Non-Field-Of-View Sound Source Localization based on Reflection and Refraction

Kuya Takami, Hangxin Liu, and Tomonari Furukawa

***Abstract*—**

I. INTRODUCTION

II. NFOV

1) Hybrid Visual/Auditory Recursive Bayesian Estimation: The proposed approach is mathematically described as follows. The SLAM technique to be used in the project is a scan-based SLAM which incorporates the grid-based scan-to-map matching technique proposed by the authors. This technique could be claimed as the most effective approach for this class of SLAM problems with further engagement, but the work will not be one of the major contributions in the proposal since this project is concerned only with the use of SLAM at a concept level. Let the state of the robot s and the map updated at time step k be $\bar{\mathbf{x}}_k^s \in \mathcal{X}^s$ and $\bar{\mathbf{m}}_k \in \mathcal{M}_k$ respectively and start the mathematical formulation of the proposed approach. Consider a target t, the state of which is given by $\mathbf{x}_k^t \in \mathcal{X}^t$, and a sequence of observations of the target t by the robot s from time step 1 to time step k given by ${}^s\tilde{\mathbf{z}}_{1:k}^t \equiv \{{}^s\tilde{\mathbf{z}}_\kappa^t | \forall \kappa \in \{1, ..., k\}\}$. The RBE represents belief on the target in the form of a probability density function and iteratively updates the belief in time and observation. Let the belief given a sequence of observations and the robot state and the map estimated by SLAM at time step $k-1$ be $p\left(\mathbf{x}_{k-1}^t | {}^s\tilde{\mathbf{z}}_{1:k-1}^t, \bar{\mathbf{x}}_{k-1}^s, \bar{\mathbf{m}}_{k-1}\right)$. Chapman-Kolmogrov equation updates the prior belief in time, or predicts the belief at time step k, by the probabilistic motion model $p\left(\mathbf{x}_k^t | \mathbf{x}_{k-1}^t, \bar{\mathbf{x}}_{k-1}^s, \bar{\mathbf{m}}_{k-1}\right)$:

$$p\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_{1:k-1}^t, \bar{\mathbf{x}}_{k-1}^s, \bar{\mathbf{m}}_{k-1}\right) = \int_{\mathcal{X}^t} p\left(\mathbf{x}_k^t | \mathbf{x}_{k-1}^t, \bar{\mathbf{x}}_{k-1}^s, \bar{\mathbf{m}}_{k-1}\right) p\left(\mathbf{x}_{k-1}^t | {}^s\tilde{\mathbf{z}}_{1:k-1}^t, \bar{\mathbf{x}}_{k-1}^s, \bar{\mathbf{m}}_{k-1}\right) d\mathbf{x}_{k-1}^t \tag{1}$$

Note that the motion model is $p\left(\mathbf{x}_k^t | \mathbf{x}_{k-1}^t, \bar{\mathbf{m}}_{k-1}\right)$ if the target is not reactive to the robot. The observation update, or the correction process, is performed using the Bayes theorem. The target belief is corrected using the new observation ${}^s\tilde{\mathbf{z}}_k^t$ as

$$p\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_{1:k}^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) = \frac{q\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_{1:k}^t, \bar{\mathbf{x}}_{k-1:k}^s, \bar{\mathbf{m}}_{k-1:k}\right)}{\int_{\mathcal{X}^t} q\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_{1:k}^t, \bar{\mathbf{x}}_{k-1:k}^s, \bar{\mathbf{m}}_{k-1:k}\right) d\mathbf{x}_k^t}, \tag{2}$$

where $q(\cdot) = l\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) p\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_{1:k-1}^t, \bar{\mathbf{x}}_{k-1}^s, \bar{\mathbf{m}}_{k-1}\right)$, and $l\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right)$ represents the observation likelihood of $\mathbf{x}_k^t$ given ${}^s\tilde{\mathbf{z}}_k^t$, $\bar{\mathbf{x}}_k^s$ and $\bar{\mathbf{m}}_k$.

