Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 34 (2015) 129–153

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Robust speaker localization for real-world robots☆

Georgios Athanasopoulos[a,*], Werner Verhelst[a,b], Hichem Sahli[a,c]

[a] Vrije Universiteit Brussel (VUB), Department of Electronics and Informatics (ETRO), Pleinlaan 2, 1050 Brussels, Belgium
[b] iMinds, Department of Future Media and Imaging, Gaston Crommenlaan 8, 9050 Ghent, Belgium
[c] Interuniversity Microelectronics Center (IMEC), Kapeldreef 75, 3001 Leuven, Belgium

Received 23 May 2014; received in revised form 15 February 2015; accepted 25 March 2015
Available online 2 April 2015

Abstract

Autonomous human–robot interaction ultimately requires an artificial audition module that allows the robot to process and interpret a combination of verbal and non-verbal auditory inputs. A key component of such a module is the acoustic localization. The acoustic localization not only enables the robot to simultaneously localize multiple persons and auditory events of interest in the environment, but also provides input to auditory tasks such as speech enhancement and speech recognition. The use of microphone arrays in robots is an efficient and commonly applied approach to the localization problem. In this paper, moving away from simulated environments, we look at the acoustic localization under real-world conditions and limitations. Our approach proposes a series of enhancements, taking into account the imperfect frequency response of the array microphones and addressing the influence of the robot's shape and surface material. Motivated by the importance of the signal's phase information, we introduce a novel pre-processing step for enhancing the acoustic localization. Results show that the proposed approach improves the localization performance in joint noisy and reverberant conditions and allows a humanoid robot to locate multiple speakers in a real-world environment.

Keywords: Microphone arrays; Acoustic localization; Time delay estimation; Steered response power; Phase spectrum enhancement

1. Introduction

In human–robot Interaction (HRI) auditory information can contribute to resolve complex problems such as the focus of attention, activity recognition, etc. It has been observed that both adults and children do not perceive humanoid robots as a mechatronic device but attribute to them characteristics similar to those attributed to living organisms (Reeves and Nass, 1998). HRI is therefore expected to employ mechanisms similar to those of humans interacting with each other. This renders acoustic localization a fundamental element of the context aware HRI.

Acoustic localization is defined as the determination of the direction of each active sound source of interest in relation to a reference point, usually the robot itself. Over the years, various general purpose acoustic localization methods have been proposed (Brandstein and Ward, 2001). In robotics, acoustic localization methods range from simulating the

☆ This paper has been recommended for acceptance by R.K. Moore.

[*] Corresponding author. Tel.: +32 2 629 3955.
E-mail addresses: gathanas@etro.vub.ac.be (G. Athanasopoulos), wverhels@etro.vub.ac.be (W. Verhelst), hsahli@etro.vub.ac.be (H. Sahli).

http://dx.doi.org/10.1016/j.csl.2015.03.009

Fig. 1. NAO, a real-world humanoid robotic platform interacting with children under the ALIZ-E project.

binaural cues of human hearing (Ferreira et al., 2009; Dávila-Chacón et al., 2012) and developing artificial ears (pinnae) (Hwang et al., 2011; Kumon and Noda, 2011) to utilizing several spatially separated microphones (microphone array). Due to their efficiency and robustness, microphone arrays are being increasingly used. Typical microphone array sizes for robots vary from two-four (small) to eight or more sensors (large).

Acoustic localization using a microphone array usually relies on time delay estimation (TDE) of the audio signal at the different microphones of the array. Different techniques for estimating the time delay can be applied (Mumolo et al., 2002; Trifa et al., 2007). The Generalized Cross-Correlation (GCC) (Knapp and Carter, 1976) is most commonly used due to its robustness and computational efficiency. TDE, however, is generally limited to a single dominant sound source. In scenarios where multiple overlapping sources need to be localized, steered-beamforming methods are employed (Valin et al., 2007; Breuer et al., 2012). Recent frameworks for localizing multiple sources have been proposed in Oualil et al. (2013) and Brutti and Nesta (2013).

As robotics technology advances, robots are expected to operate in various types of environments such as houses, schools, hospitals, etc. Fig. 1 shows an out of the lab HRI, conducted with the commercially available robot NAO (Aldebaran's, 2015) in the context of the ALIZ-E[1] project. These various environments exhibit different acoustic properties, such as the presence of ambient noise or reverberation. These properties are not always known in advance and may change dynamically. Therefore, the acoustic localization system of robots must be capable of coping with these conditions. Typically, the effect of noise and reverberation is addressed through spectral weighting of the signals coming from the microphones of the array (Abutalebi and Momenzadeh, 2011; Knapp and Carter, 1976; Rui and Florencio, 2004; Valin et al., 2007).

On the other hand, in real-world robots such as NAO, the sound wave is diffracted along the surface of the robot before reaching the microphones of the array. Different methods have been proposed for addressing the influence of the robot's shape in the acoustic localization. The Head Related Transfer Function (HRTF) can be used for modeling the diffraction of sound waves by the robot (Keyrouz, 2008). However, measuring the HRTF is not a trivial process. Besides, the HRTF depends on the room acoustics, and therefore the measurements might have to be repeated each time the robot operates in a different environment. As a simpler alternative, the Auditory Epipolar Geometry (Nakadai et al., 2001; Kim et al., 2011) and Scattering Theory (Nakadai et al., 2003) have been used in different systems. Both methods assume that the array is installed on a spherically shaped robot head and therefore their accuracy decreases when this assumption is not satisfied. On a different approach, an appropriate geometry for the microphone array can be selected so that at least one microphone pair can be used to refine the location estimate (Kwon et al., 2007).

Moving away from controlled conditions, in this paper we present an acoustic localization system designed for real-world robots and environments. Although localizing multiple speakers is our primary goal, we also cover aspects related

[1] ALIZ-E aims at designing and developing long-term adaptive social interaction between robots and child users (8–11 years old). Under ALIZ-E, HRI is deployed in real-world environments and settings (Belpaeme et al., 2012).

to TDE based techniques since they are still used in many robotic platforms due to their simplicity and computational efficiency. The main contributions of this paper are as follows:

- The role of the microphones' characteristics in the acoustic localization is investigated. A novel spectral weighting that accounts for the frequency response of the microphones is proposed. This spectral weighting is suitable for both TDE and steered-beamforming based localization.
- Smoothing of the TDE and removing estimation outliers is an important task for enhancing the acoustic localization performance. A simple, yet efficient method is formulated. The suggested method essentially relies on the properties of the speech signals and their effects on the GCC.
- Spectral weighting functions typically take into account only the magnitude of the spectrum. Motivated by the importance of the phase information in the acoustic localization, we propose a novel pre-processing approach for enhancing the signal's phase spectra. Our approach operates along the time dimension and can be readily combined with any localization algorithm.
- A framework for addressing the robot's shape and surface material influence in the acoustic localization is introduced. The proposed approach is based on a two-dimensional set of pre-measured time delays, followed by parabolic interpolation. Moreover, it can be applied to any array configuration and has the advantage of a simpler measuring methodology, when compared to the conventional HRTF based approaches.

Along with discussing our localization system, throughout this paper we also attempt to provide a short, yet comprehensive, overview of complementary approaches pertinent to addressing multiple speaker acoustic localization for robots under real-world conditions and limitations.

The remainder of this paper is organized as follows. In Section 2, we formulate the theory of TDE with emphasis on the effect of the signal's phase information and the role of spectral weighting. We look at the role of the microphone's frequency response in the spectral weighting and introduce a new TDE temporal smoothing approach. We also introduce a novel pre-processing stage which aims at enhancing the phase spectra of the array signals that are used in acoustic localization. In Section 3, we highlight the influence of the robot's shape and surface material on the acoustic localization and introduce our approach that is based on a pre-measured set of time delays, followed by parabolic interpolation. In Section 4, we formulate the theory of steered-beamforming for multiple speaker localization and discuss the specificities of its implementation in our system. Section 5 discusses the experimental evaluation both of the individual components and the overall localization system using the NAO robot under real environmental data and conditions. Finally, Section 6 concludes this paper and presents an outlook.

2. Time delay estimation

2.1. Background

2.1.1. Signal model

We consider a microphone array of M sensors and denote each microphone pair of the array as (m_i, m_j). We assume $s(n)$ to be the unknown source signal in the far-field. We also accept the room to be linear and time invariant. The signal captured by microphone m_i can then be expressed as the linear convolution of the source signal and the room's impulse response

$$x_i(n) = h_i(n) * s(n) + n_i(n) \tag{1}$$

where $n_i(n)$ is an additive noise signal, the finite length sequence $h_i(n)$ is the room's impulse response for the given source and microphone positions and n is the time index in samples.

Observing Eq. (1) in the frequency domain, we note that for each analysis frame, the available signal spectrum $X_i(k)$ at frequency bin k is approximately (due to windowing) the result of summing two complex-valued spectra: the reverberant source signal $S'_i(k) \approx H_i(k)S(k)$ and the additive noise $N_i(k)$

$$X_i(k) = |S'_i(k)|e^{j\angle S'_i(k)} + |N_i(k)|e^{j\angle N_i(k)} \tag{2}$$

where $X_i(k)$, $H_i(k)$, $S(k)$, $S_i'(k)$, $N_i(k)$ are the short-time Fourier transforms (STFT) of $x_i(n)$, $h_i(n)$, $s(n)$, $s_i'(n)$, $n_i(n)$ respectively, and $s_i'(n) = h_i(n) * s(n)$.

2.1.2. Generalized Cross-Correlation

The TDE is concerned with the estimation of the relative time difference of arrival of the source's signal between the different spatially separated sensors of the array. A well established method for calculating the TDE for a microphone pair is to locate the highest peak of the GCC function (Knapp and Carter, 1976)

$$\begin{aligned} r_{i,j}(n) &= \sum_{k=0}^{K-1} \Psi_{i,j}(k) X_i(k) X_j^*(k) e^{\dfrac{j2\pi nk}{K}} \\ &= \sum_{k=0}^{K-1} \Psi_{i,j}(k) |X_i(k)||X_j(k)| e^{j(\angle X_i(k) - \angle X_j(k))} e^{\dfrac{j2\pi nk}{K}} \end{aligned} \tag{3}$$

where $X_i(k)$, $X_j(k)$ are the STFTs of the signals $x_i(n)$, $x_j(n)$ from microphones m_i and m_j respectively, k is the frequency bin index and $(^*)$ denotes complex conjugation. The TDE $\hat{\tau}_{i,j}$ can be found as

$$\hat{\tau}_{i,j} = \arg\max_n r_{i,j}(n). \tag{4}$$

In Eq. (3), $\Psi_{i,j}(k)$ is a properly selected frequency dependent filter, known as a weighting function, that can be interpreted as the cross power spectrum of two filters $G_i(k)$ and $G_j(k)$ that are respectively applied to signals $X_i(k)$ and $X_j(k)$

$$\Psi_{i,j}(k) = G_i(k) G_j^*(k). \tag{5}$$

The selection of a suitable weighting function depends on several criteria, such as the presence of ambient noise, reverberation and even the properties of the signals (e.g., periodicity).

2.1.3. Spectral weighting functions

In order to examine the role of $\Psi_{i,j}(k)$ in deemphasizing frequency components that are expected to contribute unreliable information to the GCC function, let us first assume that no reverberation, nor noise is present and thus let us consider only the direct signal from the sound source to each microphone m_i. Eq. (1) then reduces to $x_i(n) = a_i s(n - \tau_i)$, where a_i is an attenuation factor and τ_i is the propagation time (in samples) of the sound wave to each microphone of the array. Hence, the signals $x_i(n)$, $x_j(n)$ are expected to be time shifted versions of each other with different attenuations. As can be observed from Eq. (3), the phase differences between $X_i(k)$ and $X_j(k)$ are linear and thus result in a dominant peak in the GCC function whose location corresponds to the signals' time delay. Fig. 2(a) shows a 32 ms segment of clean speech (recorded by one of the microphones of the pair), with the corresponding GCC function (for $\Psi_{i,j}(k) = 1$) being depicted in Fig. 2(b). As anticipated, the GCC function has a major peak, while secondary peaks correspond to the periodicity of the voiced speech signal.

