Author manuscript, published in "IEEE International Conference on Humanoid Robots (2013)"

hal-00861465, version 1 - 12 Sep 2013

Active-Speaker Detection and Localization with Microphones and Cameras Embedded into a Robotic Head

Jan Cech*, Ravi Mittal, Antoine Deleforge, Jordi Sanchez-Riera, Xavier Alameda-Pineda and Radu Horaud
INRIA Grenoble Rhônes-Alpes, France
*Center for Machine Perception, FEE, CTU Prague, Czech Republic

***Abstract*— In this paper we present a method for detecting and localizing an active speaker, i.e., a speaker that emits a sound, through the fusion between visual reconstruction with a stereoscopic camera pair and sound-source localization with several microphones. Both the cameras and the microphones are embedded into the head of a humanoid robot. The proposed statistical fusion model associates 3D faces of potential speakers with 2D sound directions. The paper has two contributions: (i) a method that discretizes the two-dimensional space of all possible sound directions and that accumulates evidence for each direction by estimating the time difference of arrival (TDOA) over all the microphone pairs, such that all the microphones are used simultaneously and symmetrically and (ii) an audio-visual alignment method that maps 3D visual features onto 2D sound directions and onto TDOAs between microphone pairs. This allows to implicitly represent both sensing modalities into a common *audiovisual* coordinate frame. Using simulated as well as real data, we quantitatively assess the robustness of the method against noise and reverberations, and we compare it with several other methods. Finally, we describe a real-time implementation using the proposed technique and with a humanoid head embedding four microphones and two cameras: this enables natural human-robot interactive behavior.**

I. Introduction

The ability of a humanoid robot to robustly detect and localize people that are both seen and heard is an important task which would be very useful in many human-robot interaction scenarii. In this paper we address the challenging problem of robustly and accurately combining auditory and visual sensory data for natural *untethered* interaction: the users are at some distance from the robot and they do not use any kind of wearable sensors.

There have been many approaches to identify an active speaker among a group of people. Typical techniques involve audio and vision as input modalities. Several methods exploit *audio-visual synchrony* to recognize an active speaker by detecting lip motions [1], [2]. Interesting results are achieved with a simple cross-modal correlation method to mark pixels associated with a sound by using only one camera and one microphone [3]. However, these methods require a high-resolution camera and some form of *tethered* interaction: the user must face the robot and its facial motions must be accurately detected in an image.

A large number of methods use multiple microphones and *sound-source localization* (SSL). These methods can be further divided into two groups. The first group performs *late fusion*, namely, the audio and visual features are first extracted and then they are combined, i.e. SSL and visual detection are done separately [4], [2]. The second group performs *early fusion*, i.e. at the feature level [5].

Fig. 1: The proposed method simultaneously detects and localizes an active speaker in an audio-visual scene. The humanoid robot NAO (left) turns its head towards the active speaker marked with a black disk (right). Please consult the supplementary video.

Sound source localization is often based on time difference of arrival (TDOA) between microphones. TDOA cues, also referred to as ITD (Interaural Time Difference) in the case of two microphones, is prominent for human's perception of the sound directions [6]. The main issue of these methods is to find the *correspondence* among the acoustic signals from the microphones. Finding the correspondence means to identify temporal locations of the signals which relate to the same acoustic event. The correspondence then determines the TDOA. This is generally difficult due to an intrinsic ambiguity because of:

1) **Reverberations** are due to an echoic environment thus causing false correspondences and phantom source locations.
2) **Signal dissimilarities** due to various distortions along the acoustic path. A non isotropic environment and non-linear filtering affects quite differently the frequency spectrum of the signals perceived by the microphones.
3) **Narrow band signal.** Source of harmonic tones is difficult to localize due to signal self-similarity.
4) **Low signal-to-noise ratio.** There is non-stationary noise emitted from inside the robot head which strongly affects the quality of the perceived acoustic signals.

In the recent past, several methods have been proposed to overcome these difficulties. A widely used technique is

hal-00861465, version 1 - 12 Sep 2013

cross-correlation between small temporal windows. Several methods employ biologically inspired computation, i.e. the signal is first filtered by a bank of cochlear-like gammatone filters, correlated by bands, and then aggregated again [7], [8]. Other methods overcome the ill-posed localization problem by learning directly the sensorimotor map of a steerable device and of the head-related transfer function (HRTF) [9]. A drawback is that these methods tend to overfit the model to the training environment and have difficulties to work robustly in a different setup. Alternative methods use a microphone array to reduce the ambiguity [10], [11].

We propose a method which finds the correspondence between multiple microphone signals *simultaneously*. The basic idea is to exhaustively sweep the discretized space of possible 2D directions (azimuth and elevation) and to estimate a similarity/consistency statistic of the signal correspondence. This is directly inspired from methods for finding visual correspondences in multi-view stereo [12], [13], which are shown to be more robust than pairwise correspondence search, since the matching ambiguity due to insufficient/repetitive patterns is significantly reduced and the noise is averaged out. We observed similar effects in the case of acoustic correspondences.

The proposed SSL method is similar in spirit to beamforming techniques [14] which try to align signals by steerable delays. A drawback of these methods (including ours) is a high sensitivity to precise calibration and validity of the model which maps a 3D source location onto the space of TDOAs. This is again analogous to high requirements on calibration of cameras in multi-view stereo, e.g. [15]. Nevertheless, our contribution here is in proposing a simple but efficient model which maps 2D source directions to the space of TDOAs and designing a calibration procedure which fits the model using audio-visual correspondences, as well as employing a consistency statistic borrowed from stereo vision. Estimated sound directions are then fed into a statistical model which associates them with 3D faces. The algorithm outputs the posterior probability of the speaking state over the 3D-localized candidate speakers. While this association is similar in spirit to the fusion model in [16], the novel method doesn't need an EM procedure at runtime and yields accurate and discriminative two-dimensional audio-visual localization and detection.

