A NOVEL EGO-NOISE SUPPRESSION ALGORITHM FOR ACOUSTIC SIGNAL ENHANCEMENT IN AUTONOMOUS SYSTEMS

Alexander Schmidt, Heinrich W. Löllmann, and Walter Kellermann

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

ABSTRACT

The use of autonomous systems (ASs), such as humanoid robots, drones or self-driving vehicles, has expanded significantly in recent years. For such systems, acoustic scene analysis can provide useful information about the environment and supports the AS to react appropriately. However, compared to most other application areas, analysis and enhancement of acoustic signals captured by ASs is not only complicated by external sources of signal degradation but also by very specific challenges like internal and self-created ego-noise. This paper first gives an overview of a typical acoustic scenario an AS is exposed to. Then, we consider the specific problem of ego-noise suppression and propose to use motor data to predict the characteristic time-varying harmonic structure of ego-noise. This knowledge is then incorporated into a multichannel dictionary-based algorithm. The resulting two-stage ego-noise reduction scheme is evaluated for ego-noise of a humanoid robot and outperforms a comparable method that uses no motor data but a a larger dictionary.

Index Terms— Acoustic Scene Analysis, autonomous systems, ego-noise reduction, humanoid robot

1. INTRODUCTION

An autonomous system (AS) is typically equipped with numerous sensing modalities to gather multimodal information about its environment. This is the basis for reacting on unanticipated events autonomously. A special focus is here on the acquisition of acoustic information by means of microphones which is then further processed and analysed to extract specific information from the environment. This task is generally referred to as *Acoustic Scene Analysis*, c.f. Fig. 1, and includes several subtasks such as detection and classification of acoustic events [1], source identification [2], extracting and enhancing signals of desired sources ('targets'). After localization [3] and tracking of targets [4], the extracted information is used to determine instructions to control the actuators of the AS, like motors and joints for moving but also loudspeakers for human-machine interaction. A typical sound field captured by an M microphone array of an AS contains different components, as depicted in Fig. 1. One or several targets should be extracted and enhanced out of a mixture of *background noise*, *acoustic feedback* and, very specific for ASs, *ego-noise*.

Background noise comprises diffuse and coherent interferers and reduction approaches are complicated by the absence of a noise reference signal. Common approaches for multichannel microphone apertures are beamforming [5] or other temporal filtering approaches [6], [7].

The *acoustic feedback* or *acoustic echo* results from the re-recorded loudspeaker signal and is addressed by Acoustic Echo Cancellation (AEC), e.g. [8], [9]. Since the played-back loudspeaker signal is available as reference, this can be considered as supervised signal estimation problem. Existing approaches for autonomous systems, specifically for (humanoid) robots, consider mainly a combination of beamforming and AEC, e.g. [10], [11] and [12].

Noise which is created by motors, joints and mechanical components of an AS is referred to as *ego-noise*. In contrast to background noise and acoustic echoes, ego-noise is very characteristic for an AS especially if the ASs are mobile and move actively. Typically, an AS has various ego-noise sources that generally have individual time, spectral and spatial characteristics. In many cases, there is a primary ego-noise source (*internal* ego-noise), e.g., a motor. In addition, the interaction of an AS with the environment causes *self-created* noise, e.g., the noise of the footsteps of a humanoid robot or the tire noise of a self-driving vehicle. In general, ego-noise exhibits two common properties: First, since the AS is restricted to a limited number of degrees of freedom, time, spectral and spatial properties of (internal) ego-noise should live on a manifold of according dimensionality. Second, the instantaneous internal state of the AS, e.g., angle information collected by proprioceptors of joints or the rotation speed of propellers of a drone, provides important extra information that can be exploited for

The research leading to these results was partly supported by the European Unions Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609465.

