ABSTRACT

Robot audition for humanoid robots interacting naturally with humans in an unconstrained real-world environment is a hitherto unsolved challenge. The recorded microphone signals are usually distorted by background and interfering noise sources (speakers) as well as room reverberation. In addition, the movements of a robot and its actuators cause ego-noise which degrades the recorded signals significantly. The movement of the robot body and its head also complicates the detection and tracking of the desired, possibly moving, sound sources of interest. This paper presents an overview of the concepts in microphone array processing for robot audition and some recent achievements.

Index Terms—Humanoid robots, robot audition, microphone array processing, ego-noise suppression, source tracking

1. INTRODUCTION

Developing a humanoid robot, which can interact with humans in a natural, i.e., humanoid way, is a long-lasting vision of scientists, and with the availability of increasingly powerful technologies, it turns into a realistic engineering task. With the acoustic domain as key modality for voice communication, scene analysis and understanding, acoustic signal processing represents one of the main avenues leading to a humanoid robot, but has received significantly less attention in the past decades than processing in the visual domain.

The design of a system for robot audition, which should be operated in real-world environments, starts with the observation that the recorded microphone signals are typically impaired by background and interfering noise sources (speakers) as well as room reverberation [1,2]. Thereby, the distance between robot and speaker is relatively large in comparison to, e.g., hands-free communication systems for mobile phones. In addition, the movements of a robot, its actuators (motors) and CPU cooling fan cause ego-noise (self-noise) which degrades the recorded signals significantly. Not at least, the movements of the robot body and its head also complicate the detection and tracking of the desired, possibly moving, speaker(s), cf., [3].

Finally, the implementation of algorithms on a robot is often linked to mundane hardware-related problems: The microphone, video and motor data streams are not necessarily synchronized. Besides, the interaction with a robot requires real-time processing where the limited CPU power of an autonomous robot precludes algorithms with a high computational load. A high algorithmic signal delay cannot be allowed either, as a humanoid robot should react, similar to humans, instantaneously to acoustic events in its environment.

The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no 609465. It has been conducted as part of the project EARS (Embodied Audition for RobotS).
3. ACOUSTIC SOURCE LOCALIZATION AND TRACKING

Effective robot audition requires awareness of the sound scene including the positions of sound sources and their trajectory over time. Estimates of source localization are needed, for example, to steer the beamformer and to track talkers. Time-varying acoustic maps can be used to capture this type of information. In acoustic localization, it is common that only bearing estimation can be obtained (DOA estimation), while range information is normally not available. A volumetric array comprising at least 4 microphones is required to identify a unique DOA in 2 dimensions (azimuth and inclination). A spherical harmonics (SH) representation of the sound field can be assumed for the almost spherical shape of a robot head, which suggests to perform the DOA estimation in the SH-domain. A method with low computational cost based on pseudointensity vectors (PIVs) [17][18] is attractive given the limitations of robot embedded processing. This approach has been enhanced in [19], albeit with additional computational cost, to use subspace PIVs in order to obtain better performance and robustness.

Unlike many other applications of microphone arrays, robot-based arrays move. For DOA estimation, it is therefore necessary to account for the dynamics such as in the motion compensation approach of [20]. However, movement also enables to acquire additional spatial information and this can be exploited to enhance the DOA estimation performance in comparison to static sensing [21]. In addition to the movement of the robot, DOA estimation has to

Fig. 2. a) Design drawing of the new 12-microphone prototype head, b) NAO robot with new prototype head and robomorphic array.
account for the movement of the sound sources since talkers are rarely stationary. It is advantageous to employ tracking techniques that exploit models of source movement in order to improve on the raw output of a DOA estimator. This is challenging to achieve from acoustic sensing because of the lack of range information. Bearing-only source tracking has been developed, e.g., in [22] which exploits movement of the robot to estimate the location tracks as talkers move. Tracking is also advantageous in improving robustness to missing data and estimation variance. When the robot explores the acoustic environment, it needs to determine simultaneously its own location as well as a map of other sound sources in the vicinity. Techniques for acoustic simultaneous localization and mapping (A-SLAM) are proposed in [23], which localize the moving array and infer the missing source-sensor range from the estimated DOAs. DOA estimation accuracy is commonly degraded in reverberant environments due to acoustic reflections. The direct-path dominance test and the direct-path relative transfer function are exploited in the methods of [23,25] that aim to base the DOA estimates mostly on direct path acoustic propagation rather than the acoustic reflections, thereby greatly improving robustness to the reverberation commonly encountered in use cases as for service robots.