The proposed methodology is implemented on a prototype of the NAO robot equipped with a stereoscopic camera pair and four microphones. This *cheap* humanoid has low quality microphones embedded in the head with unknown, probably complex, HRTF. Note that there is a fan mounted inside the head that causes wide-band non-stationary noise which cannot be easily filtered out. Moreover, the experiments are carried out in a standard room with no special-purpose acoustic properties. The software is implemented onto a dedicated middleware robotic architecture that allows real-time execution. To the best of our knowledge, this is the first time that a 2D active speaker localizer has been implemented on NAO.

The paper is organized as follows. The proposed methods is described in Sec. II. The calibration technique is described in Sec. III. The experiments are presented in Sec. IV. Finally, Sec. V concludes the paper.

II. Active-Speaker Detection

First we describe the method of sound source localization and subsequently its association with visual detections.

A. Sound Source Localization

We consider a robot head equipped with N microphones, which are located in positions $\boldsymbol{m}_i \in \mathbb{R}^3$ for $i \in \{1, \dots, N\}$. Let $\boldsymbol{s} \in \mathbb{R}^3$ be the unknown position of the sound source. A direction where the sound comes from can be estimated from the time difference of arrival between the microphones. Travelling time of the sound between the source $\boldsymbol{s}$ and microphone $\boldsymbol{m}_i$ is denoted by τ_i and the TDOA between two microphones $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ by $\tau_{i,j} = \tau_i - \tau_j$. Therefore we have $\binom{N}{2}$ TDOAs, but obviously only $N-1$ of them are independent. In other words, a sound source location in 3D space fully determines the *joint correspondence* among microphone signals. We denote this correspondence $(t_1, t_2, \dots, t_N)$ and without loss of generality at time instance t

$$t_1 = t,\ t_2 = t + \tau_{1,2},\ \dots,\ t_N = t + \tau_{1,N}. \tag{1}$$

Let the origin of the coordinate be set at the centroid of the microphones, $1/N \sum_{i=1}^{N} \boldsymbol{m}_i$. Assuming that the distance to the sound source $r = ||\boldsymbol{s}||$ is much larger than the distance between microphone pairs, $r \gg \max_{i,j} ||\boldsymbol{m}_i - \boldsymbol{m}_j||$, (far field assumption), the dependence of the TDOAs to the source distance is negligible [17]. This also means that under this assumption, it is not possible to estimate the distance to the source, but only the spatial direction, which we parametrize by azimuth and elevation, namely (α, β).

The algorithm requires a *sound propagation model* which relates the direction (α, β) to the set of TDOAs $\{\tau_{i,j}\}$. Let us assume that this model is provided under the form of the following mappings:

$$\tau_{1,j} = \mathbf{q}_j(\alpha, \beta),\ \forall j, 2 \leq j \leq N. \tag{2}$$

A standard model is linear propagation of acoustic waves

$$\tau_{i,j} = \frac{f_s}{c}(||\boldsymbol{m}_i - \boldsymbol{s}||_2 - ||\boldsymbol{m}_j - \boldsymbol{s}||_2), \tag{3}$$

which involves difference of Euclidean distances between the source $\boldsymbol{s} = (r \sin\alpha \cos\beta, r \sin\alpha \cos\beta, r \cos\alpha)^T$ and the microphones $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ (c denotes a sound propagation speed, f_s is the sampling frequency). Note that the TDOA is expressed in number of signal samples and not in seconds. This model assumes that the acoustic waves travel along straight lines and that the 3D microphone locations are known in the robot-head coordinate system. These locations cannot be easily determined, in particular when the microphones are inside the robot head. Therefore we propose a method to estimate a model of the form of (2) which does not require explicit microphone locations (Sec. III).

The basic idea of the proposed sound source localization method is to discretize the space of expected sound directions

hal-00861465, version 1 - 12 Sep 2013

and to evaluate each one of these direction hypotheses by a consistency statistic which locally measures the joint similarity of signals in the correspondence determined by the tested direction. At each time instance t, the final estimate is the direction that accumulated the largest value of the consistency statistic

$$(\alpha^{(t)}, \beta^{(t)})^* = \underset{(\alpha,\beta)\in\mathcal{A}\times\mathcal{B}}{\operatorname{argmax}} \ \overline{\text{corr}}(\alpha, \beta; t), \tag{4}$$

where $\mathcal{A}$ and $\mathcal{B}$ are the discretized sets of azimuth and elevation values. The consistency statistic

$$\overline{\text{corr}}(\alpha, \beta; t) = \binom{N}{2}^{-1} \sum_{(i,j)\in\binom{N}{2}} \text{corr}_{i,j}(t_i, t_j) \tag{5}$$

is the average over all pairwise correlations $\text{corr}_{i,j}(t_i, t_j)$ and $\binom{N}{2}$ enumerates all 2-combinations from N elements. Notice that $t_1 = t$ and $t_i = t + \tau_{1,i} = t + \mathbf{q}_i(\alpha, \beta)$ for $2 \leq i \leq N$ which follows from (1) and (2).

