INFORMED EGO-NOISE SUPPRESSION USING MOTOR DATA-DRIVEN DICTIONARIES

Alexander Schmidt and Walter Kellermann

Multimedia Communications and Signal Processing,
Friedrich-Alexander University Erlangen-Nürnberg,
Cauerstr. 7, 91058 Erlangen, Germany,
{alexander.as.schmidt, walter.kellermann}@fau.de

ABSTRACT

The suppression of ego-noise for (humanoid) robots is typically addressed by learning-based techniques. In this paper, we propose a novel approach which models significant parts of ego-noise spectrograms based on motor data and does not require a prior training step. Accordingly, the intrinsic harmonic structure of ego noise is taken into account by introducing a nonnegative matrix factorization (NMF) framework with motor data-driven dictionaries. Limited improvement was observed by employing an additional pre-trained small-sized dictionary accounting for the residual ego-noise. The presented approach exhibits comparable suppression performance to an audio only-based approach trained specifically to the scenario, while the number of dictionary elements which require prior learning can be reduced by a factor of two. For ego-noise resulting from previously unseen movements, the proposed method shows consistently superior suppression results while the audio only-based approach degrades drastically.

Index Terms— Ego-noise suppression, robot, motor data, dictionary learning, nonnegative matrix factorization

1. INTRODUCTION

Microphone-equipped autonomous systems are exposed to various kinds of noise, specifically to so-called *ego-noise*, i.e., self-created noise. In this paper, we consider ego-noise of a humanoid robot which is caused by its electrical and mechanical components, e.g., motors and joints. Ego-noise corrupts the recorded microphone signals and impairs the robot's capability to react autonomously to unanticipated acoustic events. This motivates techniques for ego-noise suppression, which is a key pre-processing step in robot audition [1, 2, 3]. Since the robot performs movements with varying speeds and accelerations, ego-noise is highly non-stationary, but exhibits characteristic structure in the Short-Time Fourier Transform (STFT) domain. Due to the limited number of degrees of freedom of the robot, those spectral patterns cannot be arbitrarily diverse. These two facts motivate the use of learning-based approaches, e.g., *nonnegative matrix factorization* (NMF) [4, 5] or other dictionary-based methods such as *K-SVD* [6, 7].

Aside from audio data, ego-noise suppression can exploit reference information given by the known internal state of the robot, e.g., the motor state information which is referred to as *motor data* in the following. Exemplary approaches employ spectral subtraction for ego-noise reduction where the required ego-noise power spectral densities (PSDs) are estimated from motor data and used to train a noise model. For this, a data base of ego-noise templates is proposed in [8, 9] while the authors of [10] suggest an implicit ego-noise modeling using a deep neural network.

While previous approaches exploit the use of motor data for ego-noise suppression using entirely training-based schemes, we propose in this paper the use of motor data in a parametric manner and demonstrate its benefit to a pre-trained method which employs audio but no motor data (referred to as *audio only-based* in the following). We propose to explicitly account for the harmonic structure of ego-noise and embed it into an NMF framework for ego-noise suppression. This approach is inspired by harmonic-constraint NMF as it is employed, e.g., for music analysis. In [11, 12], a multiple pitch estimator is presented by assuming that the NMF bases are given by a weighted sum of adjacent harmonics. A similar approach was followed in [13] for tone separation and vibrato modeling, where each dictionary entry represents a mixture of harmonic contributions. In both approaches, the model parameters are estimated blindly from given observations.

In the following, we suggest to model the intrinsic harmonic parts of ego-noise by using a time-varying dictionary determined by instantaneous motor data. We model the residual ego-noise components in a second dictionary that is trained using an NMF framework which takes the information from the time-varying, motor data-driven dictionary into account.

This paper is structured as follows. In Sec. 2.1, we describe the used motor data and their potential to estimate the harmonic ego-noise components. After succinctly introducing NMF, we describe our proposed ego-noise modeling ap-

proach to construct a motor data-driven dictionary in Sec. 2.2. This is then combined with an additional dictionary modeling the residual noise to jointly suppress ego-noise in Sec. 2.3. The efficacy of the method is demonstrated in Sec. 3.