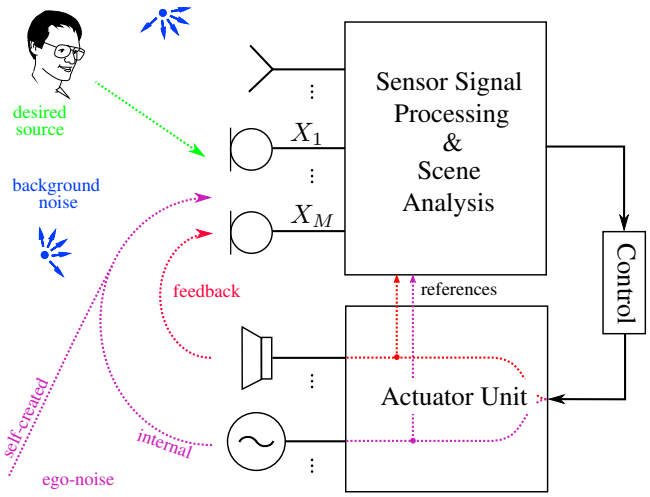

Fig. 1. Schematic signal acquisition and signal processing unit, consisting of M microphones, other sensors and one loudspeaker and motor as exemplary actuators. Here, one target source is to be extracted from a mixture of background noise, acoustic echo and ego-noise.

noise suppression. From this perspective, ego-noise suppression is conceptually located between acoustic feedback and background noise suppression since a reference parameter but not a direct reference signal is available. Dependent on the kind of ego-noise, different reduction approaches have been presented in literature. Stationary ego-noise, e.g, constant rotation speed of a cooling fan or propeller noise, qualifies (multichannel) Wiener filtering approaches for reduction, c.f. [13] and [14]. Non-stationary, but spectrally structured ego-noise can be tackled with *non-negative matrix factorization* (NMF) methods, e.g., [15], [16] for single-channel or [17], [18] for multichannel systems. In the context of humanoid robots, motor data was used in [19] as input to a deep neural network (DNN) to predict the power spectral density (PSD) of the ego-noise. [20] associates points in the motor data space with a ego-noise PSD templates which are then used for subtraction during testing.

In this paper, motor data is used to predict the intrinsic harmonic structure of ego-noise. We propose a method to incorporate this knowledge into a multichannel dictionary-based ego-noise reduction approach. The resulting two-stage suppression procedure is applied for ego-noise reduction for a humanoid robot. The results show that even with smaller overall dictionary size, better suppression results can be achieved than without using motor data.

In this paper, Sec. 2 first summarizes the considered multichannel dictionary approach. Then, it is discussed how motor data can be used to predict intrinsic harmonic structure of ego-noise patterns and how this can be incorporated to the learning stage of the algorithm. Results are presented in Sec. 3 and demonstrate the superior performance of the proposed method.

2. TWO-STAGE MOTOR DATA-BASED APPROACH FOR EGO-NOISE REDUCTION

In the following, we assume ego-noise that is primarily caused by a motor operating at varying speeds and accelerations. Examples for this are pivoting industrial roboter arms and moving humanoid roboters, as considered later. The resulting ego-noise is non-stationary but reveals typically distinctive spectral and spatial characteristics. The basic idea of a dictionary representation is to capture spatial and spectral characteristics by a collection of prototype signals, called *atoms*, collected in a dictionary. In our case, the structured ego-noise signal should be represented by a linear combination of a few atoms at each time frame. If these atoms are specifically designed to represent signals sharing spectral and spatial characteristic of ego-noise only, subtracting these atoms should remove the noise while preserving the residual target signal.

2.1. Multichannel Dictionary learning

We use a multichannel dictionary approach [21] for ego-noise suppression. An M-channel signal is considered in the Short-Time Fourier Transform (STFT) domain. Per frequency bin, the M microphone channels are concatenated, giving a signal vector of dimension MN, where N represents the number of frequency bins per channel. The dictionary is given by $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K] \in \mathbb{C}^{MN \times K}$ containing K atoms $\mathbf{d}_k \in \mathbb{C}^{MN}$ with $k = 1, \ldots, K$. The assumed signal model interprets each atom as contribution of a set of sound sources that are distributed over the body of the robot. A multichannel ego-noise spectrogram frame is then approximated by a linear combination of at most S atoms, where $S \ll K$ is assumed. S is also referred to as *sparsity level*. Beyond this, [21] introduces a time-varying phase matrix $\mathbf{\Phi}_l \in \mathbb{C}^{N \times K}$ that allows to adjust the phase of the dictionary entries, e.g., in order to compensate for time differences of arrival for the noise components of the various sources. This phase corrected dictionary is denoted by $\mathbf{D}\{\mathbf{\Phi}_l\}$, where the curled brackets indicate that nk-th element of $\mathbf{\Phi}_l$ is multiplied with the M bins associated with the frequency index n in atom k. The overall optimization problem is then given by