Pairwise similarities are measured with the normalized cross-correlation (NCC) function

$$\text{corr}_{i,j}(t_i, t_j) = \frac{\text{cov}\big(\boldsymbol{W}_i(t_i), \boldsymbol{W}_j(t_j)\big)}{\sqrt{\text{var}\big(\boldsymbol{W}_i(t_i)\big)\text{var}\big(\boldsymbol{W}_j(t_j)\big)}} \tag{6}$$

between short temporal windows $\boldsymbol{W}_i$ and $\boldsymbol{W}_j$ of L samples each, centered at t_i and t_j of signals associated with microphones $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$. The size of the window is set to 100ms and $L = 0.1 f_s$ in all our experiments.

The consistency statistic has the range $[-1, 1]$ and it is invariant to affine transformations of the signal intensity. It handles well a gain difference between the microphones, and consequently it also handles an anisotropy due to their directional characteristics. Note that the aggregated statistic (5) uses all the sensors simultaneously and symmetrically.

B. Audio-Visual Association

Similarly to the approach in [16], the proposed audio-visual fusion model relies on 3D visual features. Since the task is to localize speakers and to estimate their speaking activity status, ideally one would like to find 3D lips/mouth locations and to combine these locations with 2D sound source locations. We propose to use a face detector, e.g., [18], simply because in the case of untethered interaction their detection/localization is quite reliable. We detect faces, match the face centers between the two images, and estimate their 3D position using stereoscopic triangulation. The 3D locations thus obtained are fair approximations of lips/mouth locations. Below we explain how these visual features are associated with the sound source location estimates.

Let $\boldsymbol{s}_k = (x_k, y_k, z_k)^T$ for $k \in \{1, \ldots, K\}$ be the 3D positions of candidate speakers in the robot-head coordinate system obtained by the above procedure. Clearly there is an azimuth & elevation direction associated with each 3D face-center location. The expected direction associated with a sound source located in $\boldsymbol{s}_k$ is found by transforming the Cartesian coordinates into spherical ones

$$\boldsymbol{\mu}_k = \big(\text{azimuth}(\boldsymbol{s}_k), \text{elevation}(\boldsymbol{s}_k)\big). \tag{7}$$

Due to limited resolution in the discretization and various possible distortion of the acoustic sound source direction estimate (4), discussed in Sec. I, we model it statistically with the following mixture:

$$p(\alpha, \beta) = \sum_{k=1}^{K} p(k)\mathcal{N}(\alpha, \beta; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) + p(K+1)\mathcal{U}. \tag{8}$$

The likelihood $p(\alpha, \beta|k)$ of a sound direction estimate (4) due to the activity of speaker k is modeled by Gaussian distribution centered at $\boldsymbol{\mu}_k$ (7). The likelihood $p(\alpha, \beta|K+1)$ of a direction to be an outlier is modeled with the uniform distribution $\mathcal{U}$ over the range of all possible directions, i.e., $p(\alpha, \beta \mid K+1) = \text{const}$. For all $k \in \{1, \ldots, K+1\}$ and assuming a uniform prior distribution $p(k) = \frac{1}{K+1}$, the posterior probability is

$$p(k \mid \alpha, \beta) = \frac{p(\alpha, \beta \mid k)}{\sum_{k=1}^{K+1} p(\alpha, \beta \mid k)}. \tag{9}$$

Hence, for each time instance t the algorithm provides a distribution of posterior probabilities of activity over each speaker and the outlier class.

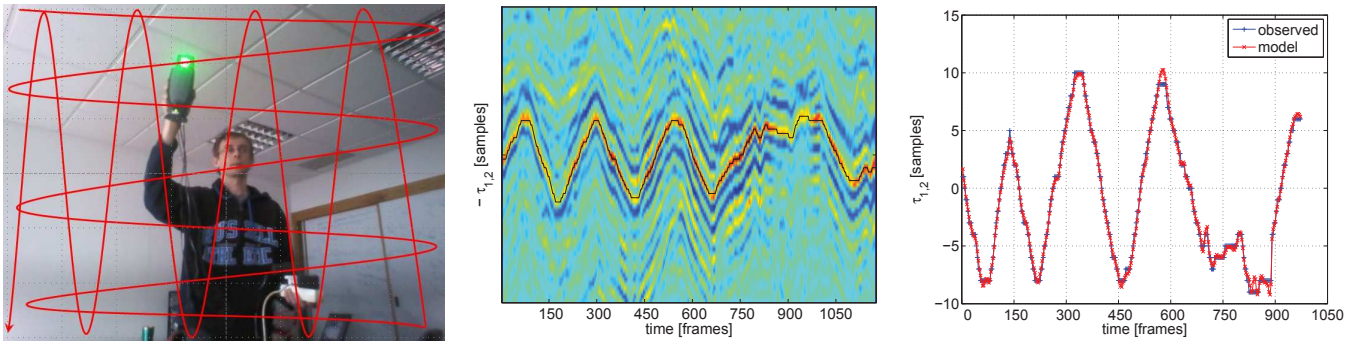
Discussion. Notice that the covariance matrix $\boldsymbol{\Sigma}$ is chosen to be the same over all Gaussians in the mixture (8). We assume that the variance of the distribution of estimates (α, β) due to speaker's activity does neither depend on the expected direction $\boldsymbol{\mu_k}$ of particular speakers nor on the emitted sound. We also assume that the estimates of α and β are not correlated and that the variance of the distribution is caused mainly by the discretization, since the system is precisely calibrated. Therefore, the covariance matrix $\boldsymbol{\Sigma}$ is set as a diagonal matrix with standard deviations as the respective discretization steps of azimuth and elevation.

All these assumptions were verified on the calibration sequence, where we have a correspondence between the ground-truth direction of the sound source and the estimate by the SSL algorithm (4), see Sec. III. We observed that either the correct maximum is selected in (4) (the one which corresponds to the true direction up to the precision given by the discretization), or rarely there is a mismatch (occurring most probably due to reverberations) which is then captured by the uniform outlier class.