Due to the linear phase differences, even if the magnitude spectrum is corrupted by noise, a major peak will appear at the correct time index. However, from Eq. (2) it is obvious that noise (and reverberation) affects the phase spectrum as well, resulting in non-linear phase differences. This can be visualized as follows. Using the previous signals of Fig. 2(a), we randomly modify the phase spectrum by adding random phase values in the range $[-\pi/4, \pi/4]$, while the original spectral magnitude is preserved. This is done for each microphone separately. The GCC function for the phase-modified speech segment is shown in Fig. 2(c). It is apparent that the random phase differences give rise to several peaks. This indicates that the GCC performance is sensitive to errors introduced in the phase spectra of the signals, whilst magnitude changes alone, have no impact on the major peak's location.

From the above, it becomes evident that in real-world conditions where both the magnitude and the phase differences between $X_i(k)$ and $X_j(k)$ are affected, the time domain signals can be barely considered as time shifted versions of each other. As a result, a peak will not necessarily exist at the location that corresponds to the actual time delay. Moreover, spurious spikes can rise in the GCC function. These effects will introduce errors in the TDE. Hence, the role of $\Psi_{i,j}(k)$ is to avoid unreliable information in $r_{i,j}(n)$ by reducing the contribution of noise and reverberation corrupted frequency bins.

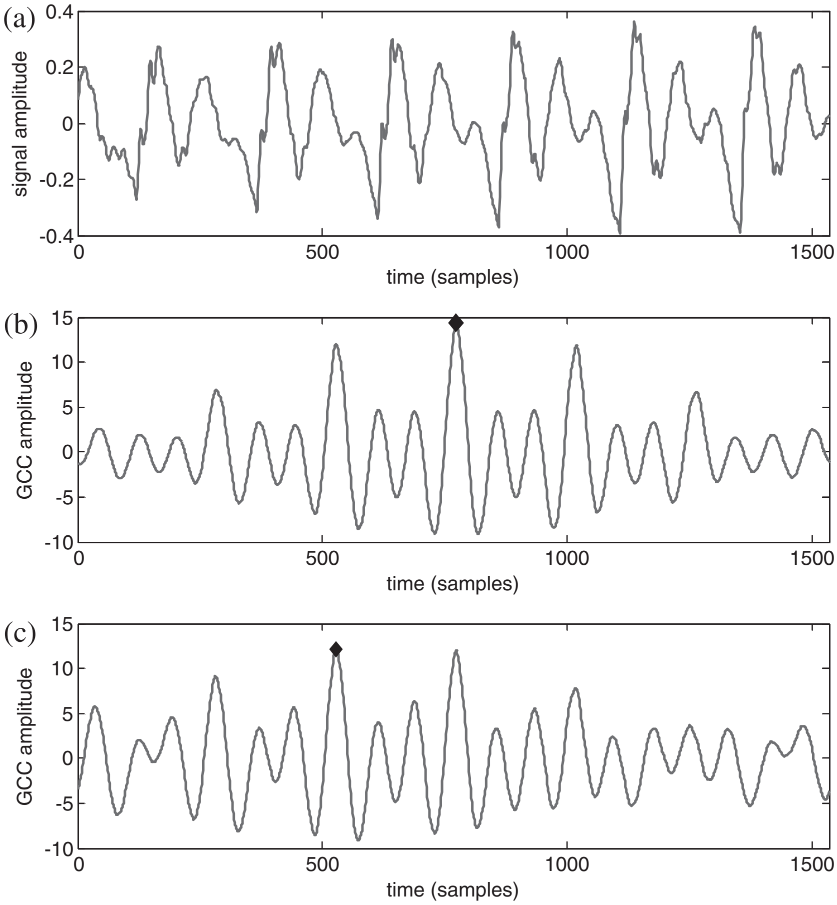

Fig. 2. A segment of (a) voiced clean speech, (b) the corresponding GCC function, and (c) the GCC function when the speech phase spectra are randomly altered. The peak of each GCC function is marked with ♦.

Choosing the most appropriate weighting function $\Psi_{i,j}(k)$ is of great importance for the GCC performance in practical implementations. Various functions have been proposed in literature. Among the most commonly used are the maximum likelihood (ML) and the phase transform (PHAT) (Knapp and Carter, 1976), defined by the following equations respectively

$$\Psi_{\text{ML}i,j}(k) = \frac{|X_i(k)||X_j(k)|}{|\hat{N}_i(k)|^2|X_j(k)|^2 + |\hat{N}_j(k)|^2|X_i(k)|^2} \tag{6}$$

$$\Psi_{\text{PHAT}i,j}(k) = \frac{1}{|X_i(k)||X_j(k)|} \tag{7}$$

with $|\hat{N}_i(k)|$, $|\hat{N}_j(k)|$ the spectral magnitudes of the additive noise estimated at microphones m_i, m_j.

The performance of ML, although optimal for high levels of uncorrelated noise, degrades when reverberation is present and when the noise picked up by the microphones is correlated, even if the signals have very high signal-to-noise ratios (SNR) (Chen et al., 2006). The PHAT weighting, on the other hand, is known to perform well under reverberant conditions and in the absence of noise, as it depends only on the room's impulse response (Chen et al., 2006; Zhang et al., 2008). However, in the presence of noise, the spectral components that are dominated by noise receive equal importance in the calculation of the GCC, resulting in a considerable decrease of its performance.

Addressing the effect of both noise and reverberation has been in the epicenter of TDE the last few years. The authors of Rui and Florencio (2004) developed the following maximum likelihood estimator for both noise and reverberation

$$\Psi_{\text{MLR}i,j}(k) = \frac{|X_i(k)||X_j(k)|}{2\rho|X_i(k)|^2|X_j(k)|^2 + (1-\rho)(|\hat{N}_j(k)|^2|X_i(k)|^2 + |\hat{N}_i(k)|^2|X_j(k)|^2)} \tag{8}$$

with $0 \leq \rho \leq 1$ and $|\hat{N}_i(k)|$, $|\hat{N}_j(k)|$ the estimated noise spectral magnitudes, like in Eq. (6). The weighting function of Eq. (8) can be seen as a combination of the ML and PHAT, controlled by the parameter ρ.

In a different approach, the authors of Valin et al. (2007) proposed a weighting function that acts as a mask on the conventional PHAT. This "reliability weighted" weighting function is based on the *a priori* SNR computed using the decision-directed approach (Ephraim and Malah, 1984). Moreover, it uses a noise estimate that includes a reverberation term that puts less weight on frequency bins where a loud sound was recently present.

More recently, noise suppression techniques were introduced to the GCC weighting problem. The authors of Abutalebi and Momenzadeh (2011) proposed a modified PHAT weighting function where the noisy frequency components are deemphasized based on a normalized quantity derived from the de-noised signal spectrum, which is estimated through spectral subtraction.

Several weighting functions, including Eqs. (6) and (8), require the estimation of the noise's spectral magnitudes $|\hat{N}_i(k)|$ from the noisy microphone array signals that are available. Although noise estimation is an important and challenging task in its own right, in the scope of this work we assume that the additive noise is quasi-stationary or slowly time varying (i.e., it changes at a relatively slower rate than speech) so that $|\hat{N}_i(k)|$ can be estimated during silent frames from the STFT of the noise-only signal $N_i(k)$ at frequency index k using the single-pole recursion

$$|\hat{N}_i^m(k)| = \gamma|\hat{N}_i^{m-1}(k)| + (1-\gamma)|N_i^m(k)| \tag{9}$$

with m denoting the frame index, $0 \leq \gamma < 1$ during silent frames and $\gamma = 1$ during speech frames. In practice, depending on the noise conditions of the environment in which the robot operates (e.g., highly non-stationary), an appropriate noise estimation method can be used (Loizou, 2007). It is also worth noting that estimating the silent/speech frames in Eq. (9) is another challenging topic in real environments. Although we do not address voice activity detection in this work, several techniques have been proposed and can be found in the literature, e.g., (Sahidullah and Saha, 2012; Eyben et al., 2013; Ghosh et al., 2011).

2.2. *Proposed enhancements*

2.2.1. *Accounting for the microphones' characteristics*

Microphone array prototypes often use high quality microphones, which typically have a nearly flat frequency response. In real-world robots most microphone arrays consist of low cost sensors. These sensors are expected to exhibit non-flat characteristics and their contribution to TDE is not uniform. For comparison, Fig. 3 shows the smoothed frequency response (Hatziantoniou and Mourjopoulos, 2000) of NAO's front microphone, measured under the same conditions as a higher quality reference microphone.

The signal that is made available by microphone m_i of the robot can be seen as the convolution of the signal arriving at m_i and the impulse response f_i of the microphone. The impulse response f_i (or equivalently, the frequency response F_i) of each microphone can be measured with any suitable technique (Muller and Massarani, 2001).

As opposed to other microphone array applications, dealing with noise is always important in real-world robots. In practice, even if a robot operates within an environment with low ambient noise, a significant amount of self-generated noise, caused by the robot's instruments, is present. Therefore, the observation that the overall SNR of a signal obtained with low cost sensors is not uniformly distributed over frequency, allows us to give more weight to frequency bins where a higher SNR can be expected. Incorporating the above in Eq. (8) results in the following modified spectral weighting, which is an improved version of the approach initially proposed by the authors in Athanasopoulos et al. (2012)

$$\Psi_{\text{MLR-FR}i,j}(k) = g_i(k)g_j(k)\Psi_{\text{MLR}i,j}(k) \tag{10}$$

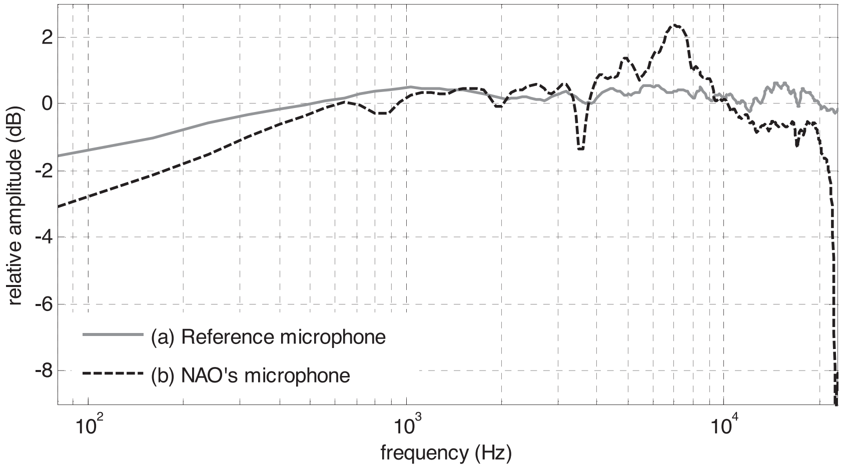

Fig. 3. Frequency response of (a) reference microphone (Earthworks M30), (b) NAO's front microphone.

where the masking functions $g_i(k)$ are defined as

$$g_i(k) = \begin{cases} 1 & \text{if } \dfrac{|F_i(k)|}{|\hat{N}_i(k)|} > R_i \\ \varepsilon & \text{otherwise} \end{cases} \tag{11}$$

with R_i a microphone specific threshold value, $F_i(k)$ the frequency response measured at microphone m_i and $0 < \varepsilon < 1$ a floor value. The masking functions $g_i(k)$, $g_j(k)$ are designed to decrease, according to the value of ε, the contribution in $\Psi_{\mathrm{MLR}i,j}(k)$ of spectral regions with a low ratio between the microphone's frequency response and the noise amplitude.

2.2.2. *TDE temporal smoothing*

In practice, the TDE produces a very high rate of estimates i.e., one per processing frame. This rate is ample for most practical applications. This redundancy allows a variety of smoothing methods to be applied for eliminating estimation outliers, for example moving average or median filtering, Kalman filtering, etc.

