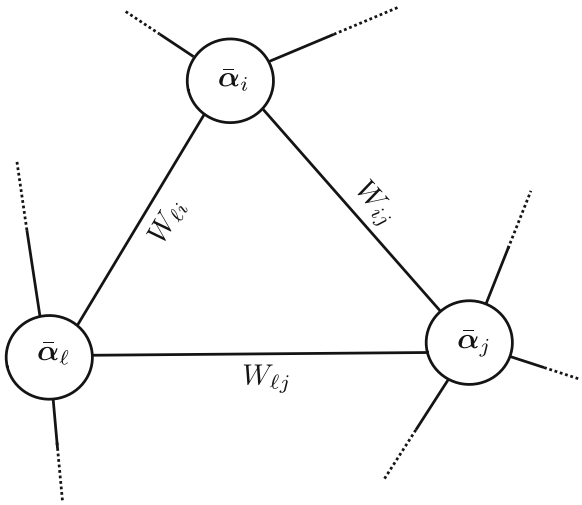

Fig. 3 Part of an undirected graph where motor data vectors $\bar{\boldsymbol{\alpha}}_\ell$, $\ell = 1,\ldots,L$ are the nodes of the graph. The weights of the edges are given by Eq. 6: the more similar two motor data vectors are, the larger the weight of the connecting edge is

$$W_{\ell j} = W_{j\ell} = \exp\left(-\frac{\|\bar{\boldsymbol{\alpha}}_\ell - \bar{\boldsymbol{\alpha}}_j\|_2^2}{2\epsilon^2}\right) \in (0,1], \tag{6}$$

with scale parameter $\epsilon \in \mathbb{R}_+$. The larger $W_{\ell j}$, the higher the affinity between two motor data samples is and we obtain $W_{\ell j} = 1$ if $\bar{\boldsymbol{\alpha}}_\ell = \bar{\boldsymbol{\alpha}}_j$. Note that by adjusting ϵ, the connectivity of the graph can be controlled, e.g., for larger ϵ, the neighbors of a node are connected with a larger weight. Therefore, ϵ can be used to control the reach of the local neighborhood of a node. Based on the affinity weights, we define the affinity matrix $\boldsymbol{W} = \boldsymbol{W}^{\mathrm{T}} \in [0,1]^{L\times L}$, where the $[\boldsymbol{W}]_{\ell j} = W_{\ell j}$. Furthermore, we introduce the diagonal matrix $\boldsymbol{Z}$ of size $L \times L$ with $Z_{\ell\ell} = \sum_j W_{\ell j} = \sum_j W_{j\ell}$ and zero else.

2.3.2 Motor data-based regularization term

The derivation of the regularization term is based on results from [24, 25]. It is assumed that the considered motor data lie on a Riemannian manifold $\mathcal{A}$. We are looking for a mapping $f : \mathcal{A} \to \mathbb{R}$, which can be interpreted as a mapping from the manifold to a line. f should preserve the local geometry of the manifold, i.e., close points on the manifold should be mapped to close points on the

line. This implies that f is allowed to vary only smoothly for similar arguments. Appropriate mappings f can be obtained by an optimization on the manifold which can be discretely approximated on the motor data graph by searching for an f which minimizes

$$\frac{1}{2}\sum_{\ell=1}^{L}\sum_{j=1}^{L}\left(f(\bar{\boldsymbol{\alpha}}_\ell)-f(\bar{\boldsymbol{\alpha}}_j)\right)^2 W_{\ell j}, \tag{7}$$

where f is a function of the nodes of the graph [24, 25].

To exploit the geometric information of the motor data manifold for the estimation of the activation vectors, we manipulate Eq. 7 and replace the abstract mapping f by the activation of atom k

$$\mathcal{R}_k = \frac{1}{2}\sum_{\ell=1}^{L}\sum_{j=1}^{L}\left(h_{k\ell}-h_{kj}\right)^2 W_{\ell j}, \tag{8}$$

where $h_{k\ell}$ denotes the ℓ-th element of $\boldsymbol{h}_k$, i.e., $h_{k\ell}$ is the scaling of atom ℓ in time frame k. The regularization term $\mathcal{R}_k$ needs to be minimized jointly with Eq. 5 with respect to the activations for every atom k, c.f. Section 2.3.3. Note that the motor data-based regularization $\mathcal{R}_k$ implicitly influences also the structure of the dictionary elements since the optimized activations directly affect the update of $\boldsymbol{D}$.

Note that in Eq. 8, affinities $W_{\ell j}$ can be interpreted as weighting parameter: if two motor data vectors $\bar{\boldsymbol{\alpha}}_\ell$ and $\bar{\boldsymbol{\alpha}}_j$ are similar, $W_{\ell j}$ is close to one according to Eq. 6 and the minimization of Eq. 8 enforces similar $h_{k\ell}$ and h_{kj}. Using the parameters defined in Section 2.3.1, Eq. 8 can be directly related to the so-called graph Laplacian $\boldsymbol{L} = \boldsymbol{Z} - \boldsymbol{W}$ [22]

$$\begin{aligned}\mathcal{R}_k &= \boldsymbol{h}_k^{\mathrm{T}}\mathbf{Z}\boldsymbol{h}_k - \boldsymbol{h}_k^{\mathrm{T}}\mathbf{W}\boldsymbol{h}_k \\ &= \boldsymbol{h}_k^{\mathrm{T}}\mathbf{L}\boldsymbol{h}_k.\end{aligned} \tag{9}$$

Summing over all atoms results in the final regularization term

$$\mathcal{R} = \sum_{k=1}^{K}\mathcal{R}_k = \mathrm{tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right), \tag{10}$$

where $\mathrm{tr}(\cdot)$ denotes the trace operator.

2.3.3 Motor data-regularized NMF

The derived regularization term Eq. 10 can be directly included into Eq. 4. We obtain as modified optimization problem

$$\begin{aligned}&\min_{\boldsymbol{D},\boldsymbol{H}} \ \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{H}\|_{\mathrm{F}}^2 + \lambda\,\mathrm{tr}\left(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{H}\right) \\ &\text{s.t.} \quad \boldsymbol{D},\boldsymbol{H} \succeq 0,\end{aligned} \tag{11}$$

where $\lambda \geq 0$ controls the influence of the motor data-based regularization.

For minimization, we form the partial derivatives with respect to $\boldsymbol{D}$ and $\boldsymbol{H}$ in Eq. 11 and obtain iterative update rules [19, 20]

Learning of $\boldsymbol{D}$ (Training)

Input:
- Ego-noise recording $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L]$
- Motor data $\bar{\boldsymbol{\alpha}}_1, \ldots, \bar{\boldsymbol{\alpha}}_L$

Learning algorithm:
- Initialize $\boldsymbol{D}, \boldsymbol{H}$ randomly
- Construct $\boldsymbol{W}, \boldsymbol{Z}$ from $\bar{\boldsymbol{\alpha}}_1, \ldots, \bar{\boldsymbol{\alpha}}_L$
- Repeat until convergence:

$$[\boldsymbol{D}]_{fk} \leftarrow [\boldsymbol{D}]_{fk} \cdot \frac{\left[\boldsymbol{Y}\boldsymbol{H}^{\mathrm{T}}\right]_{fk}}{\left[\hat{\boldsymbol{Y}}\boldsymbol{H}^{\mathrm{T}}\right]_{fk}} \quad \forall f,k$$

$$[\boldsymbol{H}]_{k\ell} \leftarrow [\boldsymbol{H}]_{k\ell} \cdot \frac{\left[\boldsymbol{D}^{\mathrm{T}}\boldsymbol{Y} + \lambda_{\mathrm{T}}\boldsymbol{H}\boldsymbol{W}\right]_{k\ell}}{\left[\boldsymbol{D}^{\mathrm{T}}\hat{\boldsymbol{Y}} + \lambda_{\mathrm{T}}\boldsymbol{H}\boldsymbol{Z}\right]_{k\ell}} \quad \forall k,\ell$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{D}\boldsymbol{H}$$