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{y}_l - \mathbf{D}\{\mathbf{\Phi}_l\}\mathbf{x}_l\|_2^2 \\ \text{subject to} \quad & \|\mathbf{x}_l\|_0 \leq S, \\ & |\phi_{n,k,l}| = 1, \ \forall k, \ \forall n, \end{aligned} \tag{1}$$

where $\phi_{n,k,l}$ is the nk-th element of $\mathbf{\Phi}_l$. Here, $\mathbf{x}_l \in \mathbb{C}^K$ *picks* the atoms from the dictionary. Since maximal S atoms are chosen and $S \ll K$ is assumed, $\mathbf{x}_l$ is sparse. In this context, $\|.\|_0$ denotes the l_0-norm that indicates the number of nonzero elements in $\mathbf{x}_l$. $\|.\|_2$ abbreviates the l_2-norm.

Eq.(1) is minimized w.r.t. different arguments, depending on which stage of the algorithm is considered. During the *training stage*, Eq. (1) is minimized w.r.t. $\mathbf{D}$, $\mathbf{x}_l$ and $\mathbf{\Phi}_l$, using the *phase-optimized K-SVD* algorithm (PO-KSVD, [21]) which is based on [22]. PO-KSVD learns a dictionary $\mathbf{D}$ from a set of training data $\{\mathbf{y}_l\}_{l=1,\dots,L}$ by alternating between a sparse coding step and a dictionary update step. During *Testing*, Eq.(1) is minimized w.r.t. $\mathbf{x}_l$ and $\mathbf{\Phi}_l$ while the dictionary $\mathbf{D}$ is fixed. The best-matching entries of $\mathbf{D}$ are searched and subtracted from a test signal $\mathbf{y}_l$, being, for example, a mixture of ego-noise and speech. Finding the best combination of dictionary entries is an NP-hard problem in K and S, and is often approximated using iterative greedy methods. Here, we use [21] that chooses orthogonal matching pursuit (OMP, after [23], [24]) due to its high accuracy and extends it by a phase optimizing step. The resulting algorithm is called PO-OMP (for *phase-optimized* OMP).

2.2. Prediction of the harmonic structure of ego-noise

The physical state of a rotating motor can be described by its rotation frequency $f_{R,l}$ at timestamp t_l. However, $f_{R,l}$ is not always directly observable. Then, a proprioceptor may provide position information in terms of angle α_l of a joint that is driven by the motor. From this, we can estimate on the rotation frequency by approximating the angular velocity from α_l at two successive timestamps

$$\dot{\alpha}_l = \frac{\alpha_l - \alpha_{l-1}}{t_l - t_{l-1}}. \tag{2}$$

and taking also the mechanical translation between motor and joint into account, given by the velocity-reduction-ratio γ. The rotation frequency is then given by $f_{R,l} = \gamma \cdot \dot{\alpha}_l$. Both position angle and angular velocity are referred to as motor data in the following.

Engine noise exhibits harmonic, deterministic structure in the spectrogram that appears mainly at multiples of half the rotation frequency of the considered motor [25]. Hence, the position of the i-th harmonic in an ego-noise motor noise spectrogram can be predicted by

$$f_{P,l}^{(i)} = \frac{f_{R,l}}{2} \cdot i = \frac{\gamma \cdot \dot{\alpha}_l}{2} \cdot i. \tag{3}$$

As example, Fig. 2 (red lines) illustrates the prediction of ego-noise harmonics for a movement of the right arm of the NAO robot (c.f. Sec. 3) [26]. The ego-noise, which is mainly present at the lower half of the frequency range, is clearly visible as harmonics whose shape follows the prediction from Eq.3 up to $i = 5$. However, there are residual parts of the ego-noise that cannot be estimated by the introduced model. In the following, this is used to train two dictionaries $\mathbf{D}^{(\text{harm})}$ an $\mathbf{D}^{(\text{res})}$, for the harmonic and the residual part of the spectrogram, respectively. During testing, both dictionaries are then applied successively to suppress the ego-noise.