If the target sources are in the field-of-view of the robot’s cameras, audio-visual localization and tracking should be exploited, e.g., [26], which is beyond the scope of this paper.

4. EGO-NOISE REDUCTION AND DEREVERBERATION

The audio signals recorded by the microphones of the robot are usually not only distorted by external sources (room reverberation, background noise etc.), but also ego-noise caused by the actuators and CPU cooling fan of the robot, e.g., [2]. Thereby, the challenging task of suppressing the non-stationary actuator ego-noise is usually accomplished by exploiting information about the motor states and a priori knowledge about the specific structure of the noise sources using, e.g., a database with noise templates [27] or ego-noise prediction based on neural networks [28] where the actual enhancement is performed by spectral subtraction.

A multichannel approach, which considers also the phase information of the ego-noise, has been proposed in [29]. The actuator ego-noise is suppressed by a phase-optimized K-SVD algorithm where the needed dictionary is learned by sparse coding using multichannel ego-noise recordings. Ego-noise samples are modeled by a sparse combination of ego-noise prototype signals in the STFT-domain and capture the spectral as well as spatial characteristics of the current ego-noise sample. The evaluation of this approach for the NAO robot in [29] shows that a better speech quality and lower word error rate (WER) is achieved in comparison to related approaches based on non-negative matrix factorization (NMF) [30] or conventional K-SVD [31]. An extension of this approach in [32] uses nonlinear classifiers to associate the current motor state of the robot to relevant sets of entries in the learned dictionary. This approach achieves a significant reduction of the computational complexity in comparison to the original approach [29] while achieving at least the same noise reduction performance.

Besides ego-noise, room reverberation causes a significant degradation of the recorded audio signals and, hence, the ASR performance. In [33,34], multichannel dereverberation is performed by MINT-filtering to enhance the performance of the subsequent signal separation by independent component analysis (ICA).

The almost spherical shape of a robot head suggests to perform the dereverberation in the SH-domain. In [35], the generalized weighted prediction error (GWPE) algorithm [36] for speech dereverberation is formulated in the SH-domain and offers computational savings over a conventional space-domain implementation when a high number of microphones is used.

5. SPATIAL FILTERING AND ACOUSTIC ECHO CANCELLATION

Spatial filtering for multichannel signal enhancement for robot audition can be realized by a data-dependent approach, e.g., [37,38], a data-independent approach, e.g., [39] or a combination of both approaches, e.g., [9]. A data-dependent approach usually achieves a higher signal enhancement than a data-independent approach at the cost of a higher computational complexity. Moreover, the required statistics, e.g., covariance matrices, need to be estimated for highly nonstationary signals in the case of robot (head) movements.

A data-dependent approach for spatial sound source separation is given by the geometric source separation (GSS) [40]. Unlike the linearly constrained minimum variance (LCMV) beamformer, which minimizes the output power subject to a distortionless constraint for the target and additional constraints for interferers, GSS minimizes the cross-talk explicitly which leads to a faster adaptation [40]. An efficient realization of this approach is presented in [37] for robot audition (online GSS) as well as [38].

A recent, more general framework, which extends the LCMV concept to higher order statistics and uses ICA for a continuous estimation of noise and suppression of multiple interferers, has been proposed in [41,42]. For the robot application, this approach can be implemented based on multiple two-channel BSS units [43] to allow for an extraction of multiple target sources [44].

A benefit of signal enhancement by data-independent fixed beamformers is their low computational complexity since the beamformer coefficients can be calculated in advance for different DOAs. However, the design of a beamformer is usually carried out by assuming a free-field propagation of sound waves, which is inappropriate for robot audition due to sound scattering effects at the robot head and torso. In [9], a minimum variance distortion-less response (MVDR) beamformer design is proposed, where the HRTFs of a robot are used instead of a steering vector based on the free-field assumption. This HRTF-based beamformer is used as pre-processing for a subsequent blind source separation (BSS) system to reduce the reverberation and background noise. The evaluation of this approach reveals that this pre-processing step leads to a significant enhancement of the signal quality for the BSS [9].