This straightforward probabilistic model can be seen as a Bayesian classifier. The most probable class (one of the speakers $1, \ldots, K$) or the outlier class $(K+1)$ is selected by (9), which is equivalent to finding the closest $\boldsymbol{\mu}_k$ to an estimated (α, β) in the sense of the Mahalanobis distance. If this distance is above a threshold (given by the covariance matrix $\boldsymbol{\Sigma}$), the outlier class is selected. Despite this simplicity, this audio-visual association turns out to be very efficient, which will be illustrated with experiments.

An alternative to the proposed model would be to relax some of the above assumptions. Namely, to hypothesize that the distribution of the estimated sound directions (around their expected means) do depend on the particular speaker. If this is the case, then the covariance matrices would

hal-00861465, version 1 - 12 Sep 2013



(a) (b) (c)

Fig. 2: Audiovisual calibration. Approximate trajectory of the loudspeaker with a light bulb superimposed in red (a), Color-coded correlogram for TDOAs with superimposed trajectory found by dynamic programming (b), Observed TDOAs from audio signals and visual directions projected into TDOA space using the estimated model (c).

be speaker/direction specific and should be estimated on-line during a short period of time during which several directions are collected. Then the GMM's parameters would be estimated via an EM procedure, e.g., by extending the model [16] from 1D to 2D sound localization. Such an approach would have required enough *independent* samples. Since a 100ms window is required in order to compute an estimate of reasonable quality, a relatively long temporal window would have been necessary in order to gather a large number of samples, which is necessary for GMM fitting. This is prohibitive within the framework of an on-line interactive process, as required by the application at hand.

III. Sound Propagation Model and AV Alignment

In this section we make explicit the sound propagation function (2), a method for estimating its parameters and the associated procedure used in practice with the robot head.

The standard model (3) requires the positions of the microphones in a head-centered coordinate system [19], [16]. We performed a number of experiments using the manufacturer's technical documentation, and it turned out that the model (3) does not hold well. Namely, we noticed a systematic bias in the TDOA estimates. For instance, we observed that the maximum value of TDOA estimated from acoustic signals was significantly larger than the maximum TDOA which would correspond to the actual baseline between microphones. The wave travels around the head to reach the microphones probably. The shape and the head's material influence the sound propagation. Therefore we propose to model (2) using regression functions, one for each microphone pair, which relates the sound source direction in the robot-head coordinate system with the TDOAs between audio signals. With the far-field assumption in mind, we suggest to use a model of the form

$$\tau_{1,j} = q_j^2\alpha + q_j^1\beta + q_j^0 \qquad (10)$$

The model parameters $\{q_j^2, q_j^1, q_j^0\}_{j=2}^N$ are estimated from a set of correspondences $\{(\alpha^{(t)}, \beta^{(t)}) \leftrightarrow \tau_{1,j}^{(t)}\}_{t=1}^T$ using a standard linear regression. Moreover, since the reference directions $(\alpha^{(t)}, \beta^{(t)})$ are provided to the calibration by the vision system, the proposed model implements the audio-visual (AV) alignment which is used in the statistical association (8).

To acquire calibration data we use a loudspeaker with a light source attached to it. This *audi-visual target* is freely moved in front of the robot. This setup was also used in [20], where an EM procedure is used to estimate the microphone positions $\boldsymbol{m}_i$ of model (3).

The visual field of view of the robot is slowly swept in a 'zig-zag' trajectory by the loudspeaker emitting random white noise while the attached light source is easily detectable in the image pair, see Fig. 2a. Since, the cameras are calibrated in a common coordinate frame using standard camera calibration, after detecting the light-bulb in left camera and matching with the right image, we reconstruct a 3D position of the sound source $\boldsymbol{s}^{(t)}$ for each time instance t. Angles $(\alpha^{(t)}, \beta^{(t)})$ are then easily derived by conversion to spherical coordinates.

Signals from the microphones perceiving the white noise correlate well in general. However, there might still be some errors if we find $\tau_{1,j}$ at each time instance t independently by shifting signals and maximizing their correlations. Therefore, we use the fact that we move the speaker slowly and TDOAs do not change abruptly. We are looking for their smooth sequence over time $t = 1, \ldots, T$ which maximizes

$$\mathrm{corr}_{1,j}(1, 1+\tau_{1,j}^{(1)}) + \sum_{t=2}^{T}\Big(\mathrm{corr}_{1,j}(t, t+\tau_{1,j}^{(t)}) + \lambda(\tau_{1,j}^{(t)} - \tau_{1,j}^{(t-1)})^2\Big). \qquad (11)$$

Regularization parameter is $\lambda = 0.1$. The optimum sequence $(\tau_{1,j}^{(1)}, \ldots, \tau_{1,j}^{(T)})$ is found by dynamic programming. See the smooth path over the table of $\mathrm{corr}_{1,2}$ in Fig. 2b. This way, we avoid outliers in the correspondence set $(\alpha^{(t)}, \beta^{(t)}) \leftrightarrow \tau_{1,j}^{(t)}$ and we use statistically efficient least squares regression to fit the model (10). The procedure is repeated for all microphone pairs, $j = 2, \ldots, N$.