On the other hand, speech signals with higher power, like voiced speech segments, result in higher peaks in the GCC function compared to unvoiced or noisy segments. Hereafter, we formulate an approach that makes use of this observation and takes the signal properties into account for smoothing the TDE output. We define the L-frames running average of GCC peak value $p_{i,j}(m)$ for microphones m_i and m_j as

$$p_{\mathrm{AV}i,j}(m) = \frac{1}{L}\sum_{i=0}^{L-1} p_{i,j}(m-i) \tag{12}$$

where m is the frame index and $p_{i,j}(m)$ is given by

$$p_{i,j}(m) = \max(r_{i,j}(n)). \tag{13}$$

The running maximum of the peak values is defined as

$$p_{\mathrm{MAX}i,j}(m) = \max(p_{i,j}(m), \beta p_{\mathrm{MAX}i,j}(m-1)) \tag{14}$$

with $0 < \beta < 1$ a forgetting factor, preventing $p_{\mathrm{MAX}i,j}(m)$ from carrying forward previous large peaks. The TDE $\hat{\tau}_{i,j}$ estimated from Eq. (4) at frame m is considered reliable if the corresponding GCC peak $p_{i,j}(m)$ is above the threshold given by

$$p_{\mathrm{TH}i,j}(m) = p_{\mathrm{AV}i,j}(m) + \zeta(p_{\mathrm{MAX}i,j}(m) - p_{\mathrm{AV}i,j}(m)). \tag{15}$$

Unreliable estimates are omitted from the output. The parameter $0 \leq \zeta < 1$ adjusts the threshold and hence controls the TDE smoothing.

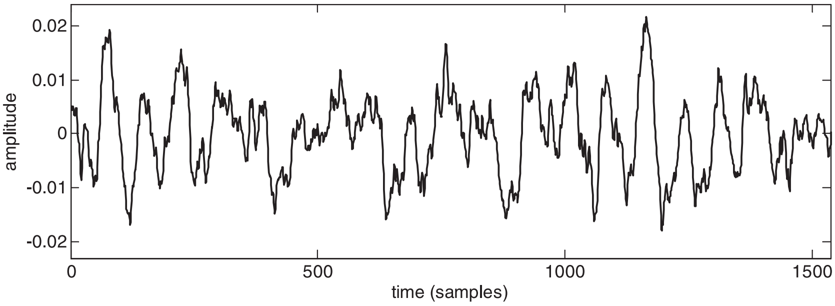

Fig. 4. Sample of internal robot noise.

The above observation remains valid in case of multiple simultaneous speakers: the GCC function will show prominent peaks as long as the received signals contain voiced speech from at least one of the speakers. However, as the number of simultaneous speakers increases, the amplitude of the GCC peaks is expected to decrease. This is due to the energy being spread throughout the GCC function. Besides, secondary peaks corresponding to the non-dominant speakers will be present, as further discussed in Section 4. Hence, the right choice of parameters becomes a very important task. As an example, smaller values of β allow for faster convergence each time a new sound source is present, while a shorter length L of the running average results in a prompter transition between active speakers.

2.2.3. *Signal pre-processing*

The audition system of a robot has to cope with two, generally broadband, types of noise. The signals made available by the microphone array contain internal robot noise that is caused by the robot's instruments (e.g., ventilation fans, gyroscopes, etc.) and ambient noise that is a composite of all variations from many noise sources located in the operational environment of the robot. Internal noise exhibits diffused characteristics caused by the multiple reflections within the robot's cover. Fig. 4 shows a sample of internal robot noise recorded using the NAO robot. Ambient noise is the result of fast pressure variations in the fluid medium, i.e., the air. The origin of these variations spans from mechanical vibrations to aerodynamic forces causing pressure perturbations. Previous studies (Kleinschmidt et al., 2011; Athanasopoulos and Verhelst, 2013) have shown that many real-world noise types can be modeled as a sum of sinusoids closely satisfying the assumption of sinusoidal stationarity across a number of frames.

Motivated by the role of the noisy signals' phase information in TDE, as highlighted in Section 2.1.3, as well as the fact that the GCC weighting functions already take the spectral magnitude into account, hereafter we look at how a more suitable approximation of the phase information can be derived for enhancing the acoustic localization performance.

Let us consider a non-speech period so that the signal $x_i(n)$ from microphone m_i in Eq. (1) contains only the noise $n_i(n)$. The observed noise for a given frame can be expressed as a sum of sinusoids (Kleinschmidt et al., 2011)

$$n_i(n) = \sum_{k=0}^{K-1} a_{i,k} \cos\left(\omega_{i,k}\frac{n}{f_s} + \varphi_{i,k}\right) \tag{16}$$

where K is the number of sinusoids, f_s the sampling frequency and $a_{i,k}$, $\omega_{i,k}$, and $\varphi_{i,k}$ are the sinusoidal amplitude, angular frequency and phase respectively. Fig. 5 shows an example of a noisy signal that is decomposed into three sinusoids.

From the example of Fig. 5, it is evident that although the amplitude and angular frequency of each sinusoid remain unchanged across frames, the phase is different at the start of each frame. Assuming a past frame in which the reference phase is known, the expected phase φ_k of each sinusoid at the start of a second frame λ can be expressed as

$$\varphi_k^{\lambda} = \varphi_k^{\lambda-\delta} + \tau\omega_k\delta \tag{17}$$

where τ is the time delay between the start of the two frames and δ is the number of frames between the reference and the current frame λ.

In practical implementations, the reference phase is calculated during non-speech intervals, and it is assumed that the sinusoids remain stationary or are slowly time-varying over a sufficient number of frames during speech intervals. Furthermore, STFT analysis can be applied in order to estimate the phase, angular frequency and amplitude

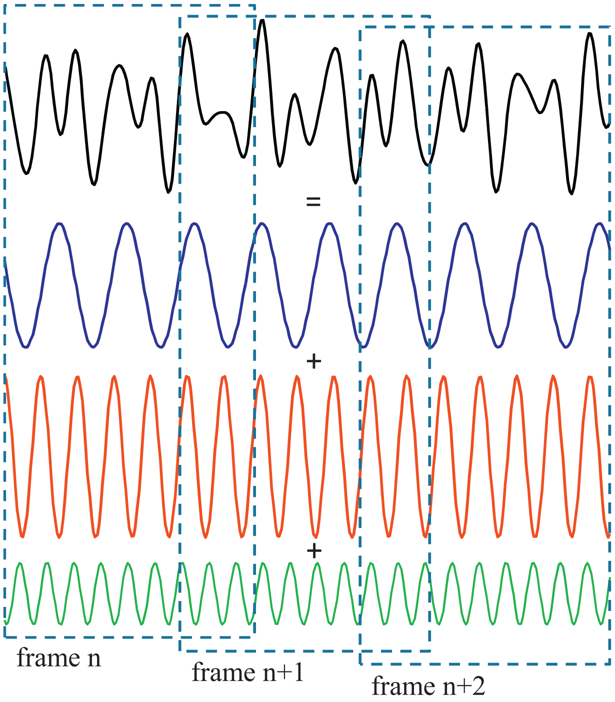

Fig. 5. The sinusoidal model (for $K = 3$ and different initial phases) and the resulting phase shift in frame-to-frame processing.

information of real signals that are modeled as a mixture of many sinusoids. In such case, a sufficiently long discrete Fourier transform (DFT) should be chosen such that the inverse DFT can be assumed to represent the actual sinusoidal decomposition of the noise signal and thus that the effect of spectral smearing is minimal.

The proposed pre-processing algorithm operates in the time domain and makes use of the correlation between frames. In essence, it uses the phase information of the last non-speech frame to approximate the phase information of the noise in subsequent speech frames by adopting the sinusoidal stationarity assumption.

Initially, the noise's phase spectrum is obtained by performing STFT analysis during the last non-speech frame (i.e., reference frame). Through Eq. (17) an estimate of the noise phase spectrum $\angle N_i(k)$ is projected to all subsequent frames that contain speech. Depending on the spectral content of the microphone signals, this operation can be performed selectively for a subset of ξ_i frequency bins of interest, leaving intact the phase of the remaining bins. The number of selected bins can be also seen as the number of sinusoids that are used to model the noise and can be adjusted at each reference frame. In practice, in order to ensure sinusoidal (quasi)stationarity, the maximum number of frames between the reference frame and the current speech frame of Eq. (17) is limited to $\delta_{\max}$.

Having estimated the noise's phase spectrum $\angle N_i(k)$, we can solve Eq. (2) with respect to the equivalent clean (yet reverberant) source signal's phase spectrum $\angle S_i'(k)$ for the current speech frame

$$\angle S_i'(k) = atan2(|X_i(k)|\sin\angle X_i(k) - |N_i(k)|\sin\angle N_i(k), |X_i(k)|\cos\angle X_i(k) - |N_i(k)|\cos\angle N_i(k)) \tag{18}$$

where the function $atan2(y, x)$ expresses the principal value of the arctangent of y/x in the range $(-\pi, \pi]$. The noise's spectral magnitude $|N_i(k)|$ in Eq. (18) is estimated during non-speech periods according to Eq. (9).

It is worth noting that, as shown in Kleinschmidt et al. (2011), the assumption of sinusoidal stationarity can also be applied for the harmonic signal model of clean speech. Therefore, whenever a frame of clean speech is available, the reference phase can be taken as the phase of the most recent frame of clean speech and can then be projected using Eq. (17) to subsequent frames of noisy speech.

The enhanced phase spectra $\angle S_i'(k)$ that are obtained via Eq. (18) are combined with the noisy magnitude spectra to synthesize an enhanced estimate of the source speech signal

$$\hat{S}_i(k) = |X_i(k)|e^{j\angle S_i'(k)}. \tag{19}$$

The phase enhanced signal $\hat{S}_i(k)$ is eventually used as input signal for performing time delay estimation as previously discussed in this section.

3. From time delay estimation to direction estimation

3.1. Overview

3.1.1. Combining TDEs from several pairs

Direction estimation is concerned with the computation of the location of an acoustic source (i.e., azimuth/elevation) from the available TDEs between different microphones, given the array's geometry.

In order to remove ambiguities, TDEs from different microphone pairs of the array are combined for estimating the direction of arrival of the sound-wave. The least-squares method (Brandstein and Ward, 2001) has the advantage that it can be applied to any microphone array configuration. Given $N = (\frac{M}{2})$ unique microphone pairs and under the far-field assumption, the least-squares estimate $\hat{s}$ of the source's direction $\vec{s}$ is found from

$$\hat{s} = \arg\min_{\vec{s}} E(\vec{\tau}, \vec{s}) \tag{20}$$

where the error criterion is defined as

$$E(\vec{\tau}, \vec{s}) = \sum_{l=1}^{N} (\hat{\tau}_l - T_{l,\vec{s}})^2. \tag{21}$$

The reference time delays $T_{l,\vec{s}}$ of pair l depend on the array configuration and are typically calculated analytically for every possible direction of the sound source $\vec{s}$, as further discussed below.

3.1.2. Robot shape and surface material influence

Many robotic platforms used in research, e.g., (Valin et al., 2007; Wu et al., 2009; Breuer et al., 2012), utilize "open-field" microphone arrays, where it is assumed that no obstacles interfere with the propagation of the sound wave between the microphones. In such case, the reference time delays $T_{l,\vec{s}}$ of Eq. (21) at the lth pair can be found by

$$T_{l,\vec{s}} = \frac{d_l \sin(\varphi_{l,\vec{s}})}{c} \tag{22}$$

where d_l is the distance between the microphones of the pair, $\varphi_{l,\vec{s}}$ is the sound wave's incidence angle relative to the axis of the microphone pair, and c is the speed of sound.

In real-world robots, the microphones are mounted on the surface of the robot and the open field array assumption is rarely satisfied. This might not pose a serious problem in robots constructed from softer materials that do not disturb the propagation of sound, as for example in Saldien et al. (2008). Most robots, however, are made from rigid materials. As a result, depending on the acoustic properties of the surface material, the sound wave is scattered and diffracted along the shape of the robot and, as the authors discussed in Athanasopoulos et al. (2012), Eq. (22) as well as the Auditory Epipolar Geometry (Nakadai et al., 2001) are no longer adequate. A more accurate modeling of the robot shape's influence can be achieved using wavefield decomposition. However, this is not a trivial task. Due to the elaborate shape of real-world robots, usually the sound wave's pressure on the robot's surface cannot be derived analytically, while finite elements analysis requires modeling of the robot's shape, as well as the acoustic properties of the robot's surface material, which is not always a straightforward process.

