Challenges in Acoustic Signal Enhancement for Human-Robot Communication

Heinrich W. Löllmann, Hendrik Barfuss, Antoine Deleforge, Stefan Meier, and Walter Kellermann

Chair of Multimedia Communications and Signal Processing (LMS), FAU Erlangen-Nürnberg, 91058 Erlangen
Email: {loellmann,barfuss,deleforge,smeier,wk}@lnt.de
Web: www.lms.lnt.de

Abstract

The construction of a humanoid robot, which can communicate with humans in a natural manner, is a worthwhile and challenging task alike. This paper discusses some major difficulties encountered in acoustic signal acquisition by the widely used humanoid robot NAO and the implications for the design of the signal enhancement algorithms that are needed for human-robot communication.

Measurements with this low-cost robot, whose microphones and loudspeakers are mounted in the head, show the challenges for the noise reduction and acoustic echo control (AEC) due to ego-noise and nonlinear loudspeaker characteristics.

It is also discussed how peripheral microphones at the limbs of the robot can mitigate these problems and offer new prospects for multi-channel signal enhancement.

1 Introduction

The desire to construct a humanoid robot is a long-lasting dream of mankind. In 1495, Leonardo da Vinci already constructed a humanoid automaton which resembles a mechanical knight [1], and numerous improved designs for a humanoid robot have been devised since then.

The ability of a humanoid robot to interact with a human in a natural way is an important and challenging feature alike. Such robots should look towards the persons they are talking to, understand the content of the conversation and react in a ‘human’ manner.

Speaker identification and localization for robots has been traditionally addressed by using video cameras, while *robot audition* only emerged in the last decade as a topic of its own [2–4]. From the viewpoint of acoustic signal processing, a robot listening to a person has to perform acoustic source detection, localization and tracking, signal extraction, and automatic speech recognition (ASR). Moreover, these algorithms should be designed for noisy and reverberant signals such that the robot can be operated in ideally unconstrained acoustic environments.

The design of speech and audio signal processing algorithms for adverse acoustical environments has received a lot of research interest, e.g., [5–7]. However, the use of such approaches for robot audition is often not straightforward as the signal acquisition by humanoid robots is subject to particular problems. As an obvious example, the actuators (i.e., motors for robot motions) cause ego-noise, which leads to a severe degradation of the recorded audio signals, cf., [4].

The goal of this contribution is to illustrate such typical challenges encountered in acoustic signal acquisition by humanoid robots by means of measurements with a widely used robot. Based on this, different concepts for acoustic signal enhancement as necessary for human-robot communication are discussed.

This paper is organized as follows: In Sec. 2, the measurement setup is described and the problem of ego-noise reduction investigated. In Sec. 3, the challenges for AEC are discussed, where Sec. 4 investigates the opportunities and difficulties of using peripheral microphones. Finally, the main conclusions are summarized in Sec. 5.

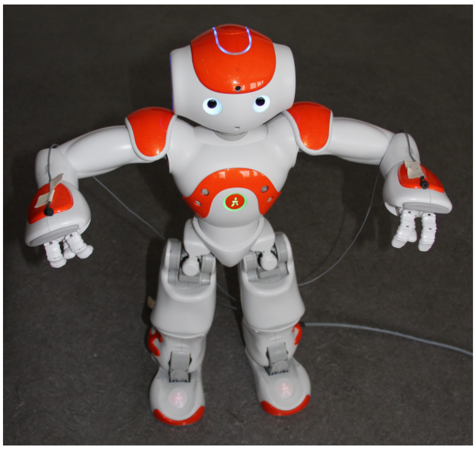

Figure 1: *NAO H25 robot with additional peripheral capsule microphones (AKG CE 20/17) attached on the back of its hands. The height of this robot is 573 mm and the shoulder-to-shoulder length is 275 mm.*

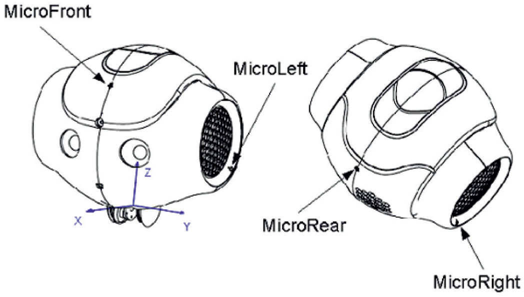

Figure 2: *Mounting of the microphones at the NAO robot head [8]. The shaded dark areas near the left and right microphone depict the covers of the stereo loudspeakers where the shaded dark area near the rear microphone shows the ventilation slot.*