*This work was not supported by any organization

[1]Kuya Takami, Hangxin Liu, Tomonari Furukawa are with Mechanical Engineering, Virginia Polytechnic Institute and State University, , USA {kuya, hangxin, tomonari}@vt.edu

One of the core technologies proposed in this project is the dual use of visual and auditory sensors. Most generically, it is given by

$$l\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) = \prod_i l_i^c\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) \prod_j l_j^a\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) \tag{3}$$

where $l_i^c(\cdot)$ and $l_j^a(\cdot)$ are the likelihoods of ith camera and jth acoustic sensor. In order to maximize information, the camera observation is used not only to detect a target if it is in the FOV but also to construct the no-detection likelihood if the target is not detected in the field of view:

$$l_i^c\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) = \begin{cases} p\left({}^s\tilde{\mathbf{z}}_k^t | \mathbf{x}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) & \exists {}^s\tilde{\mathbf{z}}_k^t \in {}^{s_i}\mathcal{X}_{FOV}^t \\ 1 - P_d\left(\mathbf{x}_k^t | \bar{\mathbf{x}}_k^s\right) & \nexists {}^s\tilde{\mathbf{z}}_k^t \in {}^{s_i}\mathcal{X}_{FOV}^t, \end{cases} \tag{4}$$

where ${}^{s_i}\mathcal{X}_{FOV}^t$ is the FOV of the ith camera. The effectiveness of Equation (4) is thoroughly investigated by the PI in the context of autonomous search and tracking.

While the derivation of $l_j^a(\cdot)$ is most challenging and thus will be dealt with separately in the next section, the advantage of Equation (3) in RBE is illustratively shown in Figure **??**. The possible locations of the target are narrowed down even though the no-detection likelihood is used in visual sensing since the likelihood clears out the joint likelihood in the FOV and dropped some peak(s) as shown in Figure **??**. Because sharpest and most Gaussian is the visual observation likelihood with detection, the prior belief is most determined by the last visual observation and remains a sharp Gaussian distribution as shown in Figure **??**. The posterior belief with the joint observation likelihood inherits this characteristics since the joint likelihood most likely captures the target location with a peak and magnifies the confidence of the prior belief with the joint likelihood.

2) Construction of Auditory NFOV Target Observation Likelihood: **Overview**: Unlike the conventional techniques, the proposed approach

- Is not based on the LOS assumption and actively utilizes the physics of sound wave propagation associated with NFOV targets;
- Does not need information such as the time of sound emission and power to be informed by the target;
- Does not need to collect acoustic cues in advance.

Figure **??** shows the overview of the approach proposed for constructing a NFOV target observation likelihood using an auditory sensor. The core of the proposed approach is to extract the first-arrival diffraction and reflection signals by taking the physics of sound wave propagation into account.

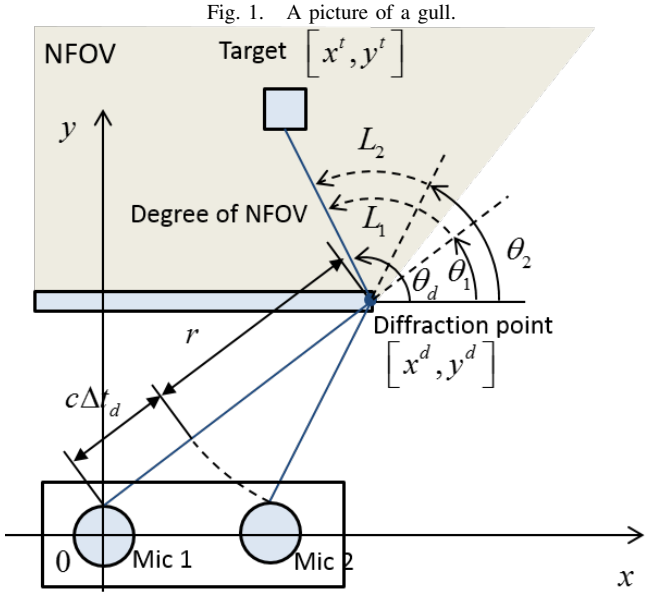
Unlike radio signals, sound signals reflect significantly without penetrating into different media while they also diffract at low frequencies [**?**]. The proposed approach begins with obtaining a time-domain signal of a relatively impulsive sound at each microphone. In each curve, notable peaks are then extracted as candidate first-arrival diffraction and reflection signals. When each candidate signal is described in the frequency domain, the first-arrival diffraction and reflection signals can be identified since they are the first signals that are correlated in low frequency range. The diffraction signal is then used to identify the so-called diffraction point by deriving the Time-Difference-Of-Arrival (TDOA) for each pair of microphones and further estimate the direction of target sound beyond the diffraction point from the loss of sound energy through diffraction, or the diffraction loss. An observation likelihood is eventually constructed by additionally estimating the distance from the sound magnitude and characteristics. The reflection signal estimates the target direction directly from the TDOA by mirroring and creating a virtual target. It also creates an observation likelihood with distance estimate by considering sound magnitude and characteristics and environmental properties. A joint observation likelihood is finally created by the fusion of the diffraction and reflection observation likelihoods.