The accuracy of the model is demonstrated in Fig. 2c. There are two curves: TDOA trajectory found from audio

hal-00861465, version 1 - 12 Sep 2013

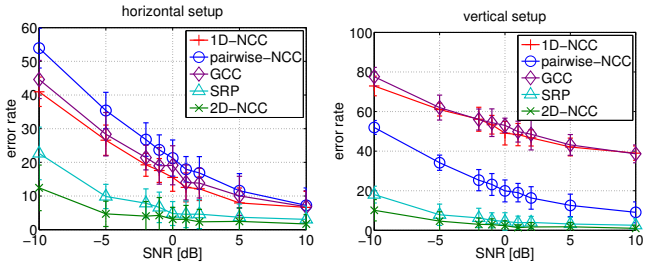


Fig. 3: Error rate against noise.

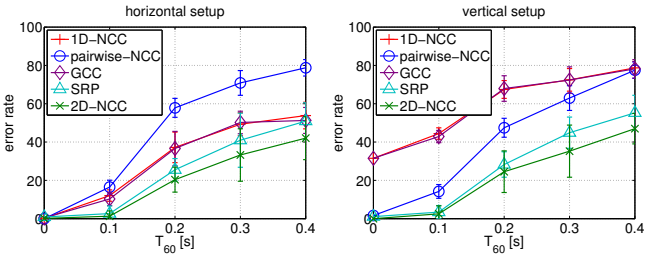


Fig. 4: Error rate against reverberations.

signal by optimization (11) (observed), and the curve where 3D trajectory of the target is projected onto the TDOA space using the estimated model (model). This shows that, for a limited range of directions (spanning the robot's visual field of view), the model (10) approximates very well the actual sound propagation.

IV. Experiments

The accuracy of the presented method was evaluated quantitatively. We measure the error rate, which is a percentage of discrepancies of the MAP estimate (9) against the ground-truth annotation.The annotation labels the active speakers who take speaking turns sequentially. Three speakers were used in all the tests, which means the maximum error rate (by pure random chance) is 75% since there is an additional class to capture outliers. The evaluation is for every frame lasting 1/30 sec, which is derived from the video frame-rate.

Besides the proposed method (*2D-NCC*), we compare with several other baseline methods. The first (*1D-NCC*) uses only two microphones and hence estimates the ITD $\tau_{1,2}$. Then, Gaussian means are projections of source direction (found by visual detection) into this space, standard deviation is set to $\sigma = 1$ sample.

The second baseline method (*pairwise NCC*) uses all four microphones. The association model lives in the space of TDOAs of $N-1$ dimensions. Gaussian means are projections of source direction into the space $\tau_{1,2} \times \tau_{1,3} \times \tau_{1,4}$. The covariance matrix Σ is set as unit matrix. All three TDOAs are found *independently* by maximizing the pairwise correlation by shifting the signals.

The last two baseline methods are classical techniques based on Fourrier domain correlation statistic. The correlation statistic is a dot product of the short-time Fourier transforms of signals which are steered by a multiplication of a complex exponential with a phase shift set according to the expected TDOA. In one-dimensional case (between two signals), it is the generalized cross-correlation (*GCC*) – it corresponds to *1D-NCC*. In the two-dimensional case (using multiple microphones), it is the steered response power (*SRP*) – it corresponds to *2D-NCC*. For further details on these methods, we refer the reader to [21], [22], [23]. In all experiments, we use the same size of the Hamming window in short-time Fourier transform as the size of the temporal window in NCC correlation (6), i.e. 100 ms.

A. Simulated Experiment

To quantitatively demonstrate the influence of noise and reverberations we performed a controlled synthetic experiment. The setup is similar as in [17]. A room of $3 \times 4 \times 2$ meters contains 3 collinear sound sources which are placed either horizontally or vertically about 2 meters away from an array of 4 microphones. The array is arranged as a tetrahedron with an edge of 10 cm placed roughly in the middle of the room. The baseline between the left and right microphones was parallel to the ground floor. The simulation which allowed to control the reverberation time[1] T_{60} was performed using the toolbox of [24]. The sound sources were repeatedly emitting, one by one, 0.5s length speech fragments randomly chosen from the TIMIT dataset with 0.25s silence gaps. All the results below are averaged out over 10 random trials of speech fragments generations. The plots in Fig. 3 and Fig. 4 have error bars of the standard deviation.

First, we tested the methods against increasing level of additive Gaussian noise over all signals independently. In Fig. 3, we can see that *1D-NCC* has problems in the vertical setup. Since it uses only two horizontal microphones and the single $\tau_{1,2}$ have almost the same value for all three sound sources. In the vertical setup, the simple method *pairwise NCC* which uses all four microphones is of course able to distinguish vertically aligned sources for low noise, but it deteriorates quickly with increasing level of noise. The reason is that a single mistake in any of three TDOAs causes an error. We can see that for sources aligned horizontally, the error rate of *pairwise NCC* increases even faster than with *1D-NCC*. However, we can see that the proposed method *2D-NCC* is consistently significantly better than both baseline methods. The reason is that all microphones are used simultaneously and the aggregated correlation statistic (5) efficiently averages out the noise.

Concerning a comparison with *GCC* and *SRP*, i.e. Fourier domain methods, we can see the *GCC* has similar performance as *1D-NCC* and *SRP* is consistently slightly worse than *2D-NCC*. This is probably due to certain effects of the Hamming window used in the short-time Fourier transform, although the temporal support was the same as in *2D-NCC*. A similar behavior is observed for tests with reverberations, see Fig. 4. The proposed method *2D-NCC* is still consistently

[1]Reverberation time T_{60} is the time in seconds required that the sound level decays by 60dB below the original level.

hal-00861465, version 1 - 12 Sep 2013

Fig. 5: People emitting sounds sequentially (top) and the corresponding aggregated statistic (5) for azimuth and elevation (bottom) roughly aligned with the image and when the leftmost person speaks. See the supplementary video for the entire sequence with the original sound track. Please notice the high level of noise in the signal due to fan inside the robot head.

better, although we can see that after $T_{60} > 0.1$s, the error rate increases quickly. The problem is that in a highly reverberant room, the reflected sound causes multiple phantom source locations which corrupt the correlation statistic. It may even happen with large T_{60} that such a phantom source overwhelms the original sound in cases of pauses in speech. Since the SSL is run in at a regular rate, large error bars are then probably due to this effect when SSL is computed close to speech pauses.