Output: $\boldsymbol{D}$

Ego-noise suppression (Evaluation)

Input:
- Mixture of ego-noise and speech $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L]$
- Motor data $\bar{\boldsymbol{\alpha}}_1, \ldots, \bar{\boldsymbol{\alpha}}_L$
- Ego-noise dictionary $\boldsymbol{D}$ from learning step

Suppression algorithm:
- Initialize $\boldsymbol{D}_{\mathrm{S}}, \boldsymbol{H}_{\mathrm{S}}, \boldsymbol{H}$ randomly
- Construct $\boldsymbol{W}, \boldsymbol{Z}$ from $\bar{\boldsymbol{\alpha}}_1, \ldots, \bar{\boldsymbol{\alpha}}_L$
- Repeat until convergence:

$$[\boldsymbol{D}_{\mathrm{S}}]_{fk} \leftarrow [\boldsymbol{D}_{\mathrm{S}}]_{fk} \cdot \frac{\left[\boldsymbol{Y}\boldsymbol{H}_{\mathrm{S}}^{\mathrm{T}}\right]_{fk}}{\left[\hat{\boldsymbol{Y}}\boldsymbol{H}_{\mathrm{S}}^{\mathrm{T}}\right]_{fk}} \quad \forall f,k$$

$$[\boldsymbol{H}]_{k\ell} \leftarrow [\boldsymbol{H}]_{k\ell} \cdot \frac{\left[\boldsymbol{D}^{\mathrm{T}}\boldsymbol{Y} + \lambda_{\mathrm{E}}\boldsymbol{H}\boldsymbol{W}\right]_{k\ell}}{\left[\boldsymbol{D}^{\mathrm{T}}\hat{\boldsymbol{Y}} + \lambda_{\mathrm{E}}\boldsymbol{H}\boldsymbol{Z}\right]_{k\ell}} \quad \forall k,\ell$$

$$[\boldsymbol{H}_{\mathrm{S}}]_{k\ell} \leftarrow [\boldsymbol{H}_{\mathrm{S}}]_{k\ell} \cdot \frac{\left[\boldsymbol{D}_{\mathrm{S}}^{\mathrm{T}}\boldsymbol{Y}\right]_{k\ell}}{\left[\boldsymbol{D}_{\mathrm{S}}^{\mathrm{T}}\hat{\boldsymbol{Y}}\right]_{k\ell}} \quad \forall k,\ell$$

$$\hat{\boldsymbol{Y}} = \begin{bmatrix}\boldsymbol{D} & \boldsymbol{D}_{\mathrm{S}}\end{bmatrix}\begin{bmatrix}\boldsymbol{H} \\ \boldsymbol{H}_{\mathrm{S}}\end{bmatrix} = \boldsymbol{D}\boldsymbol{H} + \boldsymbol{D}_{\mathrm{S}}\boldsymbol{H}_{\mathrm{S}}$$

- Compute $\left[\hat{\boldsymbol{Y}}_{\mathrm{S}}\right]_{f\ell} = \left[\boldsymbol{F}\right]_{f\ell} \cdot \left[\boldsymbol{Y}\right]_{f\ell}$ using Eq. 14

Output: estimate of desired speech signal $\hat{\boldsymbol{Y}}_{\mathrm{S}}$

Fig. 4 Overview of the proposed semi-supervised, two-stage algorithm for ego-noise suppression: Learning of $\boldsymbol{D}$ (left) and ego-noise suppression (right)

$$[\boldsymbol{D}]_{fk} \leftarrow [\boldsymbol{D}]_{fk} \cdot \frac{\left[\boldsymbol{Y}\boldsymbol{H}^{\mathrm{T}}\right]_{fk}}{\left[\hat{\boldsymbol{Y}}\boldsymbol{H}^{\mathrm{T}}\right]_{fk}}, \tag{12}$$

$$[\boldsymbol{H}]_{k\ell} \leftarrow [\boldsymbol{H}]_{k\ell} \cdot \frac{\left[\boldsymbol{D}^{\mathrm{T}}\boldsymbol{Y} + \lambda\boldsymbol{H}\boldsymbol{W}\right]_{k\ell}}{\left[\boldsymbol{D}^{\mathrm{T}}\hat{\boldsymbol{Y}} + \lambda\boldsymbol{H}\boldsymbol{Z}\right]_{k\ell}}, \tag{13}$$

where $[\boldsymbol{D}]_{fk}$ selects the fk-th element from $\boldsymbol{D}$. Similar to conventional NMF, the iterative update can be stopped, e.g., after a fixed number of iterations. In this paper, in each iteration we additionally compute the cost according to Eq. 11 and terminate updating Eqs. 12,13 after convergence.

Eqs. 12 and 13 reduce to the conventional update rules for NMF if $\lambda = 0$ [8]. Note that since the proposed method aims at enforcing similar activations for close motor data vectors, the regularization has an effect on the update rule for $\boldsymbol{H}$ only, while the update for $\boldsymbol{D}$ is unaffected.

2.4 Proposed algorithm for ego-noise suppression

As mentioned in Section 2.2, we apply a semi-supervised, two-stage strategy for ego-noise suppression [21]. We first employ audio data containing ego-noise only and train $\boldsymbol{D}$ imprinting the intrinsic geometry of the motor data space onto the model using the proposed regularization. Given a mixture of ego-noise and speech, we use $\boldsymbol{D}$ to model and suppress the current ego-noise and to obtain a speech estimate. In the following, we describe the proposed algorithm for ego-noise suppression in detail, c.f. Fig. 4 for an overview.

- **Learning $\boldsymbol{D}$:** As input, spectrograms $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots \boldsymbol{y}_L]$ are given containing ego-noise only. Per spectrogram frame $\boldsymbol{y}_\ell$, a motor data vector $\bar{\boldsymbol{\alpha}}_\ell$ is computed. $\bar{\boldsymbol{\alpha}}_\ell, \ell = 1, \ldots, L$ is used to construct the affinity and degree matrix, $\boldsymbol{W}$ and $\boldsymbol{Z}$, respectively. Subsequently, the update rules Eqs. 12 and 13 are used to compute dictionary $\boldsymbol{D}$, where the introduced regularization term is weighted by λ_{T}.
- **Ego-noise suppression:** Another dictionary $\boldsymbol{D}_{\mathrm{S}}$ of size K_{S} and according activation $\boldsymbol{H}_{\mathrm{S}}$ is initialized to model the additional speech signal in the considered mixture $\boldsymbol{Y}$. Analogously to the learning step before, $\boldsymbol{W}$ and $\boldsymbol{Z}$ are constructed from the new motor data vectors possibly representing different movements. Using the same update rules as before, $\boldsymbol{D}_{\mathrm{S}}$, $\boldsymbol{H}$ and $\boldsymbol{H}_{\mathrm{S}}$ are updated while $\boldsymbol{D}$ remains constant. The motor data-based regularization term is weighted by λ_{E}. Note that for optimizing the activations of the speech model $\boldsymbol{H}_{\mathrm{S}}$, we set $\lambda_{\mathrm{E}} = 0$ since the motor data-based regularization should affect only the estimation of the ego-noise activations. After identifying the optimum model parameters captured by $\boldsymbol{D}_{\mathrm{S}}$, $\boldsymbol{H}$ and $\boldsymbol{H}_{\mathrm{S}}$, we use a spectral enhancement filter to obtain an estimate for the desired speech signal $\left[\hat{\boldsymbol{Y}}_{\mathrm{S}}\right]_{f\ell} = [\boldsymbol{F}]_{f\ell} \cdot [\boldsymbol{Y}]_{f\ell}$ for the $f\ell$-th bin where the enhancement filter is given by

$$[\boldsymbol{F}]_{f\ell} = \frac{\left[\boldsymbol{D}_{\mathrm{S}}\boldsymbol{H}_{\mathrm{S}}\right]_{f\ell}}{[\boldsymbol{D}\boldsymbol{H}]_{f\ell} + \left[\boldsymbol{D}_{\mathrm{S}}\boldsymbol{H}_{\mathrm{S}}\right]_{f\ell}}. \tag{14}$$