2 Ego-Noise

For exemplifying the typical problems encountered in acoustic signal enhancement for human-robot interaction, we have conducted several measurements with the NAO H25 robot (version 4) of the manufacturer Aldebaran Robotics [8], which is depicted in Fig. 1. This low-cost humanoid robot is taken as a platform for various research projects related to human-robot interaction [9, 10] as well as for the Standard Platform League of the Robot Soccer World Cup (RoboCup) [11].

The NAO robot has four internal microphones as well as two loudspeakers mounted into its head (see Fig. 2). For our experiments, two peripheral capsule microphones

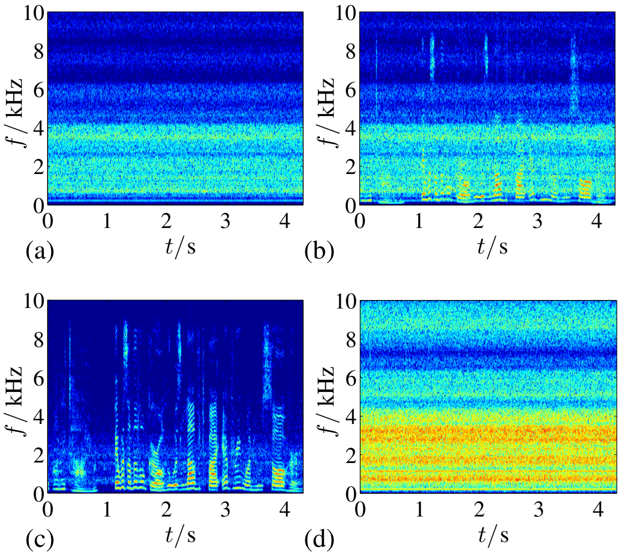

Figure 3: *Spectrograms of head microphone signals: (a) left microphone without speech, (b) speech signal recorded by the left microphone, (c) speech signal recorded by the left microphone with the fan being switched off, (d) speech signal recorded by the rear microphone.*

have also been attached to the backs of the robot's hands (see Fig. 1). All measurements with the NAO have been conducted in a low-reverberation chamber ($T_{60} \approx 50$ ms). To emulate a human speaker, a speech signal was emitted by a loudspeaker at 1 m distance with a sound pressure level (SPL) between 87 dBA and 95 dBA, which resulted in SPLs between 58 dBA and 66 dBA at the robot head.

The mounting of the robot's microphones at its head is a rather common setup as it allows to adopt approaches for binaural listening (see, e.g., [4] for an overview). However, the head of a robot is usually also housing a signal processing unit which is cooled by a fan. Therefore, aside from the desired signals, the head microphones record also unwanted ego-noise as illustrated in Fig. 3.

Fig. 3-a and Fig. 3-d show that the fan noise is stationary and non-white with most of its spectral energy concentrated in a frequency range of up to 4 kHz. Fig. 3-b and Fig. 3-d exemplify that this fan noise can distort the speech component in the microphone signal significantly. It should be noted that the signal shown in Fig. 3-c still contains some noise caused by the actuators that keep the robot upright. Fig. 3-d reveals that the rear microphone captures most fan noise (see also Fig. 2). With no loudspeaker being active, an SPL of 49 dBA was measured at the left and 52 dBA at the rear microphone, but these values do not capture the effects of structure-borne sound.

The interference by the ego-noise must be expected to lead to significant performance degradations for the localization and extraction of acoustic sources as well as the subsequent ASR. The effects on source detection and localization can be mitigated by combining auditory data of the head microphones and visual data of the head cameras, e.g., [12]. For ASR, there exist numerous techniques to improve the robustness in noisy and reverberant environments, which can be divided into two major classes: Either the acoustic model of the ASR system is adapted to the noise and reverberation, e.g., [7, 13], or noise and reverberation are tackled already at the signal level by means of speech enhancement algorithms, e.g., [14]. Our focus is here on the second class.