2. MOTOR DATA-INFORMED EGO-NOISE SUPPRESSION

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

2.1. Harmonic structure estimation

A rotating engine typically produces noise with a harmonic, deterministic structure in the spectrogram where the harmonics appear at multiples of half of the rotation frequency [14]. While this rotation frequency is usually not directly observable for motors in a moving robot, it can be derived using angular position observations collected by proprioceptors mounted to each joint of the robot. We denote the l-th observed angular position in time frame t as $\alpha_t^{(l)}$ for a given proprioceptor and approximate angular speed by

$$\dot{\alpha}_t^{(l)} = \frac{1}{\Delta T_{tl}} \cdot \left(\alpha_t^{(l)} - \alpha_t^{(l-1)} \right), \tag{1}$$

where ΔT_{tl} denotes the time difference between adjacent observations $\alpha_t^{(l)}$ and $\alpha_t^{(l-1)}$. The frequency of the i-th harmonic of the considered motor is then given by

$$i \cdot f_{0,t}^{(l)} = \tfrac{1}{2} \cdot i \cdot \dot{\alpha}_t^{(l)} \cdot \gamma, \tag{2}$$

where γ is the so-called velocity-reduction-ratio that takes the mechanical translation between motor and joint into account. The left plot in Fig. 1 shows a spectrogram of a typical arm movement of the robot NAO™ (see Sec. 3, [15]) with five harmonics and its estimated versions. Obviously, however, there are residual ego-noise parts that need to be addressed separately.

2.2. Ego-noise Modeling

We propose to model ego-noise by

$$\boldsymbol{Y} = \boldsymbol{Y}_\mathrm{P} + \boldsymbol{Y}_\mathrm{R}, \tag{3}$$

where $\boldsymbol{Y}_\mathrm{P}$ denotes the part of the spectrogram that can be estimated from motor data, while $\boldsymbol{Y}_\mathrm{R}$ models the residual ego-noise. In earlier work [16], we extracted $\boldsymbol{Y}_\mathrm{P}$ and $\boldsymbol{Y}_\mathrm{R}$ using the estimated harmonic position and modeled both parts entirely *learning-based*. In this paper, we propose a novel and more efficient NMF-based motor data-driven model for $\boldsymbol{Y}_\mathrm{P}$ that requires *no* prior learning. To account for the residual $\boldsymbol{Y}_\mathrm{R}$, we propose a second NMF-based dictionary which requires training, but is of low model complexity. Fig. 2 illustrates the proposed ego-noise modeling approach.

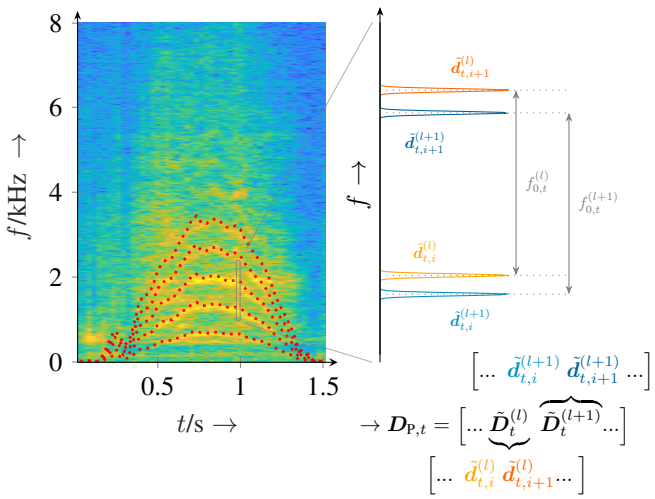

Fig. 1. **Left:** Spectrogram of an ego-noise recording for a right arm movement of the humanoid robot NAO™. Harmonic components can be estimated (red) using Eq. 2 with $i = 1, \dots, 5$. **Right:** Illustration of the proposed motor data-based modeling of $\boldsymbol{D}_{\mathrm{P},t}$ for an excerpt of the frequency range and time frame t. $L = 2$ motor data samples per time frame are considered.

2.2.1. *Nonnegative Matrix Factorization (NMF)*

The objective of NMF is to approximate the nonnegative matrix $\boldsymbol{Y}$ by a product of two nonnegative matrices

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

where $\boldsymbol{D} \in \mathbb{R}_+^{F \times K}$ is the so-called dictionary of size $F \times K$ and $\boldsymbol{H} = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_T] \in \mathbb{R}_+^{K \times T}$ is referred to as activation matrix [17, 18]. 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 $\boldsymbol{H}$. The factorization is achieved by minimizing a cost function which measures the similarity between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ with respect to the model parameters. In this paper, we consider the Itakura-Saito (IS) divergence as cost function, which is common and well suited for audio applications since it depends only on the power ratios between the true and approximated signal [4]. $\boldsymbol{D}$ and $\boldsymbol{H}$ are typically obtained using iterative update rules that can be derived using, e.g., Majorization-Minimization algorithms or heuristic approaches [17, 19].