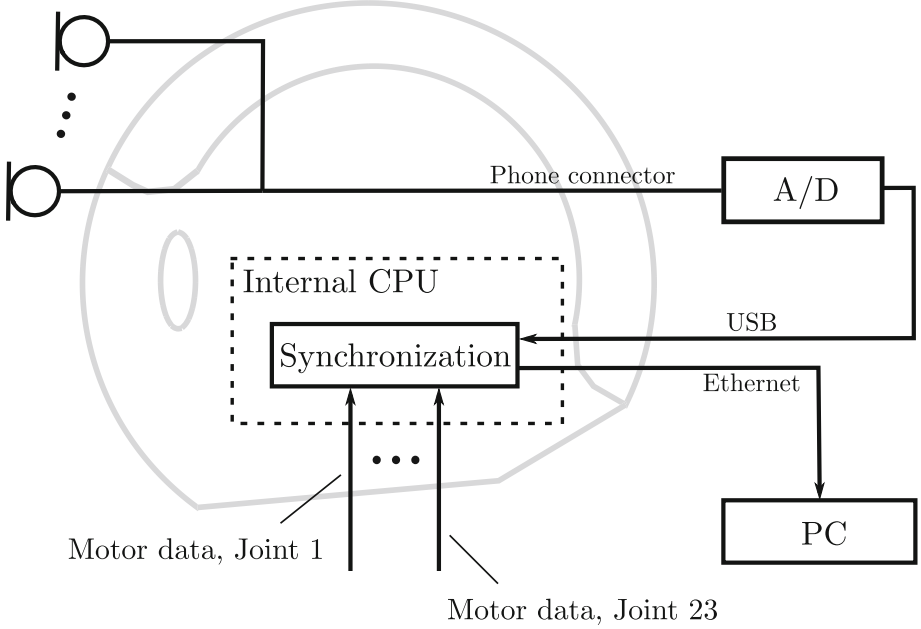

Fig. 5 Overview of recording setup. A detailed explanation is given in Section 3.1

Note that typically $\lambda_E \neq \lambda_T$ holds, i.e, the regularization terms in both steps have different weights. This is further detailed in the following section.

3 Experimental evaluation

In the following, we evaluate the proposed method using real microphone recordings. We first describe the hardware setup, the synchronization of audio and motor data and the recording scenarios, and introduce the evaluation metrics. Then, we present suppression results for ego-noise of different movements and discuss the influence of crucial parameters.

3.1 Recording setup

For our experiments, we conducted experiments with a commercially available NAO H25 robot [12]. For the audio recordings, we used a self-constructed head [26] with a microphone array of 12 sensors. For all following experiments, we used the frontmost microphone. Since the NAO platform does not provide an in-built synchronization on audio sample level, we developed a synchronization scheme which is illustrated in Fig. 5: the microphone signals are fed into an external analog-to-digital (A/D) converter using conventional phone connectors (IEC 60603-11). The sampled data is forwarded to the robot's internal CPU via USB, where it is synchronized with the motor data collected by the proprioceptors of the robot. The resulting data stream, containing audio and motor data, is finally transmitted to an external PC via Ethernet, which is used for recording.

3.2 Scenario description

The recordings were conducted in a room with moderate reverberation ($T_{60} = 200$ ms). We investigate ego-noise of different right arm movements of the robot. Compared to movement noise of other body parts, ego-noise of the arms has the most severe effect on the microphone signals due to the immediate closeness of the active joints to the microphones. In total, we recorded ego-noise of three motion sequences:

- *Sequence I* consists of repeating right arm waving movements, activating all six joints of the arm. The robot lifts the arm using the right shoulder pitch motor, while performing waving movements with the remaining five motors of the right arm.
- *Sequence II* resembles Sequence I; however, the lifting of the arm is performed with randomly varying velocity and acceleration of the right shoulder pitch motor. The number of employed joints is identical to *Sequence I*.
- *Sequence III* is a mixture of left and right arm movements where both left and right joints are controlled independently with varying speeds. Since movements of the left and right arm are considered, 12 joints are used in total, i.e., compared to *Sequence I/II* the number of joints is doubled.

While *Sequence I* is a relatively simple scenario due to its repetitive character, *Sequence II* and *Sequence III* are more challenging for a description by a dictionary. For *Sequence II*, the random accelerations of the right shoulder pitch motor result in a large variety of spectral patterns which must be captured by the dictionary. The same holds for ego-noise of *Sequence III*, where the doubling of employed joints causes more spectral diversity.

The recorded ego-noise was used for training the dictionary and evaluation, where the data for evaluation was not contained in the training data. In total, we recorded 60 s for each motion sequence and split the ego-noise data such that approximately 30 s ego-noise for the learning of $\boldsymbol{D}$ is available.

To evaluate the suppression performance, we consider a scenario in which a target source is talking to the robot. The robot is standing on the floor level while it performs different waving movements of the right arm. The microphones of the robot are at a height of 55 cm. For the speech signal, utterances from male and female speakers of the GRID corpus [27] were used. The loudspeaker was positioned at 1 m distance of the robot, at a height of 1 m. The recorded reverberant utterances were added to the ego-noise with varying signal-to-noise (SNR) ratios (see Section 3.3).

The audio signals are sampled at $f_S = 16$ kHz and transformed to the STFT domain using a Hamming window of length 64 ms with overlap of 50 %. The internal operating system of the NAO robot saves motor data samples of all joints into an internal cache which can be accessed by the user. This cache is typically updated every 10 ms. Consequently, the sampling frequency of the motor data is given by $f_S \approx 100$ Hz, i.e, typically $S_\ell = 6$ motor data samples are available per time frame.

We evaluated the overall performance of the ego-noise suppression in terms of signal-to-distortion ratio (SDR in dB) and signal-to-artifacts ratio (SAR in dB). For the

Table 1 Performance achieved by the proposed method and audio only-based NMF for different λ_T and λ_E. Ego-noise is caused by movements of Sequence I ($K = 20$, $K_S = 20$, SNR = 3 dB, $\epsilon = 5 \cdot 10^{-3}$)

		Proposed		
	NMF	$\lambda_T = 0.9$	$\lambda_T = 0.0$	$\lambda_T = 0.9$
		$\lambda_E = 0.0$	$\lambda_E = 19.0$	$\lambda_E = 19.0$
SDR [dB]	7.37 (± 0.11)	7.7 (± 0.1)	9.5 (± 0.03)	**9.77** (± 0.03)
SAR [dB]	9.63 (± 0.18)	10.09 (± 0.1)	12.1 (± 0.01)	**12.64** (± 0.03)
PESQ	1.35 (± 0.02)	1.36 (± 0.02)	1.42 (± 0.01)	**1.45** (± 0.01)

computation of both, Matlab functions provided by [28] are used. In practice, it must be expected that the ego-noise and speech estimates, i.e., $\boldsymbol{DH}$ and $\boldsymbol{D}_{\mathrm{S}}\boldsymbol{H}_{\mathrm{S}}$, contain estimation errors resulting in imperfect enhancement filter $\boldsymbol{F}$, cf. Eq. 14. As a consequence, the ego-noise cannot be removed entirely from the mixture and/or the desired speech is distorted. The severity of both effects is reflected in the selected performance criteria SDR and SAR, respectively. While SDR measures the amount of remaining ego-noise and speech distortion after processing, SAR considers introduced speech distortion only. For unprocessed data, the SDR corresponds to the SNR of the input mixture while SAR is infinite. Beside SDR and SAR, we also evaluate PESQ (perceptual evaluation of speech quality [29]). To obtain representative results, we averaged over 100 runs with random initialization of the matrices in NMF. Standard deviations for all results are given in brackets.