B. Real Data Experiment

For real experiments, we used a subset of data used in [16], which consists of 10 sequences (of about 30 sec) captured with the prototype head of NAO. There are always three persons in front of the robot in various configurations as in Fig. 5. They are sequentially uttering by counting from 1 to 15. Full ground-truth annotation was available. The room does not have any special acoustic properties and it is quite echoic. The audio track of the video in the supplementary material corresponds to NAO's sound track recorded by one of the robot's microphones, thus allowing one to listen to what the robot actually hears. The noise is mainly due to a CPU fan located inside the head, which makes the problem of robot listening particularly challenging.

The error rate for each sequence is shown in Table I. The proposed method *2D-NCC* is the best performing one for all sequences. The fluctuation of the performance is probably due to large variations in the speech loudness. Notice that the difference between *SRP* and *2D-NCC* is more significant for the real than for the synthetic data, which could be caused by higher sensitivity of the method to effects due to the HRTF

	1D-NCC	*pairwise NCC*	*GCC*	*SRP*	*2D-NCC*
seq-1	19.7	32.8	21.3	23.0	**16.4**
seq-2	13.4	22.8	17.4	8.7	**0.7**
seq-3	11.5	18.9	12.8	7.4	**4.1**
seq-4	14.6	21.9	18.5	12.6	**11.9**
seq-5	13.1	20.9	15.7	7.2	**2.0**
seq-6	22.8	30.4	22.8	7.6	**7.0**
seq-7	22.7	22.1	16.9	5.8	**1.7**
seq-8	17.0	25.5	13.1	9.8	**6.5**
seq-9	18.9	19.6	16.1	21.7	**16.1**
seq-10	10.9	18.2	11.5	12.7	**7.9**
average	16.4	23.3	16.6	11.7	**7.4**

TABLE I: Error rate for real experiments.

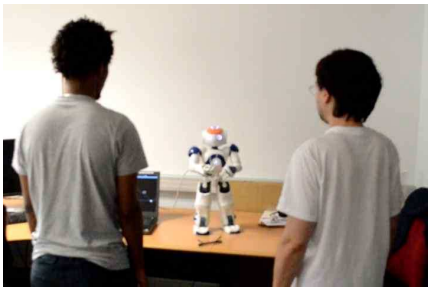


Fig. 6: Humanoid-human interaction in an unconstrained environment (background noise, reverberations, etc.). Please see the the supplementary video.

of the head being used. Certain frequency bands could be distorted in an anisotropic way, which might be more harmful for the Fourier domain method. Moreover, unlike *2D-NCC*, *SRP* is not invariant to affine transformations of the signal level. This effect may be present due to the poor quality of the microphones.

C. Interactive Behavior

The proposed method was implemented on NAO to demonstrate a possible human-robot interaction. Two demonstrations were performed, both of them run in real-time. The first demonstration was a simple reactive behaviour, where the robot turns its head towards an active speaker, see Fig. 1. The second demonstration extends the first one, such that the active speaker is engaged in a simple multimodal dialogue. The scenario and implementation details follow.

1) Proposed Scenario: There are many situations where an active-speaker detector and localizer could be useful. In this section we describe a simple scenario and its associated protocol of communication, as shown in the Fig. 7. Several people face NAO, Fig. 6, who stands up on a table, to be at approximately the same height as the persons and thus to optimally gather audio and visual information with its head-embedded cameras and microphones. Using the active-speaker detection and localization (*AV fusion* module), NAO is able to turn its head towards one of the active speakers. The speaker can then enter a dialogue by clapping his/her hands. Several modules are then activated.

First, a visual *Identity Recognition* module determines whether NAO has already seen that person and spoken to him or her. If it is the case, NAO turns its attention to another

hal-00861465, version 1 - 12 Sep 2013

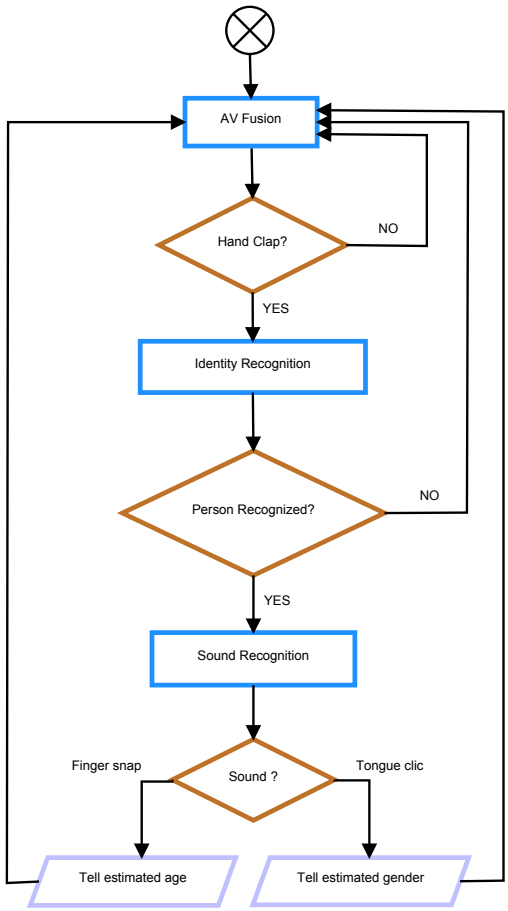


Fig. 7: Flowchart of the interactive protocol with NAO. The modules involved in the scenario are shown with blue squares. A brown rhombus illustrates a decision that NAO has to make based on the information flow received from the different modules. Purple parallelograms correspond to situations when NAO has to take an action. The *AV Fusion* module corresponds to active-speaker detection.