2.2.2. *Modeling of $\boldsymbol{Y}_P$*

To model the harmonic structure of ego-noise explicitly in each time frame, we propose to estimate $\boldsymbol{Y}_\mathrm{P}$ from Eq. 3 by

$$\hat{\boldsymbol{Y}}_\mathrm{P} = [\boldsymbol{D}_{\mathrm{P},1}\boldsymbol{h}_{\mathrm{P},1}, \dots, \boldsymbol{D}_{\mathrm{P},T}\boldsymbol{h}_{\mathrm{P},T}], \tag{5}$$

with time-varying dictionary $\boldsymbol{D}_{\mathrm{P},t}$ of size $F \times K_\mathrm{P}$ and vector $\boldsymbol{h}_{\mathrm{P},t}$ containing the K_P activation weights for time frame

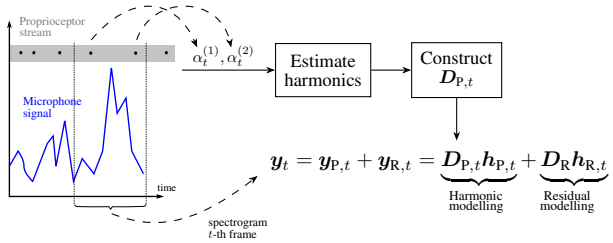

Fig. 2. Illustration of the proposed NMF-based ego-noise modeling approach for time frame t. Motor data is used to estimate the harmonic structure of ego-noise, which is then employed to construct the time-varying dictionary $\boldsymbol{D}_{\mathrm{P},t}$.

t. Note that a formulation as a single matrix-matrix multiplication according to Eq. 4 is not possible here, since $\boldsymbol{D}_{\mathrm{P},t}$ is generally different for every t.

For the construction of $\boldsymbol{D}_{\mathrm{P},t}$, we assume that a proprioceptor collects L motor data samples $\alpha_t^{(l)}, l = 1, \ldots, L$, per time frame t. We can then approximate L angular speed samples $\dot{\alpha}_t^{(l)}$ per time frame using Eq. 1, where each sample contributes to each harmonic i according to Eq. 2. For each harmonic i, the set of L samples actually forms for each harmonic i a group of L spectral components which only overlap if the temporal variation of $\dot{\alpha}_t^{(l)}$ is small. This is illustrated on the right-hand side of Fig. 1 for $L = 2$ and two harmonics i, $i+1$. We propose to model the i-th harmonic component in a spectrogram frame $\boldsymbol{y}_t$ by a set of L atoms $\tilde{\boldsymbol{d}}_{t,i}^{(l)}, l = 1, \ldots, L$. Here, $\tilde{\boldsymbol{d}}_{t,i}^{(l)}$ describes the spectral contribution of $\dot{\alpha}_t^{(l)}$ and we suggest to model it as a bell-shaped function centered around $i \cdot f_{0,t}^{(l)}$. The f-th component of $\tilde{\boldsymbol{d}}_{t,i}^{(l)}$ is then given by

$$\tilde{d}_{ft,i}^{(l)} = \exp\left(-\left(f - i \cdot f_{0,t}^{(l)}\right)^2 / 2w\right), \tag{6}$$

where w controls the width of the harmonic contribution and reflects the temporal variation of $\dot{\alpha}_t^{(l)}$. Based on this, we exploit the atoms $\tilde{\boldsymbol{d}}_{t,i}^{(l)}, l = 1, \ldots, L$, to define the sub-dictionary $\tilde{\boldsymbol{D}}_{t,i} = \left[\tilde{\boldsymbol{d}}_{t,i}^{(1)} \;\; \ldots \;\; \tilde{\boldsymbol{d}}_{t,i}^{(L)}\right]$. Assuming I harmonics in a spectrogram frame $\boldsymbol{y}_t$, we finally obtain

$$\boldsymbol{D}_{\mathrm{P},t} = \left[\tilde{\boldsymbol{D}}_{t,1} \;\; \ldots \;\; \tilde{\boldsymbol{D}}_{t,I}\right]. \tag{7}$$

The relative scaling of $\tilde{\boldsymbol{d}}_{t,i}^{(l)}$ for different i is data-dependent and determined implicitly by adapting the activation matrix $\boldsymbol{H}_{\mathrm{P}} = [\boldsymbol{h}_{\mathrm{P},1}, \ldots, \boldsymbol{h}_{\mathrm{P},T}] \in \mathbb{R}_+^{K_{\mathrm{P}} \times T}$ using the NMF update rules (c.f. Sec. 2.3).