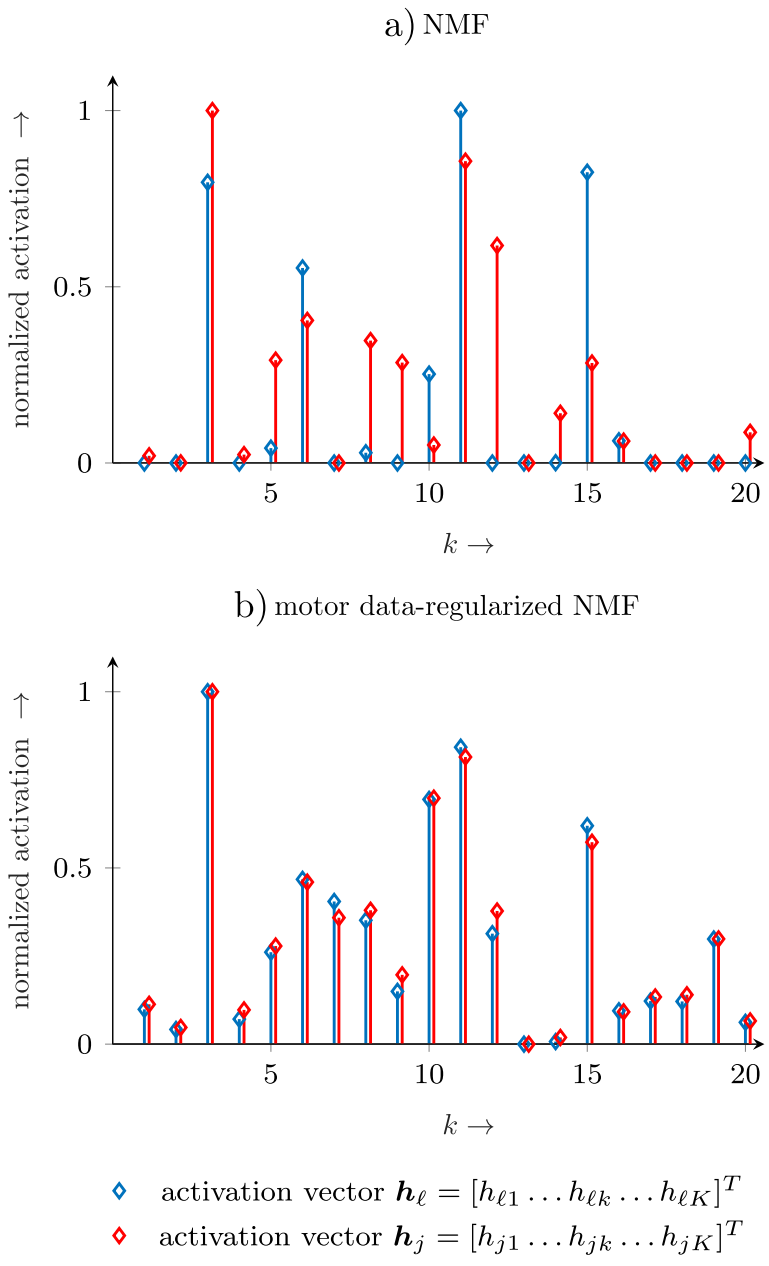

Fig. 6 Estimated ego-noise activations by **a** NMF and **b** motor data-regularized NMF for time frames ℓ and j, where ℓ and j were chosen such that $W_{\ell j}$ is large (here $W_{\ell j} \approx 0.9$). The considered mixture of ego-noise (Sequence I) and speech has SNR = 3 dB, dictionary sizes $K = 20$ and $K_{\mathrm{S}} = 20$. For motor data-regularized NMF, $\lambda_{\mathrm{T}} = 0.9$, $\lambda_{\mathrm{E}} = 19$ were used

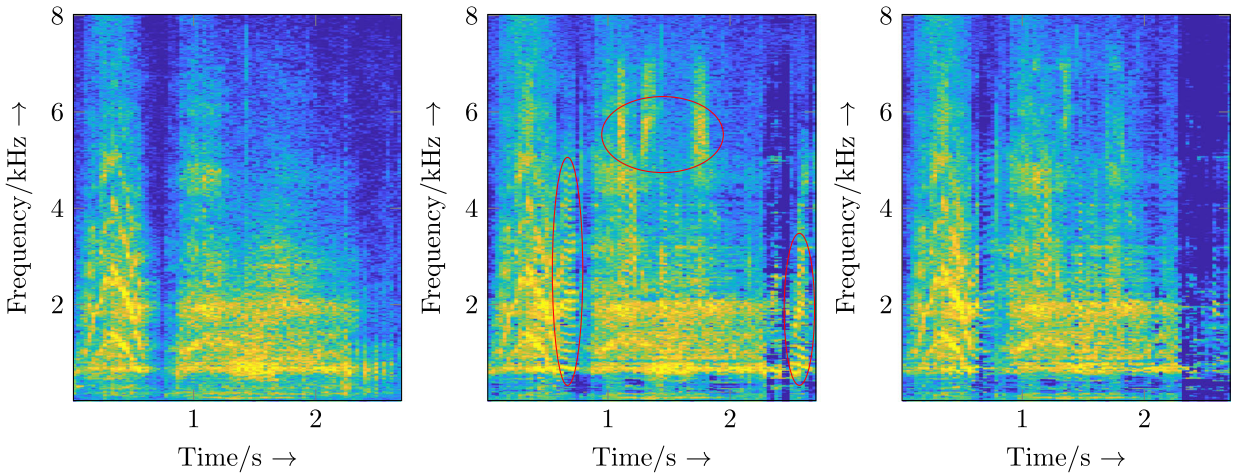

Fig. 7 Spectrograms of an exemplary ego-noise extract from Sequence II (logarithm of magnitude). The left figure shows the original ego-noise. In the center and right figures, ego-noise estimates using audio-only NMF and the proposed method are depicted, respectively. Without motor data regularization, the ego-noise is in parts incorrectly estimated and adapts to the speech signal. Corresponding components are marked with red ellipses (middle). Note that these undesired components are not or significantly less visible in the ego-noise estimate of the proposed method (right)

3.3 Evaluation and discussion of the results

For evaluating the proposed method for ego noise of motion Sequence I, II, and III, the size of the ego-noise dictionary and speech dictionary has been chosen to $K = K_S = 20$ for Sequence I and $K = 30$, $K_S = 20$ for Sequence II and III. These parameters have shown best suppression performance in terms of SDR for audio only-based NMF on the respective ego-noise recordings. We first discuss the choice of λ_T and λ_E and illustrate the effect of the regularization term $\mathcal{R}$. Subsequently, we evaluate the suppression performance for different SNRs and finally discuss alternative choices for the motor data vector $\bar{\boldsymbol{\alpha}}_\ell$.