3.2. Proposed pre-measured time delays approach

To address the robot shape's influence, we propose a set of pre-measured time delays initially introduced by the authors in Athanasopoulos et al. (2012), which we here extend to the case of non constant elevation. An advantage of this approach is that a single set of pre-measured time delays can also address cross-channel delays that are not caused by the robot's shape but that are the result of the robot's audio hardware or software (e.g., A/D converters, signal pre-filtering, etc.).

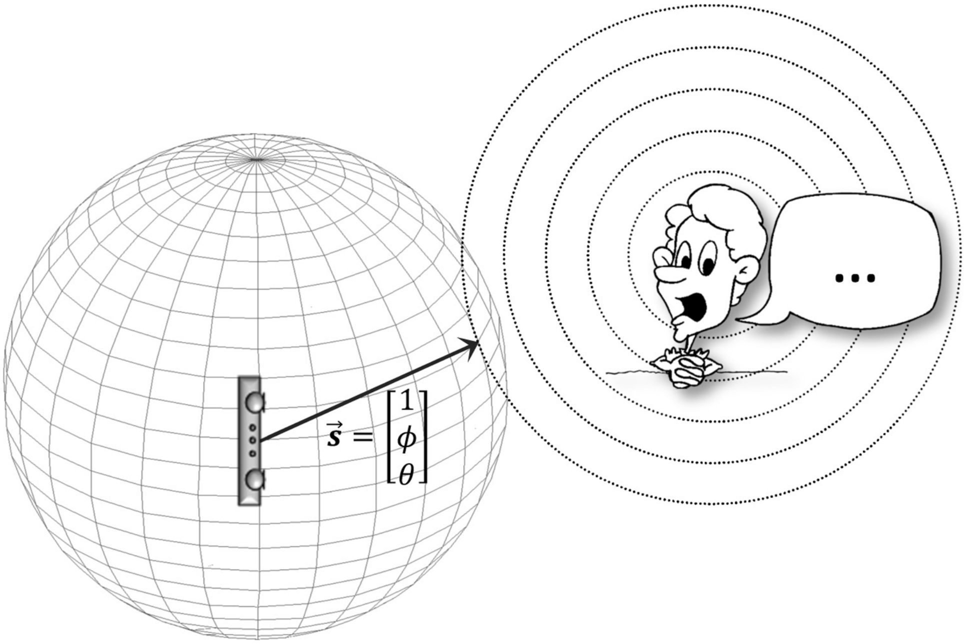
$\vec{s} = \begin{bmatrix} 1 \\ \phi \\ \theta \end{bmatrix}$
...

Fig. 6. The spherical pre-measured time delay direction grid around the robot's microphone array.

In this approach, a pre-measured set of time delays for each microphone pair l is used as the reference time delay (hereafter denoted as $T^{pm}_{l,\vec{s}}$ to differentiate it from the analytically calculated reference). The time delay is measured at discrete equidistant azimuth (ϕ) and elevation (θ) angles. Fig. 6 illustrates the measurement grid, with the microphone array at the center of the coordinate system. Each node of the grid represents a discrete direction (azimuth/elevation). Due to the relatively small size of the microphone array used in robots, usually only the direction (azimuth and/or elevation) is estimated and hence for convenience we choose $\|\vec{s}\| = 1$. By convention we consider $\phi = 0$ in front of the robot, with negative values corresponding to sound sources located on the right side and positive values corresponding to sound sources located on the left of the robot. Similarly, $\theta = 0$ corresponds to sound sources located at the same elevation as the robot's microphone array's horizontal plane, while positive values represent sound sources located higher and negative values represent sound sources located lower than the array's horizontal plane.

In the scope of this work, controlled (semi-anechoic) acoustic conditions are used during the measurement procedure. Pink noise bursts with a duration of 2.5 s are chosen as excitation signals, ensuring a low correlation between subsequent segments. The choice for pink over white noise is made in order to avoid overloading the measurement equipment. The distance between the robot and the sound source (loudspeaker) is set to approximately 1.5 m, sufficiently larger than the robot's microphone array (far-field assumption). A sufficient SNR is attained by setting the sound source's loudness at about 70 dBA (measured at the position of the robot), while the ambient noise level is kept as low as 17 dBA. The data are analyzed in frames of 64 ms with 50% overlap. The TDE is calculated for each microphone pair according to Eq. (4) with $\Psi_{i,j}(k) = 1$. The median of the obtained values for all segments is used to extract a single time delay value for each direction $\vec{s}$. Fig. 7 presents a sample of the measuring directions grid where $\phi \in [-30, 50]$ and $\theta \in [-10, 40]$ with 10° distance between each position.

The localization resolution is related to the distance between the reference time delays of Eq. (22). When a set of pre-measured time delays is used, due to the relatively small number of available values, only an approximation of the sound source direction can be estimated. Therefore, in practice, the localization accuracy depends on the distance between the speakers to be localized. In case of a relatively small distance between speakers, a more refined estimation can be attained by interpolation. Different interpolation techniques can be applied. As far as the acoustic localization accuracy is concerned, given a sufficient set of pre-measured time delays, the majority of these techniques is adequate. Hereafter, we discuss the parabolic interpolation which is computationally simple and does not require numerical approximations or look-up tables.

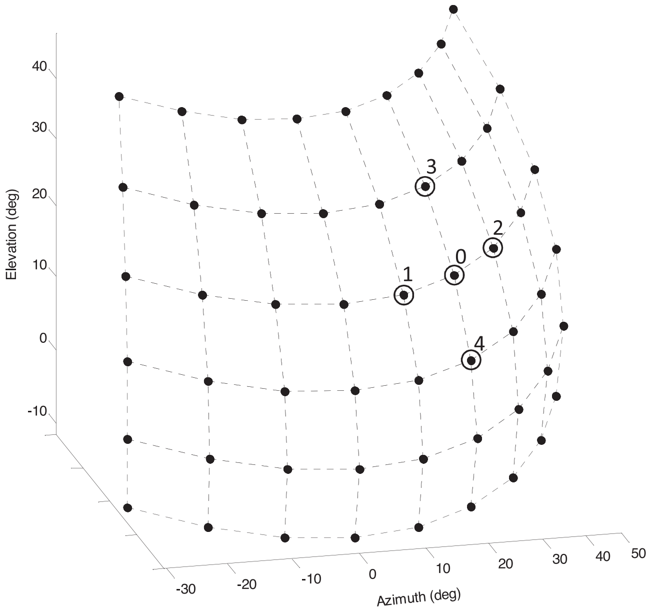

Fig. 7. Sample of pre-measured time delay locations (ϕ, θ) with 10° distance from each other.

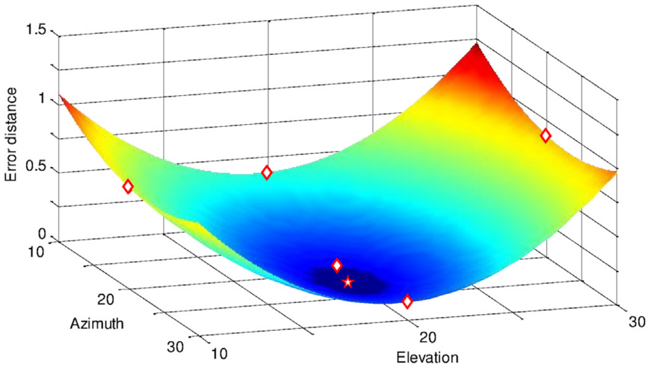

Fig. 8. The fitting bivariate quadratic function of $\widetilde{E}(\phi, \theta)$, with the reference locations marked as ⧫ and its vertex as ⋆.

Let us assume (ϕ_0, θ_0, E_0) to be the best match among the group of known locations (marked as position “0” in Fig. 7), where ϕ_0 and θ_0 are the location’s azimuth and elevation, and E_0 is the corresponding error distance of Eq. (21) between the estimated and the pre-measured time delays. Considering (ϕ_0, θ_0, E_0) and its four neighboring locations (positions “1–4” in Fig. 7), the fitting bivariate quadratic function (paraboloid) can be written as

$$\widetilde{E}(\phi, \theta) = a_4\phi^2 + a_3\theta^2 + a_2\phi + a_1\theta + a_0 \tag{23}$$

where the coefficients a_i can be computed from the five known positions through interpolation (e.g., Lagrange bivariate polynomial interpolation). Fig. 8 illustrates these five measuring directions and the fitted paraboloid.

The refined estimate of the sound source direction can now be estimated by minimizing Eq. (23). The location of the paraboloid’s vertex ($\phi_{\min}$, $\theta_{\min}$) is found by setting the partial derivatives of Eq. (23) equal to zero, resulting in

$$(\phi_{\min}, \theta_{\min}) = \left(-\frac{a_2}{2a_4}, -\frac{a_1}{2a_3}\right). \tag{24}$$

Alternatively, but with a higher computation cost, a fitting bivariate quadratic function of the general form $f(x, y) = a_8x^2y^2 + a_7x^2y + a_6xy^2 + a_5y^2 + a_4x^2 + a_3xy + a_2y + a_1x + a_0$ can be estimated by taking into account additional neighboring locations and applying a least-squares approach. The vertex of the paraboloid can be calculated numerically.

4. Multiple source localization

In realistic scenarios, robots are expected to interact, possibly simultaneously, with several users. A challenging task in such conditions is the localization of multiple, generally overlapping, active sound sources. Conventional TDE techniques like the GCC rarely provide reliable information as one of the sources tends to dominate over the others (Bechler and Kroschel, 2003). On the other hand, steered-beamformer approaches have demonstrated their capacity to fulfill the needs of multiple sources localization. A common technique of this type is the Steered Response Power (SRP) (DiBiase, 2000), also known as Global Coherence Field (GCF) (Omologo and Svaizer, 1994).

The SRP essentially utilizes several GCC measurements (i.e., not just the peak value for each microphone pair) in generating location estimates across the hypothesized location space and consists of two phases. Initially, the steered response power is computed for all possible source locations $s(r, \varphi, \vartheta)$ as

$$SRP(s) = \sum_{q=1}^{N} r_q(\tau_q(s)) \tag{25}$$

where N is, as previously, the number of unique microphone pairs of the array, such that the redundant pairs and autocorrelation terms (which lead to a constant offset) are excluded from the computation of $SRP(s)$. The value τ_q denotes the discrete relative delay between the microphones of a pair q given that the source is located at $s(r, \varphi, \vartheta)$.

Once $SRP(s)$ has been calculated for every possible position, the source location estimate $\hat{s}(\hat{r}, \hat{\varphi}, \hat{\vartheta})$, assuming a single speaker, is determined according to

$$\hat{s} = \arg\max_{s} SRP(s). \tag{26}$$

In the general case above, the search across the location space includes the distance r and therefore the near-field effects might need to be taken into account (Valin et al., 2007). As discussed previously, in our system we consider only the source direction and hence the location estimate reduces to $\hat{s}(\hat{\varphi}, \hat{\vartheta})$.

For handling multiple active sources, multiple local maxima in the $SRP(s)$ need to be extracted. In our system, we select local maxima that are not smaller than $\nu \in [0, 1]$ times the value that corresponds to the dominant source given by Eq. (26). Different approaches have been proposed for detecting the number of speakers. An overview can be found in (Oualil et al., 2012) where an alternative method is also suggested. This method is based on a probabilistic interpretation of the SRP where the GCC function is interpreted as a pdf of the TDEs and is approximated by a Gaussian Mixture Model. The number of speakers can then be detected with a static threshold.