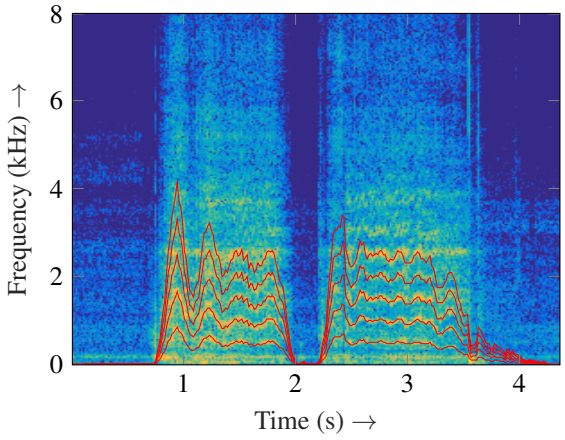

Fig. 2. Spectrogram of an ego-noise recording for a right arm movement of the humanoid robot NAO. Harmonic components can be predicted (red) using angular velocity and Eq.3 with $i = 1, \dots, 5$.

2.3. Masking for signal extraction

To extract the harmonic parts of a given spectrogram $\mathbf{y}_l$, we propose to construct a time-varying frequency mask $\mathbf{w}_l \in \mathbb{R}^{MN}$ based on the predicted harmonics from Eq.(3) which is then applied to $\mathbf{y}_l$ to obtain $\tilde{\mathbf{y}}_l = \mathbf{y}_l \odot \mathbf{w}_l$, where $\odot$ denotes element-wise multiplication. $\mathbf{w}_l$ consists of elements w_{ln}, $n = 1, \dots, N$, repeated M times for each n, i.e., for all M microphones the same weighting w_{ln} is applied for a given frequency bin. w_{ln} is computed by

$$w_{ln} = \frac{\max\limits_{i=1,\dots,I} \left\{ \exp\left(\frac{-\left(f_n - f_{P,l}^{(i)}\right)^2}{\sigma^2} \right) \right\} + \epsilon}{1 + \epsilon}. \tag{4}$$

Hence, w_{ln} is determined by evaluating the most dominant component of an ensemble of I Gaussians, where the i-th Gaussian is centered around the i-th predicted harmonic, $i = 1, \dots, I$. Parameter ϵ is typically set to a small positive number, preventing that $\tilde{\mathbf{y}}_l^{(\text{harm})}$ is not completely set to values near zero for wide areas where no Gaussian contributes significantly. The variance of the Gaussian σ^2 controls the width of the mask around the harmonics and thereby how wide the signal is extracted. It is adjusted experimentally. The optimization problem in Eq.(1) is then solved with $\tilde{\mathbf{y}}_l^{(\text{harm})} = \mathbf{w}_l \odot \mathbf{y}_l$ to obtain the harmonic-specific dictionary $\mathbf{D}^{(\text{harm})}$. Analogously, we use the inverse map $\bar{\mathbf{w}}_l = \mathbf{1} - \mathbf{w}_l$ to obtain the residual parts of the ego-noise and solve (1) with $\tilde{\mathbf{y}}_l^{(\text{res})} = \bar{\mathbf{w}}_l \odot \mathbf{y}_l$ to obtain $\mathbf{D}^{(\text{res})}$ (note that $\mathbf{1}$ is a vector of the same dimension as $\mathbf{w}_l$ with all elements being 1). The property of the mask, $\mathbf{w}_l + \bar{\mathbf{w}}_l = \mathbf{1}$, guarantees that the spectrogram can be fully reconstructed from $\tilde{\mathbf{y}}_l^{(\text{harm})}$ and $\tilde{\mathbf{y}}_l^{(\text{res})}$, i.e.,

$$\tilde{\mathbf{y}}_l^{(\text{harm})} + \tilde{\mathbf{y}}_l^{(\text{res})} = \mathbf{y}_l \odot \underbrace{(\mathbf{w}_l + \bar{\mathbf{w}}_l)}_{\mathbf{1}} = \mathbf{y}_l. \tag{5}$$

As a consequence, if $\tilde{\mathbf{y}}_l^{(\text{harm})}$ and $\tilde{\mathbf{y}}_l^{(\text{res})}$ can be approximated by $\mathbf{D}^{(\text{harm})}$ and $\mathbf{D}^{(\text{res})}$, respectively, $\mathbf{y}_l$ is also reconstructed properly.

3. RESULTS

For the experimental evaluation of the presented approach, we conducted experiments with a NAO H25 humanoid robot [26]. The robot has 26 joints in total, two in the head, twelve in the two arms, twelve in the two legs. For audio recording, we used a modified head developed during the EU FP7 Project EARS [27] with a microphone array of 12 sensors.