person. If the person was never seen before, NAO greets and asks him/her to emit a sound. Based on an isolated sound recognizer [25], a number of human-emitted acoustic signals can be recognized and NAO exhibits a different behaviour for each one of the recognized sounds. In this scenario, NAO attempts to guess either the gender or the age of that person based on facial cues.

2) Implementation on the Humanoid Robot NAO: The scenario described above was implemented using the Robotic Service Bus (RSB) middleware [26], see Fig. 8. The RSB middleware interfaces robot's sensors and actuators and hosts additional software modules which run in parallel. The RSB is in charge of module interconnection, communication and synchronization. The software modules may run in different computers and the processes communicates over the network in a way transparent for a user. All the events are automatically equipped with time stamps, which provide for introspection and synchronization abilities. Several tools

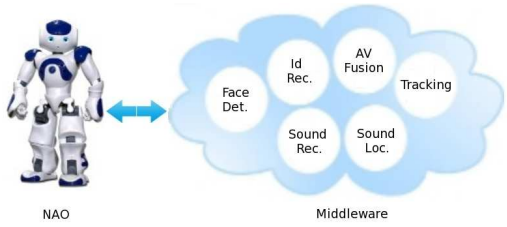


Fig. 8: The method described in this paper was implemented using the RSB (robotic service bus) middleware that interfaces NAO's sensors and actuators with the software modules.

exist which can record event flow, and replay it later, so that application development can largely be done without running the robot.

In practice the modules involved in this scenario are the followings: *Face Detector* [27], *Identity Recognition* (identity, age, and gender) [28], [29], *Face Tracking* [30], *Sound Recognition* [25], *Sound Localization* and *AV Fusion* (described in this paper).

Some of the modules are necessary to read results of other modules. For example, *Face Detector* is used in *AV Fusion* and in *Tracking*, while at the same time *Face Tracking* is used by *Identity Recognition*. All these modules are available for download[2].

For example, the *Face Detector* module collects image sequences from NAO's cameras, localizes the faces in these images, if any, and communicates to other modules, through RSB, the corresponding bounding boxes of the faces. In this manner, *Face Tracking* can be initialized and start the tracking of a face. Something similar holds for the *AV Fusion* module, which needs 2D auditory directions and 3D visual features, e.g., 3D face positions.

The active speaker detection and localization is very important in the whole application, and the fact that NAO is able to find who is speaking and move the head towards the speaker provides the sensation of a natural conversation. It would be certainly even more natural if the dialogue is driven by a speech recognition module instead of the isolated sound recognizer. However, the speech recognition in a reverberant environment with distant microphones embedded in the noisy robotic head is not reliable. This is a widely studied research topic itself which typically employs larger microphone arrays [31], [32]. Therefore we replaced the Nao's built-in speech recognition module with the sound recognition module [25], which turned out to be much more reliable in this condition.

V. Conclusions

In this paper we presented a method for the detection and localization of an active speaker that faces a robot. We devised a novel sound-source localization method that

[2]https://code.humavips.eu/

hal-00861465, version 1 - 12 Sep 2013

sums up correlation statistics from an arbitrary number of microphone pairs and which builds a 2D auditory map of possible sound directions. We also devised a simple Gaussian mixture model that combines this map with 3D visual data such that the two modalities can be fused in a principled way. A sound propagation model which captures robot's visual field of view and its calibration procedure was designed. The method was experimentally validated and tested with both simulated and real data using cameras and microphones embedded into the head of the humanoid robot NAO.

The method is robust against noise, but is still quite sensitive to reverberations. In our future work, we will investigate how to estimate sound-source direction prominently based on strong onsets detected in the perceived acoustic signals as in [33]. Furthermore, there are many possibilities in feeding the statistical model from the correlation map (4). For instance, multiple maxima or weights related to correlation could be used.

The system assumes that there is no overlapping acoustic activity of multiple subjects. When this is violated the algorithm tends to select a dominant source at a time. We are planning to investigate this challenge in future by a statistical modelling in both temporal and frequency domains.

Since the algorithm outputs posterior probabilities, the reasoning on the active speaker could be further improved by including temporal information. So far the data are processed independently, frame by frame. Therefore, one could construct a HMM over the posterior probability distribution of the speakers activity over several frames.

The human-robot interaction scenario that we described is illustrative because it encompasses a large number of situations where human-robot dialogue is based on both auditory and visual information.

Acknowledgements. The research was supported by EC project FP7-ICT-247525 HUMAVIPS and by The Czech Science Foundation Project GACR P103/12/G084.

REFERENCES

[1] H. Nock, G. Iyengar, and C. Neti, "Speaker localization using audio-visual synchrony: An empirical study," in *Proc. CIVR*, 2003.

[2] Z. Li, T. Herfet, M. Grochulla, and T. Thormahlen, "Audio-visual multiple active speaker localization in reverberant environments," in *Proc. Int. Conference on Digital Audio Effects*, 2012.