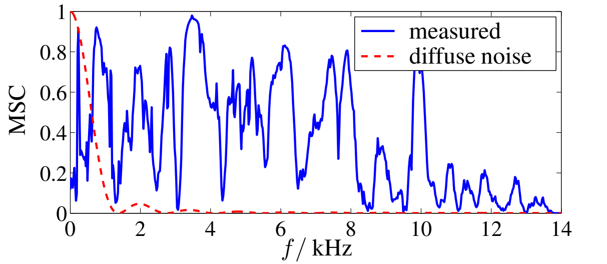

Figure 4: *MSC for the signals of the left and right microphone (without speech) of the robot head with a distance of d=0.12 m. The red dashed line marks the MSC for diffuse noise $|\Gamma_{diff}(f)|^2 = si^2(2\pi f \frac{d}{c})$ with c=330 m/s.*

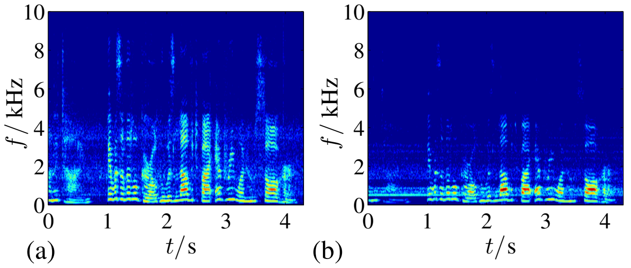

Figure 5: *Spectrograms of head microphone signals after noise reduction by an MWF using all 4 head microphones: (a) left microphone, (b) rear microphone. The noisy input signals are shown by Fig. 3-b and Fig. 3-d.*

Fig. 4 provides a plot of the magnitude-squared coherence (MSC) of the fan ego-noise. The MSC has been calculated by the Welch method [15] for half-overlapping frames of length of 2048 ($f_s = 48$ kHz). It can be observed that the fan noise cannot be described by a spatially white, diffuse, or coherent noise field model. Accordingly, the use of multi-channel noise reduction schemes which assume such a specific noise field as, e.g., [16] is not appropriate.

Instead, we have applied the speech distortion weighted multi-channel Wiener filter (SDW-MWF) [17] to reduce the fan noise. For a beamformer with N sensors, the matrix with the spectral beamformer coefficients at frequency Ω is given by [18, Chap. 9]

$$\boldsymbol{W}(\Omega) = \frac{\boldsymbol{\Phi}_{nn}^{-1}(\Omega)\,\boldsymbol{\Phi}_{xx}(\Omega) - \boldsymbol{I}}{\lambda^{-1}(\Omega) + \mathrm{tr}\left[\boldsymbol{\Phi}_{nn}^{-1}(\Omega)\,\boldsymbol{\Phi}_{xx}(\Omega)\right] - N} \qquad (1)$$

with $\boldsymbol{\Phi}_{nn}(\Omega)$ and $\boldsymbol{\Phi}_{xx}(\Omega)$ denoting the power spectral density (PSD) matrices of the noise and input signals, respectively. $\boldsymbol{I}$ denotes the $N \times N$ identity matrix and tr[.] the trace of a matrix. The factor $\lambda^{-1}(\Omega)$ controls the amount of speech distortions. No distortions occur for $\lambda^{-1}(\Omega) \equiv 0$ as then the SDW-MWF becomes equal to an MVDR beamformer. Here, a frequency-independent value of $\lambda(\Omega) \equiv 1$ was chosen and the noise PSD matrix $\boldsymbol{\Phi}_{nn}(\Omega)$ was estimated by means of pre-recorded fan ego-noise. Fig. 5 and Table 1 show that this approach leads to a significant reduction of the ego-noise.[1] This is due to the fact that the noise PSD matrix can be well estimated

[1] For the calculation of the segmental signal-to-noise ratio (SNR), the speech recordings without fan noise were taken as desired signals and frames without speech activity were excluded.