3.3.1 *Impact and choice of λ_T and λ_E*

In Table 1, the suppression results for particular choices of λ_T and λ_E are given. First, we incorporate the motor data information only into the training ($\lambda_T = 0.9$) and leave the suppression step unchanged $\lambda_E = 0$. Compared to audio only-based NMF (denoted as "NMF" in the following), this already shows a slight improvement of the results. For $\lambda_T = 0$ and $\lambda_E = 19$, we obtain significantly better results than for NMF, which shows that enforcing similar activations for similar physical states of the robot helps even if this constraint has not be learnt during the learning of the dictionary. Best results are obtained if the regularization term is included to both learning and suppression. Note that λ_T and λ_E are of different orders of magnitude, what will be further investigated and interpreted in Section 3.3.2.

The effect of the proposed regularization term is illustrated in Fig. 6. We consider two time frames ℓ and j of a mixture of ego-noise (Sequence I) and speech. Frames ℓ and j are chosen such that $W_{\ell j}$ is large, i.e., $\bar{\boldsymbol{\alpha}}_\ell$ and $\bar{\boldsymbol{\alpha}}_j$ are close indicating that the robot has similar physical states. Hence, similar activations are desired. Figure 6a shows elements of the activation vectors $\boldsymbol{h}_\ell$ and $\boldsymbol{h}_j$ obtained if audio only-based NMF is used. It is obvious that the activations differ significantly, which can be explained by the additional speech signal present in frames ℓ and j which affects the estimation of the

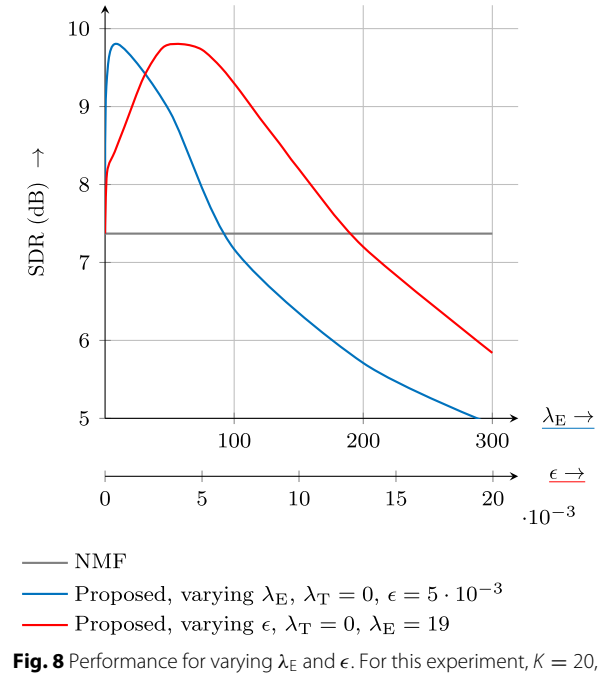

Fig. 8 Performance for varying λ_E and ϵ. For this experiment, $K = 20$, $K_S = 20$, SNR = 3 dB is chosen and ego-noise caused by movements of Sequence I is used

ego-noise activations. Figure 6 b illustrates the elements of $\boldsymbol{h}_\ell$ and $\boldsymbol{h}_j$ estimated by proposed motor data-regularized NMF. Here, in contrast to audio only-based NMF, the activations coincide even if additional speech is present.

For further illustration, Fig. 7 shows spectrograms of an ego-noise extract and its estimates using audio-only NMF and the proposed method. Without motor data regularization, the speech signal leads to additional, undesired components in the ego-noise estimate. In contrast, this effect is not or only weakly pronounced for the proposed method.

The effect of varying λ_E and ϵ on the suppression result is illustrated in Fig. 8. For $\lambda_\mathrm{E} = 0$, the regularization is ineffective and the proposed method reduces to audio only-based NMF ($\lambda_\mathrm{T} = 0$ holds during the learning of $\boldsymbol{D}$). If λ_E is chosen too large, the effect of the motor data dominates and the suppression performance degrades. $\lambda_\mathrm{E} = 19$ appears to result in the best result for the considered mixture. However, note that the optimal choice of λ_E depends on the SNR of the mixture, as will be discussed in more detail in Section 3.3.2. We now consider the suppression performance for varying scale parameter ϵ, c.f. Eq. 6. For $\epsilon \to 0$, we obtain according to Eq. 6

$$W_{\ell j} \to 0, \quad \forall \ell, j \in 1, \ldots, L,$$

i.e, all connections in the graph are set to zero. Accordingly, the regularization term in Eq. 10 equals zero and the results of the proposed method and audio only-based NMF coincide. For increasing ϵ, Eq. 6 gets less selective and the number neighbors of a node with large affinity increases. For the setup in Fig. 8, the maximum SDR is obtained for $\epsilon = 5 \cdot 10^{-3}$, which turned out to result in robust performance even for ego-noise of other movements. For larger ϵ, the suppression performance deteriorates since more and more connections between nodes obtain large weights and the discriminative nature of the graph is reduced.

3.3.2 Varying SNRs

So far, we only considered mixtures with constant SNR. In a typical human-robot interaction, the SNR is however

Table 2 Suppression performance achieved by NMF and the proposed method for varying SNRs. For this experiment, ego-noise by movements of Sequence I is used. Parameters used for proposed methods are given in the last rows. For all SNRs, dictionaries of size $K = 20, K_\mathrm{S} = 20$ are used

SNR	SDR [dB]		SAR [dB]		PESQ		Parameters
(PESQ)	NMF	Proposed	NMF	Proposed	NMF	Proposed	
− 10 dB	− 3.95 (± 0.21)	−3.95 (± 0.21)	4.53 (± 0.26)	4.53 (± 0.26)	1.09 (± 0.00)	1.09 (± 0.00)	$\lambda_\mathrm{T} = 0, \lambda_\mathrm{E} = 0$
(1.08)							$\epsilon = 5 \cdot 10^{-3}$
− 5 dB	2.76 (± 0.43)	2.82 (± 0.33)	6.01 (± 0.04)	6.52 (± 0.06)	1.19 (± 0.01)	1.22 (± 0.00)	$\lambda_\mathrm{T} = 0.1, \lambda_\mathrm{E} = 0.1$
(1.11)							$\epsilon = 5 \cdot 10^{-3}$
− 2 dB	5.82 (± 0.03)	6.19 (± 0.04)	7.86 (± 0.03)	8.77 (± 0.02)	1.27 (± 0.01)	1.32 (± 0.01)	$\lambda_\mathrm{T} = 1.1, \lambda_\mathrm{E} = 1.0$
(1.12)							$\epsilon = 5 \cdot 10^{-3}$
− 1 dB	6.24 (± 0.02)	7.04 (± 0.02)	8.31 (± 0.06)	9.49 (± 0.02)	1.28 (± 0.02)	1.37 (± 0.03)	$\lambda_\mathrm{T} = 1.1, \lambda_\mathrm{E} = 1.5$
(− 1)							$\epsilon = 5 \cdot 10^{-3}$
0 dB	6.60 (± 0.03)	7.76 (± 0.04)	8.70 (± 0.08)	10.32 (± 0.01)	1.28 (± 0.01)	1.37 (± 0.02)	$\lambda_\mathrm{T} = 1.0, \lambda_\mathrm{E} = 4.0$
(1.13)							$\epsilon = 5 \cdot 10^{-3}$
1 dB	6.89 (± 0.05)	8.41 (± 0.06)	9.05 (± 0.12)	11.12 (± 0.01)	1.3 (± 0.01)	1.4 (± 0.02)	$\lambda_\mathrm{T} = 1.0, \lambda_\mathrm{E} = 8.4$
(1.14)							$\epsilon = 5 \cdot 10^{-3}$
2 dB	7.15 (± 0.08)	9.15 (± 0.04)	9.36 (± 0.14)	11.84 (± 0.02)	1.31 (± 0.03)	1.41 (± 0.02)	$\lambda_\mathrm{T} = 1.0, \lambda_\mathrm{E} = 11.0$
(1.15)							$\epsilon = 5 \cdot 10^{-3}$
5 dB	7.58 (± 0.09)	10.92 (± 0.1)	9.96 (± 0.2)	13.88 (± 0.04)	1.35 (± 0.02)	1.45 (± 0.01)	$\lambda_\mathrm{T} = 1.1, \lambda_\mathrm{E} = 16.8$
(1.21)							$\epsilon = 5 \cdot 10^{-3}$
10 dB	8.17 (± 0.24)	13.59 (± 0.03)	10.66 (± 0.41)	16.29 (± 0.10)	1.41 (± 0.00)	1.54 (± 0.00)	$\lambda_\mathrm{T} = 1.1, \lambda_\mathrm{E} = 20.1$
(1.37)							$\epsilon = 5 \cdot 10^{-3}$