In robots, a relatively small number of microphone pairs is typically available. Furthermore, depending on the size and application of the robot, the location search space can be reduced by excluding improbable source locations (e.g., locations very close to the ground, above or perhaps behind the robot). Hence, the computational load remains reasonable in most cases. Nevertheless, a variety of approaches for reducing the computational load can be found in the literature e.g., in Dmochowski et al. (2007). Recent advances have contributed further to the robustness of the SRP. A known drawback of the SRP is that the localization can become difficult when a source dominates over the rest of the speakers. The method presented in Brutti et al. (2010) attempts to deemphasize the dominant source, after it has been detected, in order to let the other sources stand out. The deemphasis is achieved by applying a mask (notch function) on the GCC function, centered around the time delay associated with the dominant source.

From the definition of SRP, it is evident that the direction resolution is limited by the size of the search grid (cf. Fig. 6). As discussed in Section 3.2, in real-world conditions, utilizing a set of pre-measured reference time delays is more appropriate. Optionally, e.g., when the distance between the speakers to be localized is relatively small, a higher direction accuracy can be obtained by a direction refinement search around the point where a source was found (Valin et al., 2007). In such case, the grid is refined across the φ and ϑ dimensions and the corresponding time delays are calculated from the available pre-measured values through interpolation. The number of points in the refined grid depends on the desired accuracy versus the increase in computational cost. Although we have assumed the spherical rectangular grid of Fig. 6, other grid topologies can be considered such as e.g., a triangular grid sampled at uniform distances along the surface of the sphere (Valin, 2005).

Another important aspect related to the direction accuracy of the SRP method is the resolution of the GCC function. As seen from Eq. (3), for a given microphone array configuration, its resolution depends on the sampling frequency of the signals. Different solutions can be applied (Chen et al., 2006; Tervo and Lokki, 2008). In our system, we resolve

Table 1
Overview of SRP utilizing a set of pre-measured time delays.

Step	Description
1:	For each microphone pair q, calculate $r_q(n)$ up-sampled by f_{up}.
2:	For each hypothesized location s, determine $r_q(\tau_q(s))$ where $\tau_q(s)$ is found from the corresponding pre-measured time delays.
3:	Calculate *SRP(s)* and extract local maxima.
4:	Find the corresponding source location for each local maximum based upon the pre-measured time delays.
5:	If required, perform refined search (steps 2–4) around the point where a source was found using interpolation for computing the refined $\tau_q(s)$ from the pre-measured time delays.
6:	Repeat steps for the next signal segment.

this restriction by up-sampling the time-domain GCC function by a factor f_{up}. For computational efficiency, this is realized in the frequency domain through appropriate zero padding before performing the inverse STFT in Eq. (3) (Athanasopoulos et al., 2012). Table 1 presents an overview of how the pre-measured set of TDEs, introduced in Section 3.2, can be readily combined with the well-established SRP method, and summarizes its implementation in our system.

The performance of the SRP method, as seen from Eq. (25), is related to the robustness of the GCC function. Therefore, topics discussed in Section 2 such as choosing an appropriate weighting function, as well as the proposed enhancements, play an important role in multiple source localization too. Finally, although not in the scope of this work, localization robustness can be further increased by tracking the active sources e.g., by utilizing particle filtering (Valin et al., 2007; Markovic and Petrovic, 2010).

5. Evaluation & discussion

In the first part of this section, we focus on the individual evaluation of each proposed method introduced in the previous sections. The second part discusses the evaluation of the overall acoustic localization system. This evaluation is performed using the NAO robot (version 4.0) and with the intention of avoiding simulations and instead use conditions that are typically present in real-world scenarios.

5.1. Components evaluation

For the evaluation of the components of Section 2.2, we examine their effect in the TDE under various noisy and reverberant acoustic conditions. The TDE is independent of the array's geometry and robot's shape and hence its performance is assessed in terms of *TDE hit-rate*, defined as the percentage of accurate TDE estimates (i.e., exact time delay in samples) over the total number of estimates.

Different reverberant conditions are computed using the image method (Allen and Berkley, 1979). The simulated environment is a rectangular room with dimensions 5 m ×4 m ×3.8 m. The reflection coefficients for all surfaces (walls, ceiling, floor) are uniform and independent of the frequency. The reflection coefficients are chosen so that different reverberation times (RT_{60}) are achieved, spanning from non-reverberant (0 ms) to mild (120 ms), moderate (350 ms) and high (580 ms). Recorded clean natural speech utterances (~20 s long, female speaker) are convolved with the computed impulse responses to obtain the reverberant speech. The broadband noise signal is created by mixing two noise components: office noise (i.e., computers, air-conditioner, etc.) and internal robot noise caused by the robot's instruments (i.e., the processor's ventilation fan, gyroscopes, etc.). In order not to favor any of the two noise components in our evaluation, the two independently recorded noise types are mixed at an equal power ratio. The office noise is recorded using an array of omnidirectional microphones. The robot noise is recorded using the embedded microphones of NAO. The noise signal is properly scaled and added to the reverberant speech, resulting in different SNRs. All signals are sampled at 48 kHz.

The detection of non-speech frames is an important task, not only for performing the TDE, but also as discussed previously for estimating the noise spectral magnitude according to Eq. (9). As the voice activity detection is a challenging research topic on its own, in order to decouple the performance of the evaluated methods from that of the voice activity detector, especially in low SNR, the voice activity detection is performed based on the clean signals

Table 2
Parameters used in components evaluation.

Spectral weighting:	$\rho = 0.5$, $\gamma = 0.98$
Microphone masking:	$R_i = 2 \times$ mean value of $\lvert F_i(k)\rvert / \lvert \hat{N}_i(k)\rvert$, $\varepsilon = 0.3$
TDE smoothing:	$L = 20$, $\beta = 0.99$, $\zeta = 0.1$
Phase spectrum enhancement:	ξ_i = all freq. bins, $\delta_{max} = 50$

(i.e., prior to their convolution with the room's impulse response and prior to the noise addition). Therefore, the frames containing speech are identified using a simple energy threshold. Using this common voice activity baseline allows for a fair performance assessment of the different methods under diverse acoustic conditions, as the evaluation is made over the very same number of estimates.

Throughout the components' evaluation, the processing is performed using 32 ms frames with no overlap. During the evaluation of phase spectra pre-processing, a 10% frame shift is applied for the phase projection (Eqs. (17) and (18)). Hann windowing is applied to prevent spectral leakage. As evaluation baseline, we consider the TDE performance when only the weighting function of Eq. (8) is applied (hereafter referred to as the conventional approach). A neutral value for the parameter ρ is chosen so that neither the noisy nor the reverberant conditions are favored. The parameter values are summarized in Table 2. It is worth noting that these values are applied throughout the components' evaluation and no optimization took place for any of the acoustic conditions.

Specifically for the evaluation of the modified spectral weighting of Eqs. (10)–(11), the reverberant speech and ambient noise signals are further convolved with the impulse responses of the robot's microphones prior to mixing them with the robot noise. The impulse response of each microphone of NAO was independently measured using

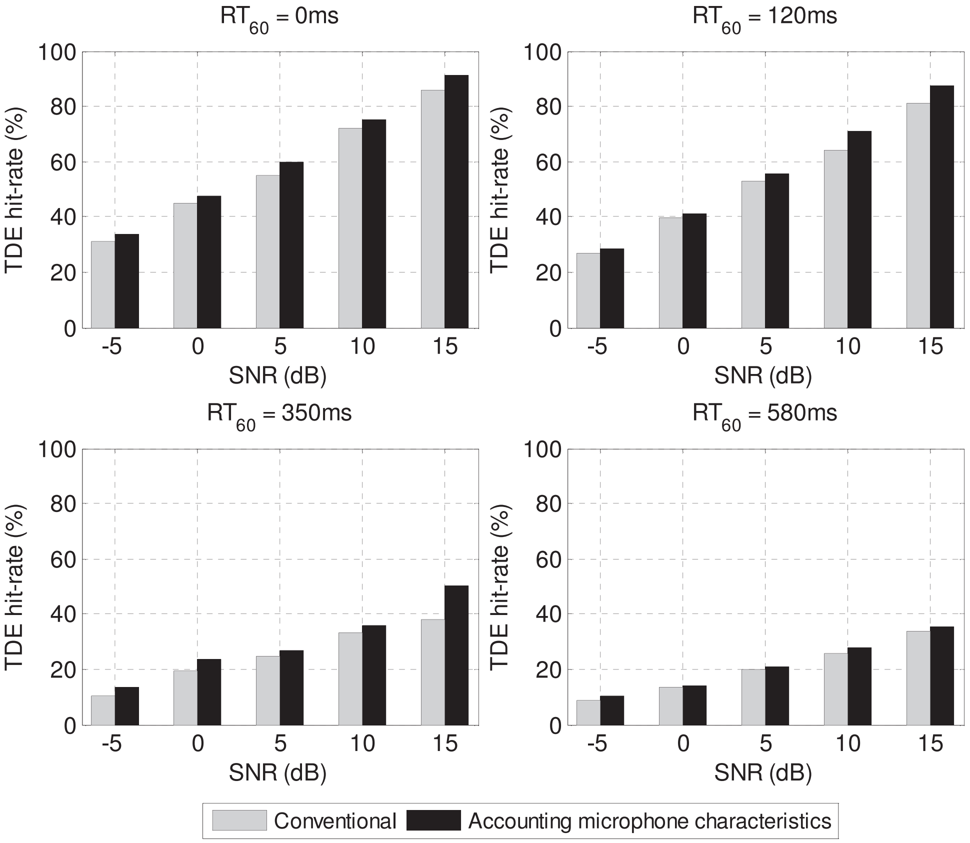

Fig. 9. Evaluation results of the modified weighting function accounting for the non-uniform microphone characteristics under different SNR and reverberation conditions.

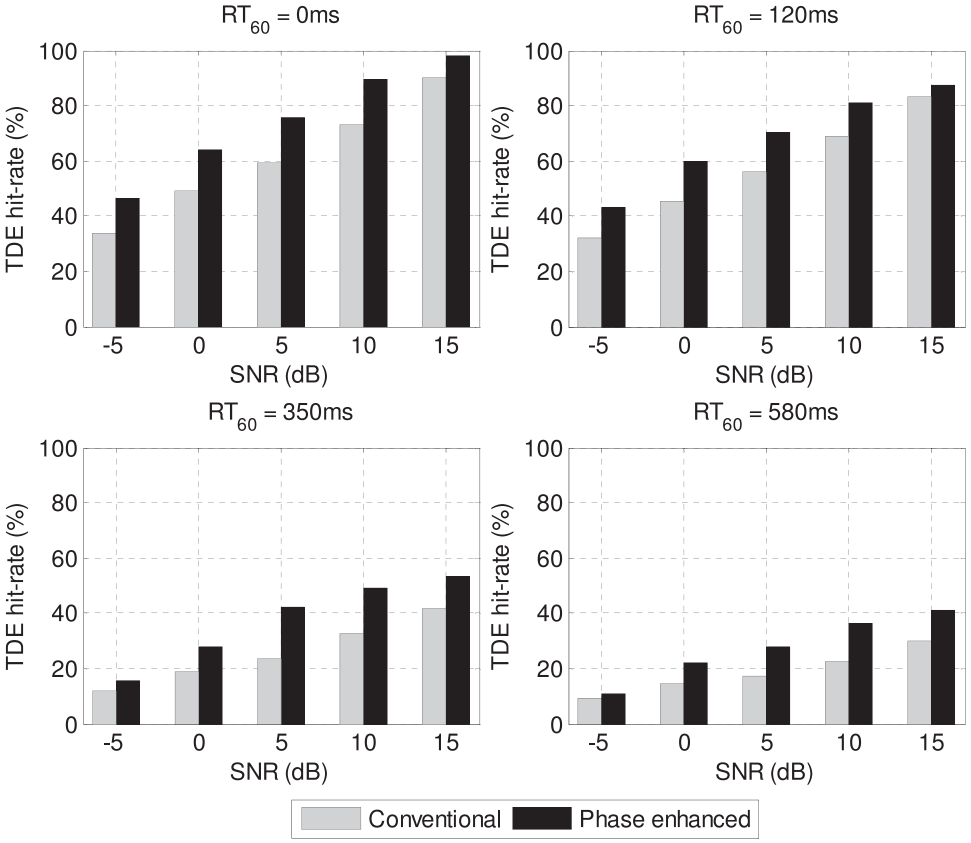

Fig. 10. Evaluation results of the phase spectrum enhancement under different SNR and reverberation conditions.

sweep excitation signals (Muller and Massarani, 2001) under controlled acoustic conditions. Hereafter, we present the aggregated evaluation results over all speech utterances.