[3] E. Kidron, Y. Schechner, and M. Elad, "Pixels that sound," in *CVPR*, 2005.

[4] P. Aarabi and S. Zaky, "Robust sound localization using multi-source audiovisual information fusion," *Information fusion*, vol. 2, 2001.

[5] C. Zhang, P. Yin, Y. Rui, R. Cutler, P. Viola, X. Sun, N. Pinto, and Z. Zhang, "Boosting-based multimodal speaker detection for distributed meeting videos," in *IEEE Trans. on Multimedia*, vol. 10, no. 8, 2008.

[6] J. Blauert, *Spatial Hearing The Psychophysics of Human Sound Localization*. MIT Press, Cambridge, UK, 1997.

[7] L. Calmes, G. Lakemeyer, and H. Wagner, "Azimuthal sound localization using coincidence of timing across frequency on a robotic platform," *J. Acoust. Soc. Am.*, vol. 21, no. 4, 2007.

[8] V. M. Trifa, A. Koene, J. Moren, and G. Cheng, "Real-time acoustic source localization in noisy environments for human-robot multimodal interaction," in *Robot and Human interactive Communication*, 2007.

[9] J. Hornstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robot - building audio-motor maps based on the HRTF," in *Int. Conference on Intelligent Robots and Systems*, 2006.

[10] J.-M. Valin, F. Michaud, J. Rouat, and D. Letourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Int. Conference on Intelligent Robots and Systems*, 2003.

[11] J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, 2003.

[12] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *International Journal of Comuter Vision*, vol. 38, no. 3, 2000.

[13] C. Strecha, W. von Hansen, L. V. Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *CVPR*, 2008.

[14] M. S. Brandstein and H. F. Silverman, "A practical methodology for speech source localization with microphone arrays," *Computer Speech and Language*, vol. 11, no. 2, 1997.

[15] Y. Furukawa and J. Ponce, "Accurate camera calibration from multi-view stereo and bundle adjustment," *International Journal of Computer Vision*, vol. 84, no. 3, 2009.

[16] J. Sanchez-Riera, X. Alameda-Pineda, J. Wienke, A. Deleforge, S. Arias, J. Cech, S. Wrede, and R. P. Horaud, "Online multimodal speaker detection for humanoid robots," in *IEEE International Conference on Humanoid Robotics*, 2012.

[17] X. Alameda-Pineda and R. P. Horaud, "Geometrically-constrained robust time delay estimation using non-coplanar microphone arrays," in *European Signal Processing Conference*, 2012.

[18] P. Viola and M. Jones, "Robust real-time face detection," *IJCV*, vol. 57, 2004.

[19] X. Alameda-Pineda, V. Khalidov, R. P. Horaud, and F. Forbes, "Finding audio-visual events in informal social gatherings," in *Proceedings of the 13th International Conference on Multimodal Interfaces*. Alicante, Spain: ACM, November 2011, pp. 247–254.

[20] V. Khalidov, F. Forbes, and R. P. Horaud, "Alignment of binocular-binaural data using a moving audio-visual target," in *IEEE Workshop on Multimedia Signal Processing (MMSP'13)*. Pula (Sardinia), Italy: IEEE Signal Processing Society, September-October 2013.

[21] A. Badali, J.-M. Valin, F. Michaud, and P. Aarabi, "Evaluating real-time audio localization algorithms for artificial audition in robotics," in *Proc. IROS*, 2009.

[22] C. H. Knapp and G. C. Carter, "Generalized correlation method for estimation of time delay," *IEEE Trans. on Acoustic, Speech and Signal Processing*, vol. 24, no. 4, 1976.

[23] M. Omologo and P. Svaizer, "Use of the crosspower-spectrum phase in acoustic event location," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 3, 1997.

[24] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," *Journal of the Acoustical Society of America*, vol. 124, no. 1, 2008.

[25] M. Janvier, X. Alameda-Pineda, L. Girin, and R. Horaud, "Sound-event recognition with a companion humanoid," in *IEEE International Conference on Humanoid Robotics*, 2012.

[26] J. Wienke and S. Wrede, "A middleware for collaborative research in experimental robotics," in *2011 IEEE/SICE Internatinal Symposium on System Integration*, IEEE. Kyoto, Japan: IEEE, 2011.

[27] J. Šochman and J. Matas, "Waldboost – learning for time constrained sequential detection," in *Proceedings of the IEEE Computer Vision and Pattern Recognition*, 2005.

[28] M. Uřičář, V. Franc, and V. Hlaváč, "Detector of facial landmarks learned by the structured output SVM," in *VISAPP*, 2012.

[29] V. Franc, S. Sonnenburg, and T. Werner, *Cutting-Plane Methods in Machine Learning*. The MIT Press, 2012, ch. 7, pp. 185–218.

[30] S. Duffner and J.-M. Odobez, "A track creation and deletion framework for long-term online multi-face tracking," in *IEEE Transaction on Image Processing*, 2013.

[31] R. Gomez, K. Nakamura, T. Kawahara, and K. Nakadai, "Multi-party human-robot interaction with distant-talking speech recognition," in *Proc. Human-Robot Interaction (HRI)*, 2012.

[32] C. T. Ishi, S. Matsuda, T. Kanda, T. Jitsuhiro, H. Ishiguro, S. Nakamura, and N. Hagita, "A robust speech recognition system for communication robots in noisy environments," *IEEE Trans. on Robotics*, vol. 24, no. 3, 2008.

[33] J. Huang, N. Ohnishi, X. Guo, and N. Sugie, "Echo avoidance in a computational model of the precedence effect," *Speech communication*, vol. 27, 1999.