changing due to, e.g., varying distances between desired source and robot or different power levels of the signal of interest. Therefore, a robust ego-noise suppression at different SNRs is of high importance.

In the following, we evaluate the proposed approach for SNR $\in \{\pm 10, \pm 5, \pm 2, \pm 1, 0\}$ dB of the input mixture. For this, we added scaled versions of the speech signal to the ego-noise. Note that for the considered NAO robot SNR=10 dB is an unlikely scenario since it corresponds to an human-robot distance of only a couple of centimeters or a very loud human voice. We acknowledge, however, that for robots which emit less loud ego-noise, such a high SNR could be realistic.

Results for ego-noise Sequence I are given in Table 2. In the right part of Table 2, parameters λ_T, λ_E, and ϵ used for the proposed method are summarized. Interestingly, for SNR=− 10 dB, the proposed method shows best result if the regularization is ineffective. Consequently, it does not show any benefit compared to audio only-based NMF. For larger SNRs, SDR and SAR increase both for audio only-based NMF and the proposed method. Motor data-regularized NMF consistently shows superior performance, while the relative improvement between the proposed approach and NMF increases for growing SNR, e.g., for SNR=+ 2 dB, a gain of + 2 dB in SDR and 2.5 dB in SAR is achieved. This effect can be explained by the fact that for audio only-based NMF, the estimation of the activations is more severely impaired by the additional speech signal. This effect becomes more pronounced for increasing SNR. Since motor data is a non-acoustic reference signal, the regularization term is not affected by the increasing power of the additional speech component. This also explains why almost no benefit could be observed for low SNR when the additional speech signal does not have an impact on the ego-noise estimation.

While ϵ is constant for all SNRs, especially λ_E has to be increased continuously for larger SNRs. By this, the influence of the motor data-dependent regularization gets more aggressive compensating the increasingly negative impact of the speech on the estimation of the ego-noise activations.

Table 3 Suppression performance achieved by NMF and the proposed method for varying SNRs. For this experiment, ego-noise by movements of Sequence II is used. Parameters used for proposed methods are given in the last rows. For all SNRs, dictionaries of size $K = 30, K_S = 20$ are used

SNR	SDR [dB]		SAR [dB]		PESQ		Parameters
(PESQ)	NMF	Proposed	NMF	Proposed	NMF	Proposed	
− 10 dB	− 2.59 (± 0.09)	− 2.59 (± 0.09)	3.48 (± 0.05)	3.48 (± 0.05)	1.16 (± 0.02)	1.16 (± 0.02)	$\lambda_T = 0, \lambda_E = 0$
(1.14)							$\epsilon = 5 \cdot 10^{-3}$
− 5 dB	2.33 (± 0.04)	2.45 (± 0.05)	5.40 (± 0.01)	5.41 (± 0.02)	1.19 (± 0.00)	1.19 (±0 .00)	$\lambda_T = 0.1, \lambda_E = 0.5$
(1.15)							$\epsilon = 5 \cdot 10^{-3}$
− 2 dB	4.82 (± 0.02)	5.36 (± 0.03)	6.77 (± 0.01)	7.64 (± 0.04)	1.22 (± 0.01)	1.23 (± 0.01)	$\lambda_T = 1, \lambda_E = 1$
(1.15)							$\epsilon = 5 \cdot 10^{-3}$
− 1 dB	5.27 (± 0.02)	6.16 (± 0.03)	7.15 (± 0.02)	8.25 (± 0.04)	1.25 (± 0.04)	1.28 (± 0.02)	$\lambda_T = 1.5, \lambda_E = 1.5$
(1.15)							$\epsilon = 5 \cdot 10^{-3}$
0 dB	5.60 (± 0.03)	6.82 (± 0.04)	7.52 (± 0.08)	8.75 (± 0.07)	1.26 (± 0.01)	1.31 (± 0.01)	$\lambda_T = 0.9, \lambda_E = 2.0$
(1.16)							$\epsilon = 5 \cdot 10^{-3}$
1 dB	5.82 (± 0.01)	7.45 (± 0.03)	7.8 (± 0.01)	9.48 (± 0.08)	1.26 (± 0.03)	1.37 (± 0.02)	$\lambda_T = 1.0, \lambda_E = 4.0$
(1.16)							$\epsilon = 5 \cdot 10^{-3}$
2 dB	5.89 (± 0.01)	8.05 (± 0.01)	6.12 (± 0.02)	10.38 (± 0.06)	1.27 (± 0.01)	1.37 (± 0.01)	$\lambda_T = 3.0, \lambda_E = 9.1$
(1.18)							$\epsilon = 5 \cdot 10^{-3}$
5 dB	6.37 (± 0.03)	10.00 (± 0.04)	8.6 (± 0.03)	12.26 (± 0.18)	1.30 (± 0.02)	1.39 (± 0.00)	$\lambda_T = 2.1, \lambda_E = 14.9$
(1.24)							$\epsilon = 5 \cdot 10^{-3}$
10 dB	6.69 (± 0.03)	12.20 (± 0.15)	9.06 (± 0.04)	14.09 (± 0.21)	1.40 (± 0.03)	1.48 (± 0.00)	$\lambda_T = 2.5, \lambda_E = 19.5$
(1.35)							$\epsilon = 5 \cdot 10^{-3}$

Table 4 Suppression performance achieved by NMF and the proposed method for varying SNRs. For this experiment, ego-noise by movements of Sequence III is used. Parameters used for proposed methods are given in the last rows. For all SNRs, dictionaries of size $K = 30, K_S = 20$ are used