2.2.3. *Modeling of $\boldsymbol{Y}_R$*

For $\boldsymbol{Y}_{\mathrm{R}}$, we propose a conventional NMF-based model according to Eq. 4

$$\hat{\boldsymbol{Y}}_{\mathrm{R}} = \boldsymbol{D}_{\mathrm{R}}\boldsymbol{H}_{\mathrm{R}}, \tag{8}$$

with dictionary $\boldsymbol{D}_{\mathrm{R}}$ incorporating K_{R} atoms. $\boldsymbol{D}_{\mathrm{R}}$ needs a prior learning step, however has to capture only the residual part of ego-noise and can be therefore of small size.

2.3. Proposed Algorithm for Ego-noise Suppression

We propose a two-stage algorithm for ego-noise suppression. First, we use audio data containing ego-noise only and employ $\boldsymbol{D}_{\mathrm{P},t}$ to model the harmonic components (see Subsec. 2.3.1) while concurrently training $\boldsymbol{D}_{\mathrm{R}}$ on the residual part of the ego-noise. Given a mixture of ego-noise and speech, we employ $\boldsymbol{D}_{\mathrm{P},t}$ and $\boldsymbol{D}_{\mathrm{R}}$ to model and suppress current ego-noise and to obtain a speech estimate (see Subsec. 2.3.2).

2.3.1. *Learning $\boldsymbol{D}_R$*

As input, spectrograms $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$ are given containing ego-noise only. Per spectrogram sample, motor speed $\dot{\alpha}_t^{(l)}$ is present which is used to construct $\boldsymbol{D}_{\mathrm{P},t}$ using Eqs. 2, 6 and 7. Using $\boldsymbol{Y}$, we employ the concept presented in [18] and update $\boldsymbol{D}_{\mathrm{R}}$, $\boldsymbol{H}_{\mathrm{R}}$ and $\boldsymbol{H}_{\mathrm{P}} = [\boldsymbol{h}_{\mathrm{P},1}, \ldots, \boldsymbol{h}_{\mathrm{P},T}]$ using update rules proposed in [17]. The set of t-dependent dictionaries $\boldsymbol{D}_{\mathrm{P},t}$ remains constant during multiple iterations of optimizing $\boldsymbol{D}_{\mathrm{R}}$.

2.3.2. *Ego-noise suppression*

Another dictionary $\boldsymbol{D}_{\mathrm{S}}$ with activation $\boldsymbol{H}_{\mathrm{S}}$ is initialized to model the additional speech signal in the considered mixture. Using the same update rules as in the previous learning step, $\boldsymbol{D}_{\mathrm{S}}$, $\boldsymbol{H}_{\mathrm{P}}$, $\boldsymbol{H}_{\mathrm{R}}$ and $\boldsymbol{H}_{\mathrm{S}}$ are updated while the set of t-dependent dictionaries $\boldsymbol{D}_{\mathrm{P},t}$ and $\boldsymbol{D}_{\mathrm{R}}$ remain constant. After identifying the optimum model parameters captured by $\boldsymbol{D}_{\mathrm{S}}$, $\boldsymbol{H}_{\mathrm{P}}$, $\boldsymbol{H}_{\mathrm{R}}$ and $\boldsymbol{H}_{\mathrm{S}}$, we employ a spectral enhancement filter to obtain an estimate of the desired speech signal $\hat{\boldsymbol{Y}}_{ft}^{(\mathrm{speech})} = \boldsymbol{W}_{ft}^{(\mathrm{speech})} \odot \boldsymbol{Y}_{ft}$ for the ft-th bin, where $\odot$ denotes element-wise multiplication and where the enhancement filter matrix $\boldsymbol{W}_{ft}^{(\mathrm{speech})}$ is given by

$$\boldsymbol{W}_{ft}^{(\mathrm{speech})} = \frac{\hat{\boldsymbol{Y}}_{\mathrm{S},ft}}{\hat{\boldsymbol{Y}}_{\mathrm{P},ft} + \hat{\boldsymbol{Y}}_{\mathrm{R},ft} + \hat{\boldsymbol{Y}}_{\mathrm{S},ft}}, \tag{9}$$

with $\hat{\boldsymbol{Y}}_{\mathrm{S}} = \boldsymbol{D}_{\mathrm{S}}\boldsymbol{H}_{\mathrm{S}}$.