Table 1: *Segmental SNR values before and after denoising.*

microphone position	left	right	front	rear
seg. input SNR/ dB	-1.6	-3.6	-16.1	-13.7
seg. output SNR/ dB	15.4	15.6	14.3	14.8

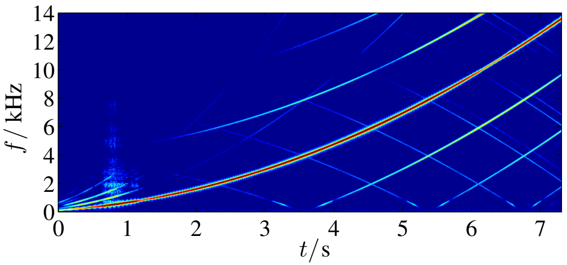

Figure 6: *Spectrogram for a quadratic chirp signal emitted by the right head loudspeaker and recorded by a measurement microphone at 0.1 m distance.*

for stationary noise sources. In contrast, reducing the ego-noise caused by the actuators by this approach has shown much less satisfactory results, since the estimation of the noise PSD matrix for such nonstationary noise sources is much more challenging. Various approaches have been proposed to address this problem. In [19], it is proposed to perform spectral subtraction where the needed noise PSD is predicted by neural networks from the status of the joints. Another approach is to estimate the noise PSD for spectral subtraction by identifying noise patterns (noise templates) with the help of ego-noise databases [20].

As in other typical hands-free communication scenarios, the recorded microphone signals are typically also distorted by environmental noise. Moreover, the speech signal is usually also rather reverberant since robot and human will typically be separated by one meter or more. The design of ASR systems for humanoid robots in such conditions is treated, e.g., in [21, 22].

3 Acoustic Echo Control (AEC)

A speaking robot records its own voice by the internal microphones. Therefore, an AEC has to be performed to avoid that this signal is mistaken for the desired signal by the ASR system. The design of the AEC system for a microphone-loudspeaker setup as shown in Fig. 2 is very challenging due to the following reasons: First, the microphones simultaneously capture the previously discussed ego-noise. Second, there is only a short distance between the microphones and the loudspeaker, which results in a high level for the feedback signal. A third major problem is that the acoustic feedback path cannot be modeled sufficiently well by a linear transfer function.

The spectrogram for a quadratic chirp signal emitted by a head loudspeaker is depicted in Fig. 6. The volume of the loudspeaker was set to 50% of its maximum and the measured SPLs close to the head loudspeaker were between 92 dBA and 100 dBA.

At higher frequencies above 2 kHz, significant aliasing effects can be noticed. Between 0.6 and 1.1 ms on the time axis, distortions which are very likely due to a resonating housing can be observed. At low frequencies, additional harmonics can also be observed, which indicate a nonlinear transfer characteristic and call for nonlinear AEC schemes. The nonlinear characteristic of the echo path can be modeled by power filters, Volterra filters, or Hammerstein group models, e.g., [23–25]. Obviously, the use of such schemes instead of a linear AEC system causes usually an increased computational load.

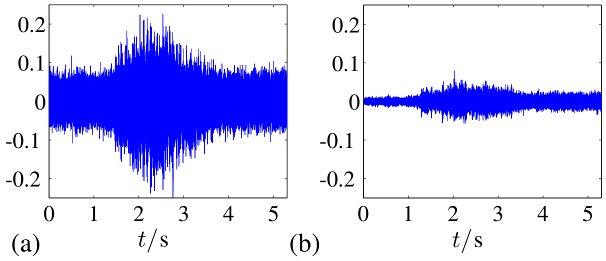

Figure 7: *Self-noise recordings while the NAO is in a position as shown in Fig. 1 and rotating its head: (a) left head microphone, (b) left peripheral microphone.*

4 Peripheral Microphones

An approach to tackle the problem of recording ego-noise caused by the fan and actuators in the robot head is to use peripheral (external) microphones, which are mounted outside the head, e.g., on the backs of the robot's hands as shown by Fig. 1. The opportunities and difficulties of this approach are discussed in the following.

Fig. 7 reveals that the peripheral microphone captures much less ego-noise of the fan and head actuators than the head microphones.[2] For this measurement, the maximal SPLs near the external and head microphone were both around 53 dBA, whereas a maximum value of 70 dBA was measured at the ventilation slot near the rear microphone.

For AEC, the use of peripheral microphones instead of those in the head has two major advantages. First, the level of the feedback signal is lower due to the higher loudspeaker-microphone distance. Second, the fan-related ego-noise has a much lower impact. However, this is achieved at the cost of a time-varying echo path when the robot moves its limbs, which will challenge the adaptation performance of the AEC algorithm.