SNR	SDR [dB]		SAR [dB]		PESQ		Parameters
(PESQ)	NMF	Proposed	NMF	Proposed	NMF	Proposed	
− 10 dB (1.05)	− 2.34 (± 0.75)	− 2.34 (± 0.75)	3.46 (± 0.09)	3.46 (± 0.09)	1.07 (± 0.00)	1.07 (± 0.00)	$\lambda_T = 0, \lambda_E = 0$ $\epsilon = 5 \cdot 10^{-3}$
− 5 dB (1.05)	3.7 (± 0.29)	3.91 (± 0.17)	6.00 (± 0.05)	6.14 (± 0.1)	1.11 (± 0.05)	1.13 (± 0.01)	$\lambda_T = 0.9, \lambda_E = 0.9$ $\epsilon = 5 \cdot 10^{-3}$
− 2 dB (1.06)	5.53 (± 0.02)	5.81 (± 0.02)	7.46 (± 0.04)	8.06 (± 0.03)	1.2 (± 0.01)	1.25 (± 0.01)	$\lambda_T = 0.6, \lambda_E = 2.5$ $\epsilon = 5 \cdot 10^{-3}$
− 1 dB (1.06)	5.88 (± 0.03)	6.40 (± 0.02)	7.88 (± 0.04)	8.65 (± 0.05)	1.25 (± 0.01)	1.3 (± 0.01)	$\lambda_T = 0.3, \lambda_E = 4.4$ $\epsilon = 5 \cdot 10^{-3}$
0 dB (1.07)	6.08 (± 0.02)	6.97 (± 0.04)	8.14 (± 0.03)	9.26 (± 0.09)	1.26 (± 0.01)	1.32 (± 0.01)	$\lambda_T = 0.4, \lambda_E = 6.9$ $\epsilon = 5 \cdot 10^{-3}$
1 dB (1.08)	6.34 (± 0.03)	7.46 (±0.01)	8.48 (± 0.03)	9.78 (± 0.03)	1.29 (± 0.04)	1.38 (± 0.02)	$\lambda_T = 0.5, \lambda_E = 8.0$ $\epsilon = 5 \cdot 10^{-3}$
2 dB (1.09)	6.55 (± 0.05)	7.98 (± 0.01)	8.76 (± 0.04)	10.38 (± 0.04)	1.31 (± 0.03)	1.39 (± 0.01)	$\lambda_T = 0.5, \lambda_E = 12.5$ $\epsilon = 5 \cdot 10^{-3}$
5 dB (1.12)	6.99 (±0.06)	9.10 (± 0.02)	9.38 (± 0.05)	11.59 (± 0.05)	1.35 (± 0.01)	1.41 (± 0.01)	$\lambda_T = 0.5, \lambda_E = 22.5$ $\epsilon = 5 \cdot 10^{-3}$
10 dB (1.23)	7.73 (± 0.1)	10.5 (± 0.05)	9.95 (± 0.04)	13.2 (± 0.06)	1.42 (± 0.0)	1.51 (± 0.02)	$\lambda_T = 0.4, \lambda_E = 29.5$ $\epsilon = 5 \cdot 10^{-3}$

Table 5 Performance degradation of proposed method for suboptimal parameter settings. We chose $\lambda_T = 0.5 = \text{const.}$ and $\lambda_E = 8.0 = \text{const.}$ which are optimal for SNR= 1 dB and evaluated SNRs − 1 dB, . . . , 3 dB. For each performance value of the proposed method, the degradation relative to optimum parameter settings for λ_E and λ_T is given in parentheses. For this experiment, ego-noise by movements of Sequence III is used

		SNR=− 1 dB PESQ=1.06	SNR=0 dB PESQ=1.07	SNR=+ 1 dB PESQ=1.08	SNR=+ 2 dB PESQ=1.09	SNR=+ 3 dB PESQ=1.1
NMF	SDR [dB]	5.88	6.08	6.34	6.55	6.72
	SAR [dB]	7.88	8.14	8.48	8.76	9.00
	PESQ	1.25	1.26	1.29	1.31	1.33
Proposed	SDR [dB]	6.25 (− 0.15)	6.87 (− 0.1)	7.46 (± 0.0)	7.87 (− 0.11)	8.19 (− 0.22)
	SAR [dB]	8.41 (− 0.24)	9.36 (− 0.1)	9.78 (± 0.0)	10.19 (− 0.19)	10.51 (− 0.38)
	PESQ	1.26 (− 0.04)	1.30 (− 0.02)	1.38 (± 0.0)	1.37 (− 0.02)	1.39 (− 0.02)

Table 6 Performance for different designs of $\bar{\boldsymbol{\alpha}}_\ell$: if no derivatives are considered, $\bar{\boldsymbol{\alpha}}_\ell$ is composed of angular positions only. If one derivative is considered, $\bar{\boldsymbol{\alpha}}_\ell$ contains angular positions and their first order temporal derivatives. For this experiment, ego-noise caused by movements of Sequence II is used, SNR=0 dB, $K = 30$, $K_S = 20$, $\epsilon = 5 \cdot 10^{-3}$, $\lambda_T = 0.9$, and $\lambda_E = 2$

	NMF	Proposed				
# of derivatives of $\bar{\boldsymbol{\alpha}}_\ell$		0	1	2	3	4
SDR [dB]	5.60 ($\pm$ 0.03)	3.87 ($\pm$ 0.08)	6.63 ($\pm$ 0.06)	6.82 ($\pm$ 0.04)	6.68 ($\pm$ 0.04)	6.35 ($\pm$ 0.05)
SAR [dB]	7.52 ($\pm$ 0.08)	5.46 ($\pm$ 0.07)	8.3 ($\pm$ 0.02)	8.75 ($\pm$ 0.07)	8.60 ($\pm$ 0.06)	8.26 ($\pm$ 0.06)
PESQ	1.23 ($\pm$ 0.01)	1.18 ($\pm$ 0.02)	1.29 ($\pm$ 0.03)	1.31 ($\pm$ 0.01)	1.3 ($\pm$ 0.02)	1.29 ($\pm$ 0.01)

We conducted the same experiments for ego-noise caused by motions of Sequence II and Sequence III. The results are summarized in Tables 3 and 4. In principle, the results obtained for Sequence I can be confirmed: the proposed method outperforms audio only-based NMF consistently, especially for high SNRs, and λ_E shows a significant dependence on the SNR. Interestingly for Sequence II, the absolute values for λ_E have to be chosen slightly smaller for optimum performance than for Sequence I. Overall, the suppression results are $\approx$ 1 dB (Sequence II) and $\approx$ 0.5 dB (Sequence III) worse than for Sequence I, which can be explained by the more complex movements, c.f., movement description in Section 3.2.

Note that for an optimal choice of λ_E knowledge of the SNR is required for which a single-channel or multichannel SNR estimator can be employed. However, it must be expected that the SNR estimation is imperfect leading to a suboptimal choice of λ_E. The resulting effect on the suppression performance is shown in Table 5. We chose $\lambda_E = 8.0$ and $\lambda_T = 0.5$, i.e, optimal parameters for SNR= 1 dB, and evaluated the proposed method for SNRs -1 dB, ..., 3 dB, simulating imperfect SNR estimates. Overall, a suboptimal parameter choice leads to a degradation of the suppression performance. However, the proposed method still shows superior results compared to audio only-based NMF.

3.3.3 *Alternative choices for $\bar{\boldsymbol{\alpha}}_\ell$*

In the previous experiments, the motor data vector $\bar{\boldsymbol{\alpha}}_\ell$ was composed of the angular position and its first- and second-order temporal derivatives, i.e., angular velocity and acceleration. By complementing angular position by its first and second order derivatives, we implicitly added temporal information to our model since not only current, but also past motor data samples are taken into account for the construction of $\bar{\boldsymbol{\alpha}}_\ell$, c.f. Eq. 1.

In the following, we evaluate how the performance of the proposed method depends on the amount of temporal context included into $\bar{\boldsymbol{\alpha}}_\ell$.