Fig. 9 shows the evaluation outcome for the modified spectral weighting. The experiments demonstrate that the proposed approach compensates for the effect of the non-flat frequency response of the microphones. As a result, the corresponding TDE hit-rate is consistently improved by over 3.5% on average with a variance of 0.06% over all acoustic conditions. The drastic decrease of the TDE hit-rate for lower SNR in our experiments is justified by the fact that the bandwidth of the speech signal is significantly smaller than that of the chosen noise signal. In practical applications, appropriate bandpass filtering of the signals would increase the TDE performance. In the general case, however, the sound source of interest might be another broadband source (e.g., music). Hence, in the scope of this work we opt to study the performance of the TDE without band-limiting the signals.

Next, we look at the performance of the proposed smoothing approach, i.e., Eqs. (12)–(15). As a reference, the smoothing effect of a median filter is also considered. The filter length L_{med} is chosen so that the final number of estimates is close to that of the proposed approach. The average smoothing rate of the proposed approach during our experiments is found to be one estimate per 2.81 input estimates, with a variance of 0.14. Therefore, the median filter length is set to $L_{med} = 3$. The evaluation outcome is shown in Fig. 11. The experiments demonstrate the efficiency of the suggested approach, which results in a TDE hit-rate increase of 25.4% on average with a variance of 1.32% compared to the unsmoothed TDE. Contrary to the suggested approach, which demonstrates a very good performance under all acoustic conditions considered, the median filter based smoothing degrades drastically as the number of outliers increases (i.e., higher reverberation and noise levels). On the other hand, and as it is visually demonstrated in the example of Fig. 12, the proposed approach does not maintain a constant output rate. This is due to the nature of the proposed approach that relies on the properties of the speech signals, as discussed in Section 2.2.2.

Lastly, as shown in Fig. 10, the TDE hit-rate significantly improves when the phase spectra pre-processed signals of Eq. (19) are applied. The increase in the amount of accurate TDE estimates is once again consistent over all acoustic

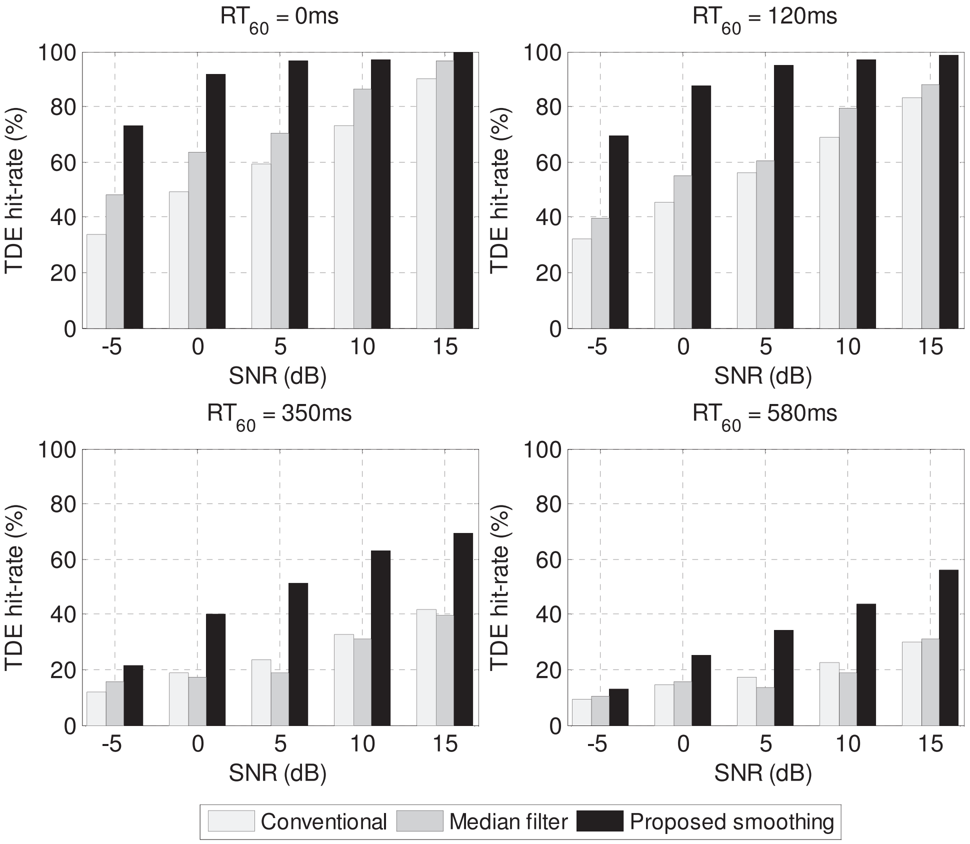

Fig. 11. Evaluation results of the proposed TDE smoothing under different SNR and reverberation conditions.

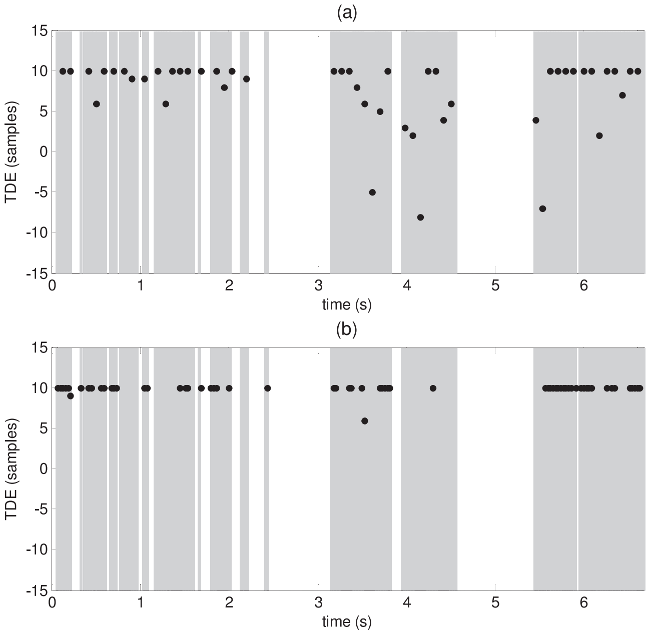

Fig. 12. A segment of TDE output (0 dB SNR, $RT_{60} = 120$ ms) smoothed by (a) median filtering, (b) the proposed approach. The gray shaded areas indicate the presence of speech.

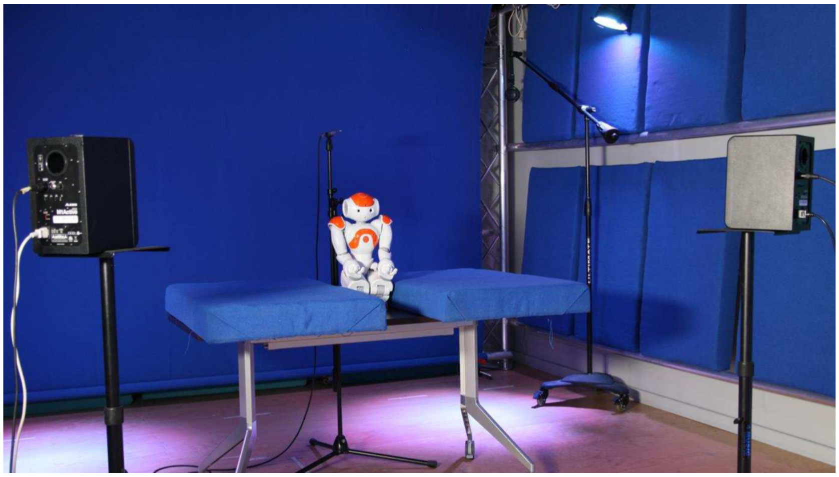

Fig. 13. The evaluation set-up at ETRO Audio-Visual Lab. The removable sound absorbing panels are visible on each side of the robot as well as the studio wall.

conditions. The proposed approach contributes an average TDE hit-rate increase of about 11.5% with a variance of 0.2%.

5.2. System evaluation

In this part of the evaluation, we first look at the performance of the proposed direction estimation approach of Section 3.2, followed by the assessment of the overall acoustic localization system, which we compare with a baseline system. The localization system comprises the GCC-MLR of Eq. (8) and the proposed enhancements of Sections 2.2 and 3.2. Due to the limited processing resources of NAO's on-board computer, the localization system is designed to run on a remote computer as described in Kruijff-korbayová et al. (2012). The baseline system relies on the GCC-PHAT of Eq. (7), a commonly used approach in acoustic localization for robots. The evaluation took place at the ETRO Audio-Visual Lab (Nosey Elephant Studios, 2015), an environment that allows us to approximate real-world conditions (in terms of SNR and reverberation time) in a controlled manner. The robot was placed in the center of the recording room. Two loudspeakers were placed in front of the robot at known locations (i.e., 30° on the left and 45° on the right of the robot, both at 1.5 m distance). This set-up is shown in Fig. 13.

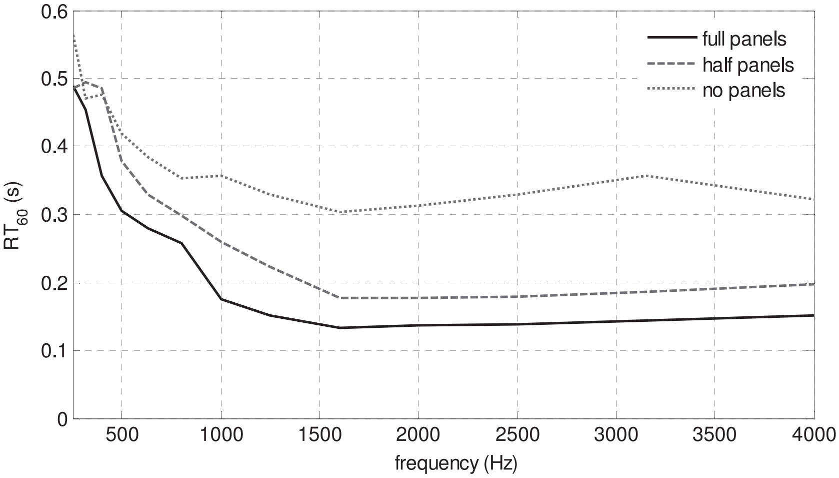

Fig. 14. The different reverberation conditions (RT_{60}) used during the evaluation.

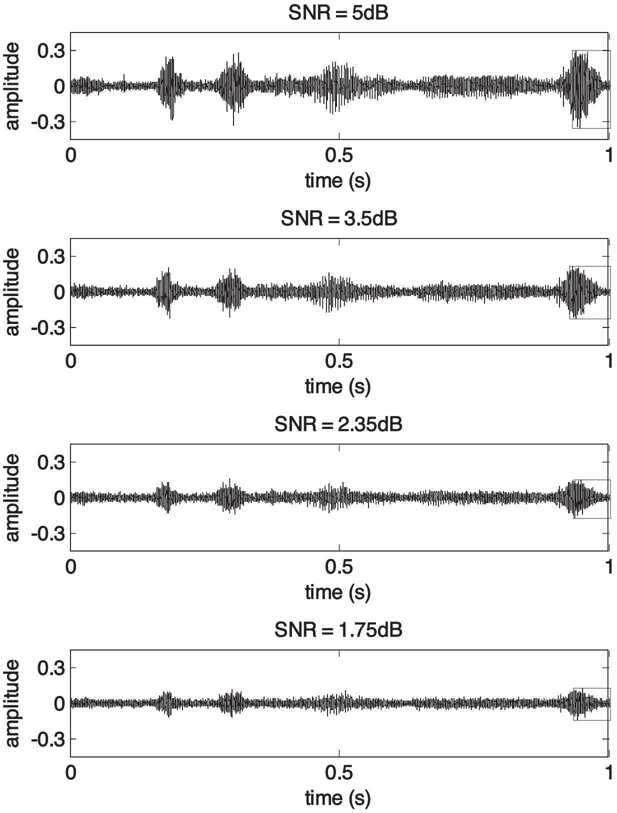

Fig. 15. Samples of captured signals by the robot's right microphone for different SNR conditions (the reference SNR value corresponds to the average over all microphones).