In [39], the robust least-squares frequency-invariant (RLSFI) beamformer design of [45] has been extended by incorporating HRTFs of a robot head as steering vectors into the beamformer design to accounts for the sound scattering of a robot’s head. An evaluation of this HRTF-based RLSFI beamformer design for the NAO robot head with five microphones has shown that a significantly better speech quality and lower WER can be achieved in comparison to the original free-field-based design as long as the HRTFs match the position of the target source [46]. An extension of the HRTF-based RLSFI beamformer design to the concept of polynomial beamforming is presented in [47], which allows for a flexible steering of the main beam without significant performance loss. In addition, the HRTF-based RLSFI beamformer design [39] has been extended such that the beamformer response can be controlled for all directions on a sphere surrounding the humanoid robot [48].

As suggested before, the almost spherical shape of a humanoid robot motivates to perform the beamforming in the SH-domain. The SH transformation of the sound pressure at the head microphone can be computed by using a boundary-element model for the robot head [49]. Based on this, well-known beamformers such as the maximum directivity beamformer or the delay-and-sum beamformer can
be implemented in the SH-domain [50]. To address the spatial aliasing problem for spherical arrays [12], a new general framework has been developed which can also be applied to robot heads [13].

The single-channel output signal of the spatial filtering system can be further enhanced by post-filtering. The needed noise power spectral density (PSD) is usually estimated by the input signals of the spatial filter. In [37] and [51], a post-filter is proposed whose filter weights are calculated by the MMSE amplitude estimator of [52]. The needed noise PSD is estimated by assuming that the transient components of the corrupting sources are caused by the leakage from other channels in the process of GSS. An evaluation on the SIG2 robot platform has revealed that this post-filtering approach achieves a significant reduction of the WER [51].

A humanoid robot needs a system for AEC such that it can listen to a person while speaking at the same time to allow for a so-called “barge-in”. Most approaches for robot audition are based on a combination of spatial filtering and AEC, as already investigated in [53]. In [54], the AEC is performed on the input signals of a generalized sidelobe canceler (GSC) and the adaptation of the AEC filters is controlled by a double-talk detection, which considers the ratio of the PSDs of beamformer output and echo signal. In [33], the AEC is realized by means of ICA. Recently, it has been shown in [55] that a combination of GSC and AEC, where the AEC filter is operated in parallel to the interference canceler of the GSC according to [56], can also be successfully employed for robot audition.

6. CONCLUSIONS & OUTLOOK
The development of multichannel systems and algorithms for robot audition has received increased research interest in the last decades. The needed microphones are usually mounted to the robot head whose almost spherical shape motivates the use of SH-domain processing for spatial filtering and target source localization and tracking, if a large number of microphones is available. The optimal microphone positions can be found by numerical optimization (maximizing the effective rank of the GHRTF-matrix). The head microphone array might be extended by microphones integrated into the limbs of the body of the robot to increase the array aperture. A major challenge of this approach is to account for the varying sensor spacings due to robot movements.

Techniques for A-SLAM allow to localize the moving array and to infer the missing source-sensor range from the estimated DOAs. Such approaches have still a rather high computational complexity, but show promise for providing sophisticated acoustic awareness for robots in the future. The ego-noise reduction is usually performed by exploiting a priori knowledge about the specific structure of the noise sources and by incorporating information about the motor states. Recent works suggest that it is beneficial to consider also the relative phase of the ego-noise components in the multichannel recordings. Advanced techniques for dereverberation and spatial filtering can also be employed for robot audition, where the HRTFs of the robot head should be considered in the design of such systems. The adaptive AEC, which is needed to allow for a “barge-in”, is usually designed jointly with the spatial filtering to ensure a fast convergence.

A promising direction for future robot audition systems is to benefit from external sensors, which may be provided, e.g., by all kinds of voice communication devices, smart home environments or other robots.

Acknowledgment
The authors would like to thank Hendrik Barfuss, Alexander Schmidt, Christine Evers and all other coworkers in the EARS project for their contributions which form the background of this paper.

7. REFERENCES