In the following, we consider a scenario in which a target source is talking to the robot while it is waving with the right arm. The movement includes all six joints of the right arm. We used exclusively the motor data of the right shoulder pitch joint to approximate angular speed and predict the harmonics according to Eq.(2) and Eq.(3), respectively. For the recordings, we used four microphones, two located at the front side of the head, one on the top and one at the back side. The sampling frequency of the motor data is given by $f_M \approx 200\,\text{Hz}$, the audio signals are sampled with $f_S = 16\,\text{kHz}$. The audio recordings are transformed to STFT domain using a Hamming window of length $32\,\text{ms}$ with overlap of $50\,\%$. To associate a motor data value to each STFT frame, we computed the arithmetic average of all motor data samples falling into the duration of an STFT frame.

NAO performed its movements in a room with moderate reverberation ($T_{60} = 200\,\text{ms}$). For testing, 200 utterances from the GRID corpus [28] were recorded. The loudspeaker was positioned at $1\,\text{m}$ distance of NAO, at a height of $1\,\text{m}$. The recorded utterances were added to out-of-training movement noise. These mixtures were then used to evaluate the ego-noise suppression algorithms described above.

We recorded $30\,\text{s}$ training data which was used to train $\mathbf{D}$ on audio data alone, and to train $\mathbf{D}^{(\text{harm})}$, $\mathbf{D}^{(\text{res})}$, respectively, jointly on audio and motor data as proposed above. The frequency mask was parameterized with $\epsilon = 0.1$ and $\sigma = 2.4$. During testing, the recorded utterances were added to out-of-training movement noise. Afterwards, ego-noise reduction was compared using PO-OMP with $\mathbf{D}$ on the one hand and $\mathbf{D}^{(\text{harm})}$, $\mathbf{D}^{(\text{res})}$ applied successively on the other hand. For both, dictionary sizes and sparsity level were chosen such that optimum results were obtained (c.f. caption Table 1).

The overall performance of the ego-noise suppression is measured in terms of Signal-to-Inference-Ratio (SIR in dB) and Signal-to-Distortion-Ratio (SDR in dB), using Matlab functions provided by [29]. As all individual source signals are required for computing SDR and SIR, ego-noise and speech utterances were recorded separately and added to evaluate the proposed approaches.

While SIR measures the overall noise cancellation, SDR also incorporates information about how much speech is distorted by the suppression algorithm. Additionally, we measure speech keyword error rate (WER), using *pocketsphinx* [30] in the GRID corpus [28], as defined by the CHiME challenge [31].

Table 1. PO-OMP, suppression results using dictionaries $\mathbf{D}$ ($K = 20$, $S = 3$), $\mathbf{D}^{(\text{harm})}$ ($K^{(\text{harm})} = 5$, $S^{(\text{harm})} = 2$), $\mathbf{D}^{(\text{res})}$ ($K^{(\text{res})} = 10$, $S^{(\text{res})} = 2$)

	SIR [dB]	SDR [dB]	WER [%]
$\mathbf{D}^{(\text{harm})}$, $\mathbf{D}^{(\text{res})}$	13.72	4.72	29.9
$\mathbf{D}^{(\text{harm})}$	11.45	3.95	37.0
$\mathbf{D}$	11.68	4.03	34.6
Unprocessed	-3.45	-4.02	59.9

Table 1 shows the suppression results after successive use of $\mathbf{D}^{(\text{harm})}$ and $\mathbf{D}^{(\text{res})}$ compared to using $\mathbf{D}$ trained with audio data alone, e.g, without motor data. All approaches clearly improve all evaluated metrics. However, the two-stage motor data-based method outperforms the approach using only $\mathbf{D}$. Interestingly, even if evaluating the reduction with $\mathbf{D}^{(\text{harm})}$ only, metrics are close to those of $\mathbf{D}$. The splitting of the ego-noise reduction into two successive stages allows to parameterize $\mathbf{D}^{(\text{harm})}$ and $\mathbf{D}^{(\text{res})}$ differently. Thereby, dictionary size and sparsity level can be tuned individually on the specific properties of the ego-noise. In this example, dictionary $\mathbf{D}^{(\text{harm})}$ requires a smaller size ($K^{(\text{harm})} = 5$) to match the harmonic structure, while $\mathbf{D}^{(\text{res})}$ needs a size of $K^{(\text{res})} = 10$ to approximate the residual part best. In contrast, $\mathbf{D}$ has a size of $K = 20$, what is effectively larger than $K^{(\text{harm})} + K^{(\text{res})}$. This shows that, despite a smaller overall dictionary size, the splitted approach outperforms a joint reconstruction of the entire ego-noise spectrogram using only $\mathbf{D}$. It is worth mentioning that although $S^{(\text{harm})} + S^{(\text{res})} > S$, both approaches are computationally equal efficient since the phase optimization of the third atom requires most computational effort.