Results are given in Table 6. First, we consider a motor data vector $\bar{\boldsymbol{\alpha}}_\ell$ which contains only angular positions, i.e., no derivatives are used. The suppression result lags significantly behind audio only-based NMF. This drop in performance is not surprising since by considering angular position alone it cannot be distinguished whether, e.g., the robot raises or drops its arm if up- and downwards movements have the same trajectory. Consequently, the ego-noise caused by these two movements is assessed as similar. The results improve drastically if angular velocity is added to motor data vector $\bar{\boldsymbol{\alpha}}_\ell$. If also angular acceleration is included, the results further improve; however, the additional gain is clearly smaller compared to that of adding the first derivative. Adding higher order derivatives to $\bar{\boldsymbol{\alpha}}_\ell$ does not offer further benefit as results get slightly worse with increasing number of derivatives incorporated to $\bar{\boldsymbol{\alpha}}_\ell$.

4 Summary and outlook

In this paper, we proposed motor data-regularized NMF and used it in a semi-supervised manner for ego-noise suppression.

The basic idea of the presented method is to improve the approximation of the ego-noise by taking motor data describing the physical state of the robot into account. We propose to construct a motor data graph which encodes the similarities between motor data samples. Based on this, a regularization term can be derived and added to the conventional, audio only-based NMF cost function. It enforces the activation of similar dictionary entries when the robot is in similar physical states. We evaluated the proposed method for mixtures of desired speech signals and ego-noise of different movements and considered various SNRs of the mixture. The presented approach showed superior performance in all scenarios, especially for high SNR when the power of the additional speech signal is large and the estimation of the ego-noise activations based on audio data-only is challenging. Consequently, the weighting of the motor data-dependent regularization term has to be increased for larger SNR.

For future work, we plan to evaluate the proposed method for other NMF cost functions, such as Itakuro-Saito and Kullback divergence. Furthermore, we plan to evaluate the presented concept for multichannel NMF, where dictionary-activation-modeling of single-channel NMF is extended by a spatial covariance matrix for each atom and frequency bin.

Authors' contributions
AS has conducted the research on this paper. AB, TH, and WK contributed valuable feedback on the conceptual idea and assisted the work intensively. All authors read and approved the final manuscript.

Funding
This work was partially supported by the DFG under contract no <Ke890/10-2> within the Research Unit FOR2457 "Acoustic Sensor Networks".

Competing interests
The authors declare that they have no competing interests.

Received: 8 January 2020 Accepted: 10 June 2020
Published online: 31 July 2020

References

1. K. Nakadai, T. Lourens, H. G. Okuno, H. Kitano, in *Proc. 17th Nat. Conf. Artificial Intell. (AAAI)*. Active audition for humanoid (AAAI, Austin, TX, 2000), pp. 832–839
2. H. G. Okuno, K. Nakadai, in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*. Robot audition: its rise and perspectives (IEEE, South Brisbane, QL, Australia, 2015), pp. 5610–5614
3. J. Even, H. Saruwatari, K. Shikano, T. Takatani, in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*. Semi-blind suppression of internal noise for hands-free robot spoken dialog system (IEEE, St. Louis, MO, 2009), pp. 658–663
4. M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process. **54**(11), 4311–4322 (2006)
5. A. Deleforge, W. Kellermann, in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*. Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures (IEEE, South Brisbane, QL, Australia, 2015), pp. 355–359
6. D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization. Nature. **401**(6755), 788–791 (1999)
7. D. D. Lee, H. S. Seung, in *Proc. 13th Int. Conf. Neural Inform. Process. Syst. (NIPS)*. Algorithms for non-negative matrix factorization (NeurIPS, Denver, CO, 2000), pp. 535–541
8. C. Févotte, J. Idier, Algorithms for non-negative matrix factorization with the β-divergence. Neural Comput. **23**(9), 2421–2456 (2011)
9. T. Tezuka, T. Yoshida, K. Nakadai, in *Proc. IEEE Int, Conf. Robotics and Automation (ICRA)*. Ego-motion noise suppression for robots based on semi-blind infinite non-negative matrix factorization (IEEE, Florence, Italy, 2014), pp. 6293–6298
10. H. Sawada, H. Kameoka, S. Araki, N. Ueda, Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE/ACM Trans. Audio, Speech, Language Process. **21**(5), 971–982 (2013)
11. T. Haubner, A. Schmidt, W. Kellermann, in *Proc. ITG Fachtagung Sprachkommunikation*. Multichannel nonnegative matrix factorization for ego-noise suppression (VDE-Verlag, Oldenburg, Germany, 2018), pp. 136–140
12. Clean PNG, NAO, der humanoide Roboter. https://de.cleanpng.com/png-m5r7ur/ Accessed 20 May 2020
13. A. Ito, T. Kanayama, M. Suzuki, S. Makino, in *Proc. European Conf. Speech Communication and Technology (INTERSPEECH - Eurospeech)*. Internal noise suppression for speech recognition by small robots (ISCA, Lisbon, Portugal, 2005), pp. 2685–2688
14. A. Schmidt, W. Kellermann, in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*. Informed ego-noise suppression using motor data-driven dictionaries (IEEE, Brighton, UK, 2019), pp. 116–120
15. Y. Nishimura, M. Ishizuka, K. Nakadai, M. Nakano, H. Tsujino, in *Proc. IEEE/RAS Int, Conf. Humanoid Robots (Humanoids)*. Speech recognition for a humanoid with motor noise utilizing missing feature theory (IEEE, Cancun, Mexico, 2006), pp. 26–33
16. G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, J. Imura, in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*. Ego-noise suppression of a robot using template subtraction (IEEE, St. Louis, MO, 2009), pp. 199–204
17. G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, in *Proc. IEEE Int, Conf. Robotics and Automation (ICRA)*. Imura: A hybrid framework for ego noise cancellation of a robot (IEEE, Anchorage, AK, 2010), pp. 3623–3628
18. A. Schmidt, A. Deleforge, W. Kellermann, in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*. Ego-noise reduction using a motor data-guided multichannel dictionary (IEEE, Daejon, South Korea, 2016), pp. 1281–1286
19. D. Cai, X. He, X. Wu, J. Han, in *Proc. 8th IEEE Int, Conf. on Data Mining*. Non-negative matrix factorization on manifold (IEEE, Pisa, Italy, 2008), pp. 63–72
20. D. Cai, X. He, J. Han, T. S. Huang, Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. Pattern Anal. and Mach. Intell. **33**(8), 1548–1560 (2011)
21. M. N. Schmidt, J. Larsen, F.-T. Hsiao, in *Proc. IEEE Workshop Mach. Learning Signal Process*. Wind noise reduction using non-negative sparse coding (IEEE, Thessaloniki, Greece, 2007), pp. 431–436
22. U. von Luxburg, A tutorial on spectral clustering. Statistics and Computing. **17**(4), 395–416 (2007)
23. F. R. K. Chung, *Spectral graph theory*, 1st edn, vol. 1. (American Mathematical Soc., Providence, RI, 1997)
24. M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and uUnlabeled examples. J. Mach. Learn. Research. **7**, 2399–2434 (2006)
25. M. Belkin, *Problems of learning on manifolds. PhD Thesis*. (The University of Chicago, Chicago, 2003)
26. Seventh Framework Programme, 'Embodied Audition for RobotS' (EARS). https://robot-ears.eu/. Accessed 25 Sept 2018
27. M. Cooke, J. Barker, An audio-visual corpus for speech perception and automatic speech recognition. J. Acoustical Society of America. **120**(5), 2421–2424 (2006)
28. C. Févotte, R. Griboval, E. Vincent, in *Technical Report 1706*. BSS EVAL toolbox user guide (IRISA, Rennes, France, 2005). Software available at http://www.irisa.fr/metiss/bsseval/
29. ITU-T Recommendation P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs. Recommendation, ITU (November 2007)

Publisher's Note