Speech enhancement with microphones whose positions are unknown can be addressed, e.g., by noise reduction algorithms for sensor networks [26]. However, a robot can control the position of its limbs and, by this, the position of the peripheral microphones, which offers new possibilities for the design of the speech enhancement system.

The option to control the microphone position and, thus, the array aperture suggests to adapt the microphone spacings to improve the performance of a multi-channel speech enhancement system [27]. A simple incarnation of this concept is given in the following: An array with three microphones is divided into two sub-arrays with two microphones and spacings $d_1^{(j)}$ and $d_2^{(j)}$. Each sub-array performs a two-channel blind signal enhancement (BSE) according to [28]. This algorithm employs a blind signal separation (BSS) scheme, which allows to measure the signal extraction performance by calculating the average MSC between the two respective output signals. The microphone spacing of the sub-array with the inferior BSE performance $d_{\text{inf}}^{(j)}$ is then changed in the

[2] The signal of the peripheral microphone was equalized to provide the same frequency-dependent amplification as the left head microphone.

next iteration step $j+1$ according to the adaptation rule $d_{\text{inf}}^{(j+1)} = \left(\frac{1+n}{2+n}\right)^{(-1)^{n+1}} d_{\text{sup}}^{(j)}$ with $d_{\text{sup}}^{(j)}$ denoting the microphone spacing of the superior sub-array. The procedure starts with $n = 1$ and after the adaptation phase of the BSS algorithm, n is increased by one until the performance of the inferior sub-array becomes superior and so on, such that the two noise suppression algorithms compete against each other. The further development and evaluation of this concept is subject of current research [10, 27].

5 Conclusions

This paper provides an overview of specific problems encountered in robot audition on the basis of acoustic signal measurements with a widely used humanoid robot, and discusses known as well as novel concepts to address these challenges.

The microphones in the robot head capture ego-noise caused by the cooling-fan and actuators. Even though the fan noise cannot simply be described by common noise field coherence models, it can be well suppressed by a SDW-MWF due to its stationarity. In contrast, the reduction of the nonstationary actuator ego-noise is much more challenging as the estimation of the needed noise PSD-matrix is more difficult.

The realization of an AEC system for humanoid robots, which is needed to enable ASR while a robot is talking, turns out to be a very challenging task. One major problem is the nonlinear transmission characteristic of the loudspeaker. Another major problem is the small distance between microphones and loudspeakers in the head as well as the robot ego-noise. These two problems can be alleviated by peripheral microphones, e.g., at the robot hands or arms which, however, comes at the cost of a time-varying echo path when the robot moves its limbs.

It has also been outlined how the ability to control the microphone array topology can be exploited to improve the performance of a two-channel speech enhancement scheme. The further development of this concept is subject of ongoing research.

Acknowledgment

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 609465. It has been conducted as part of the project EARS (Embodied Audition for RobotS).

References

[1] M. E. Rosheim, *Leonardo's Lost Robots*, Springer, Berlin, Heidelberg, 2006.

[2] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active Audition for Humanoid," in *Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX, USA, July 2000, pp. 832–839.

[3] A. Deleforge and R. Horaud, "The Cocktail Party Robot: Sound Source Separation and Localisation with an Active Binaural Head," in *Proc. of Intl. Conference on Human-Robot Interaction (HRI)*, Boston, MA, USA, Mar. 2012, pp. 431 – 438.

[4] S. Argentieri, A. Portello, M. Bernard, P. Danés, and B. Gas, "Binaural Systems in Robotics," in *The Technology of Binaural Listening*, J. Blauert, Ed., Modern Acoustics and Signal Processing, pp. 225–253. Springer, Berlin, Heidelberg, 2013.

[5] E. Hänsler and G. Schmidt, Eds., *Speech and Audio Processing in Adverse Environments*, Springer, Berlin, New York, 2008.

[6] T. Virtanen, R. Singh, and B. Raj (Eds.), *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, Oct. 2012.

[7] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making Machines Understand Us in Reverberant Rooms: Robustness Against Reverberation for Automatic Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[8] Aldebaran Robotics, *NAO Software 1.14.5 documentation*, `https://community.aldebaran-robotics.com`.

[9] "Website of FB 7 EU project HUMAVIPS (grant agreement no. 247525)," `http://humavips.inrialpes.fr`.