Different reverberation times are achieved by adjusting the number of sound absorbing panels that were placed on the flour, walls and ceiling of the recording room. Three distinct reverberation conditions are considered with RT_{60} at 1 kHz ranging from 170 ms (full panels), to 260 ms (half panels) and 360 ms (no panels). Fig. 14 shows the overall reverberation times measured with the interrupted pink noise method (ISO 3382-2, 2015) at the position where the robot was placed.

Several recorded natural speech utterances (adding up to ∼1.5 min of speech) in English and French, uttered by one female and one male speaker, were played over the loudspeakers, acting as the sound sources to be localized. By adjusting the loudspeakers loudness we obtain signals of different SNR at the robot's microphones. We consider four SNR sets of 5 dB, 3.5 dB, 2.35 dB and 1.75 dB. These values correspond to the average SNR values over all microphones. An example of these four sets, which visually demonstrates the amount of noise in the signals, is shown in Fig. 15.

The processing is performed using frames of 30 ms with 90% overlap for the phase spectrum enhancement and no overlap for computing the TDE, thus resulting in a total of ∼2.5 k location estimates for each source position and each noise/reverberation condition. All signals are sampled at 48 kHz. Hann analysis windowing is used to minimize spectral leakage in the STFT calculation. As discussed in Section 2.2.3, we restrict the noise phase spectrum projection

Table 3
Parameters used in system evaluation.

Spectral weighting:	$\rho = 0.3$, $\gamma = 0.98$
Microphone masking:	$R_i = 2 \times$ mean value of $\lvert F_i(k)\rvert / \lvert \hat{N}_i(k)\rvert$, $\varepsilon = 0.6$
TDE smoothing:	$L = 20$, $\beta = 0.99$, $\zeta = 0.1$
Steered Response Power:	$\nu = 0.85, f_{up} = 8$
Phase spectrum enhancement:	$\delta_{\max} = 50$

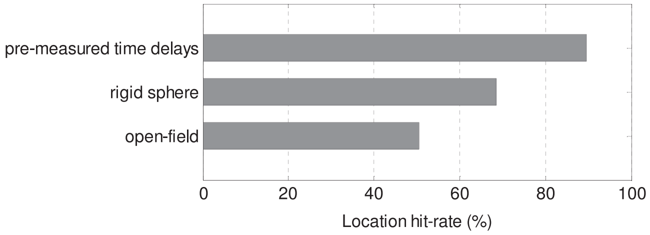

Fig. 16. Comparison of three methods for calculating the location direction estimates from the TDE.

to a subset of frequency bins of interest (ξ_i) by selecting those bins with ratio $|X_i(k)|/|\hat{N}_i(k)|$ higher than a threshold μ_i, thus excluding regions where no speech signal is present or where the noise is dominant. In our evaluation, the value of μ_i is set to 1.5 times the mean value of $|X_i(k)|/|\hat{N}_i(k)|$ across all frequency bins for each processing frame. Table 3 summarizes the remaining parameters used during the evaluation. In real-life situations, the acoustic conditions can dynamically change during human–robot interaction. With this in mind, we again opt not to optimize the parameters

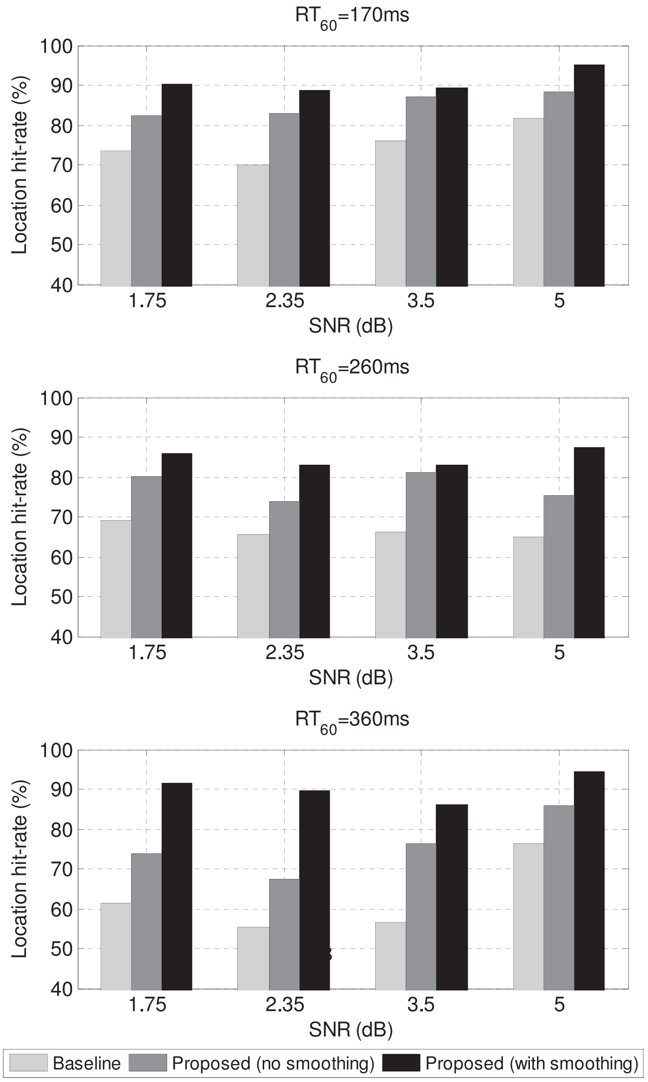

Fig. 17. System evaluation results for different SNR and reverberation conditions.

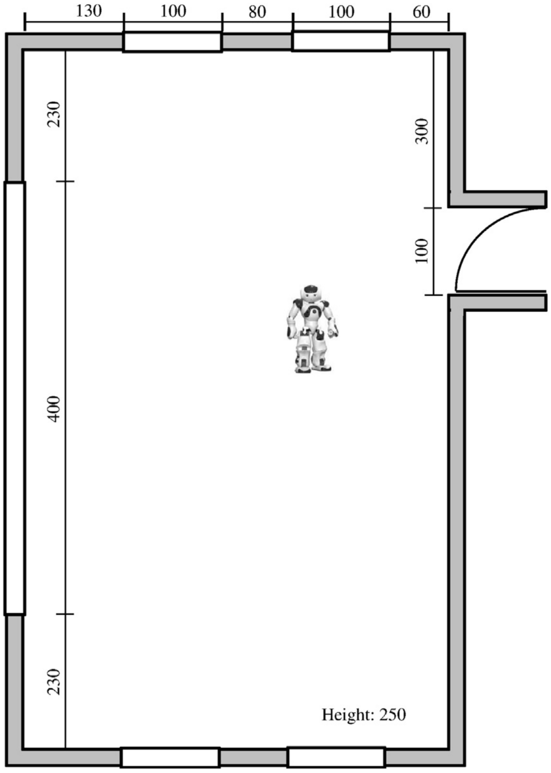

Fig. 18. The dimensions (in cm) of the untreated room used for the evaluation. The robot icon indicates the rough position of the robot.

to each acoustic condition. As our intention is to evaluate the overall system as presented in previous sections, no other pre-processing (e.g., bandpass filtering) nor post-processing (e.g., speaker tracking) is applied for further improving the localization performance.

As the output of the localization system consists of estimations of the source's location, we assess its performance in terms of *location hit-rate*, which is defined as the percentage of accurate location estimates over the total number of estimates. As accurate location estimates, we consider all the estimates within ±3° (target accuracy) from the ground truth.

Before proceeding with the performance evaluation of the overall system, we look at the conversion from TDE to direction estimates as discussed in Section 3. Besides the proposed pre-measured time delays approach, we also consider the open-field array configuration i.e., Eq. (22), as well as the assumption that NAO's head can be approximated as a rigid sphere with the microphones mounted on its surface (Athanasopoulos et al., 2012). Fig. 16 presents the aggregated localization performance of the baseline system for different source positions and SNR. Although the spherical approximation better addresses the influence of the robot's shape than the open-field array, using a set of pre-measured time delays followed by parabolic interpolation clearly outperforms the other approaches, and it is therefore suitable for scenarios that require high resolution speaker localization with a relatively small number of pre-measured references.

Hereafter, for obtaining a fair comparison, we utilize the approach of pre-measured time delays for both the baseline and the proposed system. Moreover, we consider two instances of the proposed system, with and without TDE smoothing i.e., Eqs. (12)–(15). When no smoothing is applied, the output rate is identical to that of the baseline system, hence allowing a direct comparison of the two systems. Fig. 17 presents the results of the evaluation for each SNR and reverberation condition. The location hit-rate percentage for each condition corresponds to the average for all speech utterances and source locations.

The evaluation outcome highlights the robustness of the proposed system under diverse acoustic conditions. From the results presented in Fig. 17, it is evident that the proposed system (when no TDE smoothing is applied) outperforms the baseline system in all considered acoustic conditions. The average improvement on the acoustic localization

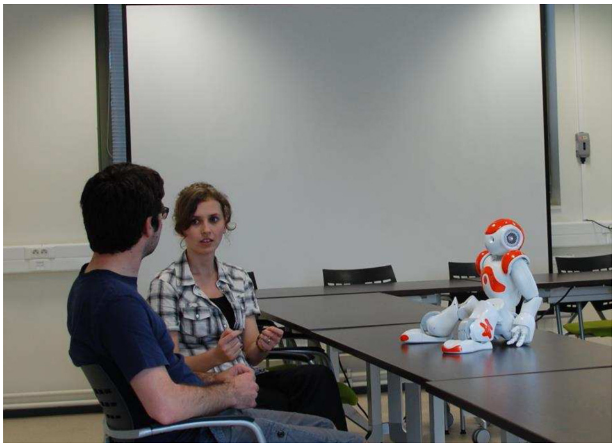

Fig. 19. Two speaker human–robot interaction during the acoustic localization evaluation.

performance is about 11.5% with a variance of 0.1%. Furthermore, when mild TDE smoothing is applied (i.e., a relatively small value for the parameter ζ), an additional improvement by over 9% is achieved with a variance of 0.32%, resulting in an overall location hit-rate improvement by almost 20% compared to the baseline system.

In the final part of this evaluation, we demonstrate the performance of multiple speaker localization in uncontrolled acoustic conditions. For this purpose, the experiment was carried out in large meeting room. The room is not acoustically treated. Typical ambient noise was present, while the glass surface of the windows results in higher reverberation times. The room's layout and dimensions are shown in Fig. 18. For consistency, the same parameter values as previously were used in the system.

One male (A) and one female (B) adult speakers were asked to interact with the robot. An instance of this set-up is shown in Fig. 19. The raw output of the acoustic localization is shown in Fig. 20. In this segment, speaker A starts talking. After approximately 6 s, speaker B also starts talking, and both speakers talk simultaneously for about 11 s. Finally, speaker A stops talking, leaving speaker B to conclude alone.