4. CONCLUSION AND OUTLOOK

Among the various tasks in acoustic scene analysis for ASs we pointed to the special role of ego-noise relative to other multimicrophone acquisition systems. We proposed a method to incorporate ego-noise harmonics into a multichannel dictionary-based ego-noise reduction method, which results in a better performance than exploiting no motor data. For future work, we plan to verify the proposed method using other learning-based ego-noise suppression approaches like (multichannel) NMF. Beyond this, it appears promising to extend the motor data model such that not only harmonics but also other parts of the ego-noise can be predicted.

5. REFERENCES

[1] D. Stowell *et al.*, "Detection and Classification of Acoustic Scenes and Events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.

[2] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, May 2002, vol. 4, pp. IV–4072–IV–4075.

[3] J.H. DiBiase, *A High-Accuracy, Low-Latency Technique for Talker Localization in Reverberant Environments Using Microphone Arrays*, PhD Thesis, Brown University, Providence, Rhode Island, 2000.

[4] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*, Artech House, Boston, Aug. 1999.

[5] B.D. Van Veen and K.M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[6] S. Gannot *et al.*, "A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 4, 2017.

[7] M. Brandstein and D. Ward, *Microphone Arrays - Signal Processing Techniques and Applications*, Springer, Berlin, Heidelberg, 1st edition, 2006.

[8] C. Breining *et al.*, "Acoustic echo control. An application of very-high-order adaptive filters," *IEEE Signal Processing Mag.*, vol. 16, no. 4, pp. 42–69, July 1999.

[9] W. Kellermann, "Strategies for combining acoustic echo cancellation and adaptive beamforming microphone arrays," in *IEEE Int. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, Apr. 1997, vol. 1, pp. 219–222 vol.1.

[10] A. El-Rayyes *et al.*, "Acoustic echo control for humanoid robots," in *Proc. 43rd Annu. Conf. Acoust. (DAGA)*, Mar 2016.

[11] R. Takeda *et al.*, "ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, Apr. 2009, pp. 3677–3680.

[12] J. Beh *et al.*, "Combining acoustic echo cancellation and adaptive beamforming for achieving robust speech interface in mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Syst. (IROS)*, Sept. 2008, pp. 1693–1698.

[13] S. Doclo *et al.*, "Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," *Speech Communication*, vol. 49, no. 7, pp. 636–656, 2007.

[14] A. Spriet *et al.*, "Spatially pre-processed speech distortion weighted multi-channel wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367–2387, 2004.

[15] M. Kim and P. Smaragdis, "Mixtures of local dictionaries for unsupervised speech enhancement," *IEEE Processing Lett.*, vol. 22, no. 3, pp. 293–297, 2015.

[16] 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.

[17] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures with application to blind audio source separation," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2009, pp. 3137–3140.

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

[19] 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.

[20] 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.

[21] 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.

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

[23] T.T. Cai and L. Wang, "Orthogonal matching pursuit for sparse signal recovery with noise," *IEEE Trans. Inform. Theory*, vol. 57, no. 7, pp. 4680–4688, 2011.

[24] Y.C. Pati *et al.*, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Syst. and Comput.*, 1993, pp. 40–44.

[25] 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 Processing Conference (EUSIPCO)*, Sept. 2000, pp. 1–4.

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

[27] "Seventh framework programme 'Embodied Audition for RobotS' (EARS)," `https://robot-ears.eu/`, Accessed: 2017-09-25.

[28] M. Cooke *et al.*, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[29] 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/`.

[30] D. Huggins-Daines *et al.*, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *IEEE Trans. Acoust., Speech, Signal Processing*. IEEE, 2006, vol. 1, pp. 185–188.

[31] E. Vincent *et al.*, "The second 'chime' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *IEEE Workshop on Automat. Speech Recognition and Understanding (ASRU)*, 2013, pp. 162–167.