3. EVALUATION

3.1. Experimental setup

To evaluate our approach, we conducted experiments with a NAO™ H25 humanoid robot [15]. For audio recordings, we used a modified head developed during the EU FP7 Project EARS [20] with a microphone array of 12 sensors. For all experiments, we used the frontmost microphone.

The measurements were conducted in a room with moderate reverberation ($T_{60} = 200\,\mathrm{ms}$). We recorded ego-noise of right arm movements, including six joints. All movements

Table 1. Performance in dB achieved by the proposed method and audio only-based NMF.

		Unproc.	proposed $K_{\mathrm{R}}=0$	proposed $K_{\mathrm{R}}=5$	NMF $K=5$	proposed $K_{\mathrm{R}}=10$	NMF $K=10$	proposed $K_{\mathrm{R}}=20$	NMF $K=20$	proposed $K_{\mathrm{R}}=30$	NMF $K=30$
Scenario I	SDR	-3.0	5.39	6.05	2.41	6.20	5.75	6.31	6.2	6.47	5.76
	SIR	-2.96	9.28	11.42	4.19	11.61	10.64	13.05	15.5	13.0	16.5
	SAR	275.16	7.60	7.8	7.3	7.78	7.6	7.9	7.93	7.86	7.45
Scenario II	SDR	-3.0	5.2	5.44	-0.42	5.56	-0.37	5.81	0.39	5.89	1.2
	SIR	-2.99	8.6	9.91	-1.1	11.1	-0.9	12.1	0.91	12.29	5.21
	SAR	279.15	7.49	7.58	6.3	7.62	6.43	7.81	6.28	7.76	6.68

consist of lifting the arm using the right shoulder pitch motor, while performing waving movements with the remaining five motors of the right arm. The recorded ego-noise was then used for training and testing, where the testing data was not contained in the training data. In total, we recorded $60\,\mathrm{s}$ of ego-noise, the ratio of training to test data was $3:1$. In the following, we investigate two scenarios differing in the structure of the test data. In *Scenario I*, test data contains shoulder pitch speeds similar to some contained in the training data. In contrast, we consider shoulder pitch speeds in *Scenario II* that were not contained in the training data. However, movements of the remaining five joints are similar in both testing and training. For testing, we consider a scenarios in which a target source is talking to the robot while the robot performs different waving movements of the right arm. For the speech, utterances from the GRID corpus [21] were used. The loudspeaker was positioned at $1\,\mathrm{m}$ distance of NAO™, at a height of $1\,\mathrm{m}$. The recorded utterances were added to the movement noise.

The audio signals are sampled at $f_S = 16\,\mathrm{kHz}$ and transformed to the STFT domain using a Hamming window of length $64\,\mathrm{ms}$ with overlap of $50\,\%$. The sampling frequency of the motor data is given by $f_M \approx 100\,\mathrm{Hz}$, i.e., $L \approx 6$ motor data samples are available per time frame. We use exclusively motor data of the right shoulder pitch joint to approximate angular speed and estimate the harmonics. We chose to model $I = 15$ harmonics, where each harmonic contribution has width $w = 2$. Both I and w were found heuristically to result in best performance for *Scenario I*, however both have shown in our experiments to generalize well for other movements. We compare our method to audio only-based NMF. We evaluated both algorithms for different dictionary sizes K_{R} for the proposed method and K for NMF, i.e., we modify the size of those dictionaries of both approaches that require a prior training step. During testing, a speech dictionary with size $K_{\text{speech}} = 20$ for both approaches was chosen.

We evaluated the performance in terms of Signal-to-Inference-Ratio (SIR in dB), Signal-to-Distortion-Ratio (SDR in dB) and Signal-to-Artifacts-Ratio (SAR in dB), using Matlab functions provided by [22]. All results are averaged over 100 runs.