The illustration of Fig. 20 demonstrates the ability of the proposed system to satisfactorily distinguish and localize two simultaneous speakers in a real environment. The relatively few outliers that are still present are primarily due to the low SNR of the robot's microphones (∼3.65 dB at the front microphone, measured over the total signal when both speakers are active). In practice, outliers could be further suppressed through post processing and source tracking

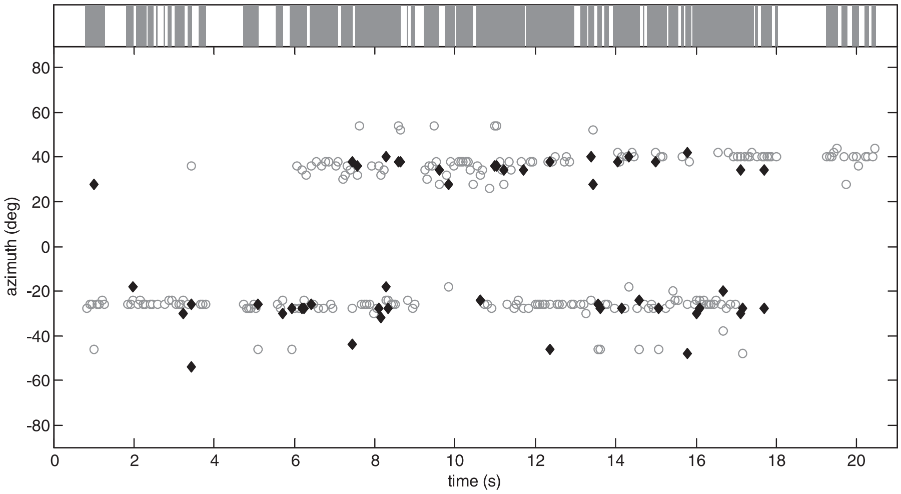

Fig. 20. The raw output of the acoustic localization for two speakers (primary SRP peaks are marked as ○, while secondary peaks are marked as ♦). The gray shaded areas in the bar on the top of the figure indicate the frames containing speech.

(Valin et al., 2007; Markovic and Petrovic, 2010). Overall, the experiments show that the proposed system functions properly on NAO for various types of voices, up to a distance of about 4 meters. This limitation is mainly due to the amount of noise from the robot's processor fan (which is located very close to the microphones), the small dynamic range and the inferior quality of the robot's microphones.

6. Conclusion

In this paper, we have shared our experience in bridging the gap between developing acoustic localization algorithms for robots in simulated or controlled environments, and deploying them in real-life applications. We have also provided a comprehensive overview of the different approaches for addressing multiple speaker acoustic localization for robots under real-world conditions.

Moreover, we have identified the various, often overlooked, limitations that are typically met in real-world applications, such as the influence of the robot's shape and surface material, the non-uniform microphone characteristics, the amount of robot-generated noise present, etc. Motivated by these observations, we proposed a series of novel enhancements for increasing the localization performance.

Finally, we have described a complete system that has been deployed on the commercially available robot NAO. Despite the poor quality of the microphone signals of the robot (i.e., small dynamic range and low SNR), we have demonstrated that our system is capable of performing reliable multiple speaker acoustic localization in the wild and that it can thus contribute to the making of autonomous decisions by the robot in real-life human–robot interaction.

Despite of significant progress over the last few years, many fascinating challenges for acoustic localization in real-world conditions remain on the way to an ideal HRI. Some future directions include the evaluation of the system with multiple, alternating, overlapping and moving speakers, as well as considering tracking techniques for the multiple speakers scenario. In particular, model-based tracking algorithms such as Kalman and Particle filters contain a smoothing function. It would be interesting to compare the different smoothing methods and to assess the model-based smoothing methods in combination with the proposed signal properties-based smoothing. Other scenarios that deserve further research are those considering diverse noise types and environments under which a robot could operate, as well as those addressing the challenge of a walking and gesturing robot, which causes additional varying self-generated noise.

Acknowledgments

This research work was carried out at the Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel. Parts of this research were performed in the context of the EU-FP7 project ALIZ-E (ICT-248116) and in the context of the project SINS (IWT-SBO 130006).

References

Abutalebi, H., Momenzadeh, H., 2011. Performance improvement of TDOA-based speaker localization in joint noisy and reverberant conditions. EURASIP J. Adv. Signal Process., 621390.

Aldebaran's NAO humanoid robotic platform. Available at: http://www.aldebaran.com/en/humanoid-robot/nao-robot

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65, 943–950.

Athanasopoulos, G., Verhelst, W., 2013. A phase-modified approach for TDE-based acoustic localization. In: 14th Annual Conference of the International Speech Communication Association (Interspeech), ISCA, pp. 2890–2894.

Athanasopoulos, G., Dekens, T., Brouckxon, H., Verhelst, W.,2012. The effect of speech denoising algorithms on sound source localization for humanoid robots. In: 11th International Conference on Information Sciences, Signal Processing and their Applications. IEEE, pp. 369–374.

Athanasopoulos, G., Brouckxon, H., Verhelst, W., 2012. Sound source localization for real-world humanoid robots. In: 11th International Conference on Signal Processing, WSEAS, pp. 131–136.

Bechler, D., Kroschel, K., 2003. Considering the second peak in the GCC function for multi-source TDOA estimation with a microphone array. International Workshop on Acoustic Echo and Noise Control, 315–318.

Belpaeme, T., Baxter, P., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., Racioppa, S., Kruijff-Korbayová, I., Athanasopoulos, G., Enescu, V., Looije, R., Neerincx, M., Demiris, Y., Ros-Espinoza, R., Beck, A., Cañamero, L., Hiolle, A., Lewis, M., Baroni, I., Nalin, M., Cosi, P., Paci, G., Tesser, F., Sommavilla, G., Humbert, R., 2012. Multimodal child–robot interaction: building social bonds. J. Human–Robot Interact. 1, 33–53.

Brandstein, M., Ward, D., 2001. Microphone Arrays Signal Processing Techniques and Applications. Springer.

Breuer, T., Macedo, G.R.G., Hartanto, R., Hochgeschwender, N., Holz, D., Hegger, F., Jin, Z., Müller, C., Paulus, J., Reckhaus, M., Álvarez Ruiz, J.A., Plöger, P.G., Kraetzschmar, G.K., 2012. Johnny: an autonomous service robot for domestic environments. J. Intell. Robot. Syst. 66, 245–272.

Brutti, A., Nesta, F., 2013. Tracking of multidimensional TDOA for multiple sources with distributed microphone pairs. Comput. Speech Lang. 27, 660–682.

Brutti, A., Omologo, M., Svaizer, P., 2010. Multiple source localization based on acoustic map de-emphasis. EURASIP J. Audio Speech Music Process.

Chen, J., Benesty, J., Huang, Y., 2006. Time delay estimation in room acoustic environments: an overview. EURASIP J. Appl. Signal Process., 170–189.

Dávila-Chacón, J., Heinrich, S., Liu, J., Wermter, S.,2012. Biomimetic binaural sound source localisation with ego-noise cancellation. In: 22nd International Conference on Artificial Neural Networks, LNCS 7552. Springer, pp. 239–246.

DiBiase, J.H., 2000. A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays (Ph.D. thesis). Brown University.

Dmochowski, J.P., Benesty, J., Affes, S., 2007. A generalized steered response power method for computationally viable source localization. IEEE Trans. Audio Speech Lang. Process. 15, 2510–2526.

Ephraim, Y., Malah, D., 1984. Speech enhancement using minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32, 1109–1121.

Eyben, F., Weninger, F., Squartini, S., Schuller, B., 2013. Real-life voice activity detection with LSTM recurrent neural networks and an application to Hollywood movies. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 483–487.

Ferreira, J.F., Pinho, C., Dias, J., 2009. Implementation and calibration of a Bayesian binaural system for 3D localisation. In: IEEE International Conference on Robotics and Biomimetics, pp. 1722–1727.

Ghosh, P., Tsiartas, A., Narayanan, S., 2011. Robust voice activity detection using long-term signal variability. IEEE Trans. Audio Speech Lang. Process. 19, 600–613.

Hatziantoniou, P.D., Mourjopoulos, J.N., 2000. Generalized fractional-octave smoothing of audio and acoustic responses. J. Audio Eng. Soc. 48, 259–280.

Hwang, S., Park, Y., sik Park, Y., 2011. Sound direction estimation using an artificial ear for robots. Robot. Auton. Syst. 59, 208–217.

Keyrouz, F., 2008. Efficient Binaural Sound Localization for Humanoid Robots and Telepresence Applications (Ph.D. thesis). Technische Universitäat München.

Kim, U.-H., Mizumoto, T., Ogata, T., Okuno, H.G., 2011. Improvement of speaker localization by considering multipath interference of sound wave for binaural robot audition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2910–2915.

Kleinschmidt, T., Sridharan, S., Mason, M., 2011. The use of phase in complex spectrum subtraction for robust speech recognition. Comput. Speech Lang. 25, 585–600.

Knapp, C.H., Carter, G.C., 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. 24, 320–327.

Kruijff-korbayová, I., Cuayáhuitl, H., Kiefer, B., Schröder, M., Cosi, P., Paci, G., Sommavilla, G., Tesser, F., Sahli, H., Athanasopoulos, G., Wang, W., Enescu, V., Verhelst, W., 2012. Spoken language processing in a conversational system for child–robot interaction. In: 3rd Workshop on Child Computer Interaction.

Kumon, M., Noda, Y., 2011. Active soft pinnae for robots. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 112–117.

Kwon, B., Kim, G., Park, Y., 2007. Sound source localization methods with considering of microphone placement in robot platform. In: 16th IEEE International Symposium on Robot and Human Interactive Communication, pp. 127–130.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press, Boca Raton, FL.

Markovic, I., Petrovic, I., 2010. Speaker localization and tracking with a microphone array on a mobile robot using von Mises distribution and particle filtering. Robot. Auton. Syst. 58, 1185–1196.

Measurement of room acoustic parameters, Reverberation time in ordinary rooms, EN ISO 3382-2:2008.

Muller, S., Massarani, P., 2001. Transfer-function measurement with sweeps. J. Audio Eng. Soc. 49, 443–471.

Mumolo, E., Nolich, M., Vercelli, G., 2002. Algorithms for acoustic localization based on microphone array in service robotics. Robot. Auton. Syst. 42, 69–88.

Nakadai, K., Okuno, H., Kitano, H., 2001. Epipolar geometry based sound localization and extraction for humanoid audition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1395–1401.

Nakadai, K., Matsuura, D., Okuno, H., Kitano, H., 2003. Applying scattering theory to robot audition system: robust sound source localization and extraction. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1147–1152.

Nosey Elephant Studios, ETRO Audio-Visual Lab. Available at: http://www.etro.vub.ac.be/research/Nosey_Elephant_Studios

Omologo, M., Svaizer, P., 1994. Acoustic event localization using a cross power-spectrum based technique. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 273–276.

Oualil, Y., Magimai-Doss, M., Faubel, F., Klakow, D., 2012. Joint detection and localization of multiple speakers using a probabilistic interpretation of the steered response power. In: Statistical and Perceptual Audition Workshop.

Oualil, Y., Magimai-Doss, M., Faubel, F., Klakow, D., 2013. A probabilistic framework for multiple speaker localization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3962–3966.

Reeves, B., Nass, C., 1998. The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places. Cambridge University Press.

Rui, Y., Florencio, D., 2004. Time delay estimation in the presence of correlated noise and reverberation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2. IEEE, pp. 133–136.

Sahidullah, M., Saha, G., 2012. Comparison of Speech Activity Detection Techniques for Speaker Recognition.

Saldien, J., Goris, K., Yilmazyildiz, S., Verhelst, W., Lefeber, D., 2008. On the design of the huggable robot probo. J. Phys Agents 2 (2), 3–12.

Tervo, S., Lokki, T., 2008. Interpolation methods for the SRP-PHAT algorithm. In: 11th International Workshop on Acoustic Echo and Noise Control.

Trifa, V., Koene, A., Moren, J., Cheng, G., 2007. Real-time acoustic source localization in noisy environments for human–robot multimodal interaction. In: 16th IEEE International Symposium on Robot and Human Interactive Communication, pp. 393–398.

Valin, J.-M., Michaud, F., Rouat, J., 2007. Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. Robot. Auton. Syst. 55, 216–228.

Valin, J.-M., 2005. Auditory System for a Mobile Robot (Ph.D. thesis). Department of Electrical & Computer Engineering Université de Sherbrooke.

Wu, X., Gong, H., Chen, P., Zhong, Z., Xu, Y., 2009. Surveillance robot utilizing video and audio information. J. Intell. Robot. Syst. 55, 403–421.

Zhang, C., Florencio, D., Zhang, Z., 2008. Why does PHAT work well in low noise, reverberative environments? IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2565–2568.