[10] "Website of FB 7 EU project EARS (grant agreement no. 609465)," `http://robot-ears.eu`.

[11] "Website of Robot Soccer World Cup (RoboCup) 2013," `http://www.robocup2013.org`.

[12] J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, and R. Horaud, "Active-Speaker Detection and Localization with Microphone and Cameras Embedded into a Robotic Head," in *Proc. of Intl. Conference on Humanoid Robots*, Atlanta, GA, USA, Oct. 2013.

[13] A. Sehr, R. Maas, and W. Kellermann, "Reverberation Model-Based Decoding in the Logmelspec Domain for Robust Distant-Talking Speech Recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1676–1691, Sept. 2010.

[14] R. Gemello, F. Mana, and R. de Mori, "A Modified Ephraim-Malah Noise Suppression Rule for Automatic Speech Recognition," in *Proc. of Intl. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, Mar. 2004, pp. 956–960.

[15] P. D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms," *IEEE Trans. on Audio and Electroacoustics*, vol. 15, pp. 70–73, 1967.

[16] I. A. McCowan and H. Bourland, "Microphone Array Post-Filter Based on Noise Field Coherence," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 6, pp. 709–716, Nov. 2003.

[17] A. Spriet, M. Moonen, and J. Wouters, "Spatially Pre-processed Speech Distortion Weighted Multi-Channel Wiener Filtering for Noise Reduction," *Signal Processing*, vol. 84, pp. 2367–2387, Dec. 2004.

[18] I. Cohen, J. Benesty, and S. Gannot, *Speech Processing in Modern Communication*, Springer, Berlin, Heidelberg, 2010.

[19] A. Ito, T. Kanayama, M. Suzuki, and S. Makino, "Internal Noise Suppression for Speech Recognition by Small Robots," in *Proc. of European Confernce on Speech Communication and Technology (Interspeech)*, Lisbon, Portugal, Sept. 2005, pp. 2685–2688.

[20] G. Ince, K. Nakadai, T. Rodemann, Y. Hasegawa, H. Tsujino, and J. ichi Imura, "Ego Noise Suppression of a Robot Using Template Subtraction," in *Intl. Conference on Inteligent Robots and Systems (IEEE/RSJ)*, St. Louis, MO, USA, Oct. 2009, pp. 199–204.

[21] F. Kraft and M. Wölfel, "Humanoid Robot Noise Suppression by Particle Filters for Improved Automatic Speech Recognition Accuracy," in *Proc. of Intl. Conferene on Intelligent Robots and Systems (IROS)*, San Diego, CA, USA, Oct. 2007.

[22] R. Gomez, K. Nakamura, T. Kawahara, and K. Nakadai, "Multi-Party Human-Robot Interaction with Distant-Talking Speech Recognition," in *Proc. of Intl. Conference on Human-Robot Interaction (HRI)*, Boston, MA, USA, Mar. 2012, pp. 439–446.

[23] F. Küch and W. Kellermann, "Orthogonalized Power Filters for Nonlinear Acoustic Echo Cancellation," *Signal Processing*, vol. 86, pp. 1168–1181, June 2006.

[24] M. Zeller and W. Kellermann, "Advances in Identification and Compensation of Nonlinear Systems by Adaptive Volterra Models," in *Proc. of Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, Nov. 2010.

[25] C. Hofmann, C. Hümmer, and W. Kellermann, "Significance-Aware Hammerstein Group Models for Nonlinear Acoustic Echo Cancellation‘," in *Proc. of Intl. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 2014.

[26] Y. Zeng and R. C. Hendriks, "Distributed Delay and Sum Beamformer for Speech Enhancement in Wireless Sensor Networks via Randomized Gossip," in *Proc. of Intl. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4037–4040.

[27] H. Barfuss and W. Kellermann, "An Adaptive Microphone Array Topology for Target Signal Extraction with Humanoid Robots," in *Intl. Workshop on Acoustic Signal Enhancement (IWAENC)*, Antibes, France, Sept. 2014.

[28] K. Reindl, Y. Zheng, A. Schwarz, S. Meier, R. Maas, A. Sehr, and W. Kellermann, "A Stereophonic Acoustic Signal Extraction Scheme for Noisy and Reverberant Environments," *Computer Speech and Language (CSL)*, vol. 27, no. 3, pp. 726–745, May 2012.