3.2. Discussion of Results

The results are summarized in Table 1. For both scenarios, using only the motor data-driven dictionary ($K_{\mathrm{R}} = 0$) (recall that no prior training is needed for this) shows already a significant gain compared to unprocessed data for all evaluation criteria. For *Scenario I*, the proposed method outperforms audio only-based NMF for small dictionary sizes, since the residual part of ego-noise can be well modeled with a small number of bases. In contrast, an NMF-trained dictionary has to capture both harmonics and residual and therefore requires approximately twice the number of bases to produce comparable performance results. For *Scenario II*, the proposed approach clearly outperforms NMF for all dictionary sizes since the evaluated shoulder pitch speeds were not contained in the training data. As the noise contribution of the remaining arm components was similar in both training and testing, employing the residual dictionary slightly improves the result. For the same reason, audio only-based NMF shows a slight improvement for increasing K, since for larger K the learned dictionary models harmonics and residual in separate bases. However, this discriminative effect is only weakly pronounced.

4. CONCLUSION AND OUTLOOK

In this paper, we presented a semi-supervised NMF-based ego-noise suppression approach where the harmonic ego-noise components were modeled using a time-varying, motor data-driven dictionary which does not require a prior training step. The proposed method shows robust suppression results even for ego-noise that was not seen during training. For future work, we plan to apply the presented concept to multichannel NMF, where the variance modeling is extended by a spatial covariance matrix for each atom and frequency bin, c.f. [5, 23].

5. REFERENCES

[1] K. Nakadai et al., "Active audition for humanoid," in *Proc. 17th Nat. Conf. Artificial Intell. (AAAI)*, July 2000, pp. 832–839.

[2] H. G. Okuno and K. Nakadai, "Robot audition: Its rise and perspectives," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2015, pp. 5610–5614.

[3] S. Argentieri et al., "Binaural systems in robotics," in *Modern Acoustics and Signal Processing*, J. Blauert, Ed., pp. 225–253. Springer, Berlin, Heidelberg, 2013.

[4] C. Févotte et al., "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[5] H. Sawada et al., "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. Audio., Speech, Language Process.*, vol. 21, no. 5, pp. 971–982, 2013.

[6] M. Aharon et al., "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[7] A. Deleforge and W. Kellermann, "Phase-optimized K-SVD for signal extraction from underdetermined multichannel sparse mixtures," in *Proc. of IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2015, pp. 355–359.

[8] G. Ince et al., "Ego noise suppression of a robot using template subtraction," in *Proc. IEEE Int. Conf. Intelligent Robots and Systems (IROS)*, 2009, pp. 199–204.

[9] G. Ince et al., "Assessment of general applicability of ego noise estimation," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, May 2011, pp. 3517–3522.

[10] A. Ito et al., "Internal noise suppression for speech recognition by small robots," in *Proc. European Conf. Speech Communication and Technology (INTERSPEECH - Eurospeech)*, 2005, pp. 2685–2688.

[11] E. Vincent et al., "Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2008, pp. 109–112.

[12] E. Vincent et al., "Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[13] R. Hennequin et al., "Time-dependent parametric and harmonic templates in Non-negative Matrix Factorization," in *Proc. 13th Intern. Conf. Digital Audio Effects (DAFx)*, 2010, pp. 8–13.

[14] H. Puder and F. Steffens, "Improved noise reduction for hands-free car phones utilizing information on vehicle and engine speeds," in *Proc. 10th European Signal Process. Conf. (EUSIPCO)*, Sept. 2000, pp. 1–4.

[15] D. Gouaillier et al., "Mechatronic design of NAO humanoid," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, May 2009, pp. 769–774.

[16] A. Schmidt et al., "A novel ego-noise suppression algorithm for acoustic signal enhancement in autonomous systems," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Mar. 2018.

[17] C. Févotte and J. Idier, "Algorithms for non-negative matrix factorization with the β-divergence," *Neural Comput.*, vol. 23, no. 9, pp. 2421–2456, 2011.

[18] M. N. Schmidt et al., "Wind noise reduction using non-negative sparse coding," in *Proc. IEEE Workshop Mach. Learning Signal Process.*, 2007, pp. 431–436.

[19] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Inform. Process. Syst.*, vol. 13, 2001.

[20] "Seventh framework programme 'Embodied Audition for RobotS' (EARS)," `https://robot-ears.eu/`, Accessed: 2018-09-25.

[21] M. Cooke et al., "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoustical Soc. America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[22] C. Févotte et al., "BSS eval toolbox user guide," Technical Report 1706, IRISA, Rennes, France, April 2005, Software available at `http://www.irisa.fr/metiss/bsseval/`.

[23] D. Kitamura et al., "Determined Blind Source Separation Unifying Independent Vector Analysis and Non-negative Matrix Factorization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 9, pp. 1626–1641, Sept. 2016.