The proposed approach infers the location of the sound target using both the first-arrival diffracted and reflected sound signals. The next subsection describes the extraction of the first-arrival diffraction and reflection signals, followed by the target estimation using the diffracted and reflected signals in the subsequent two subsections. The final goal of this project is to develop a probabilistic RBE based framework, but the preliminary study has succeeded in the proof-of-concept in deterministic formulations. The two subsections will present the deterministic NFOV target estimation using diffraction and reflection sound waves. The final subsection derives the joint observation likelihood as a result of data fusion.

III. Extraction of First-arrival Diffraction and Reflection Signals

Figure **??** shows the extraction process of the diffraction and reflection signals proposed in this project illustratively in one of the simplest scenarios where a robot carrying two microphones receives sound emitted by a target in the NFOV in a two-dimensional indoor environment with three walls (Figure **??**). As shown in the figure, sound waves emitted from the target reach the robot first through diffraction and second through reflection and, if the sound is relatively impulsive, the first-arrival diffraction and reflection signals can be extracted clearly. Extraction becomes challenging for complex environments, but various existing techniques proposed to extract signals or select thresholds for extraction reportedly achieve successful extraction and identify candidate diffraction and reflection signals [**?**], [**?**], [**?**], [**?**]. Figure **??** shows not only the sound pressure of sound in the time domain, $P_i(t)$, but also the magnitude of the resulting first-arrival diffraction and reflection signals in the frequency domain, $M_i^d(\omega)$ and $M_i^r(\omega)$, where $i \in \{1, 2\}$ is the index of microphone. Note that the magnitude is scaled to examine correlation. Signals are considered from the same sound source if they share the same characteristics in low frequency because low-frequency signals reflect and diffract. The proposed approach thus select the first set of signals that have the same low-frequency characteristics but are dissimilar in high frequency as the first-arrival diffraction and reflection signals of all the candidate signals. Diffraction signals have little high-frequency components whilst reflection signals see components in all frequencies. Needless to say, each of Microphones 1 and 2 constructs a different data set.

Fig. 1. A picture of a gull.



Estimation of Sound Direction from Diffraction Signals: Figure **??** shows the notations used for target estimation from diffraction signals in the scenario introduced in the last subsection. Since the diffraction sound Microphones 1 and 2 receive is originated from the LOS location at which the sound diffracts, the proposed approach starts target estimation from diffraction signals with the selection of diffraction point from all candidates, which are corners of all structures. The measured quantity used for the selection is the TDOA, $\Delta t_d = t_{d2} - t_{d1}$, where t_{d1} and t_{d2} are the Times-of-Arrival (TOAs) at Microphones 1 and 2 respectively. The diffraction point can be easily found from candidates as it satisfies the following equation:

$$\left(x^d\right)^2 + \left(x^d\right)^2 = \left(c\Delta t_d + r\right)^2. \tag{5}$$

where $\left[x^d, y^d\right]$ is the location of a candidate diffraction point, c is the speed of sound and r is a shorter distance between a microphone and the candidate diffraction point. With the diffraction point identified, the proposed approach further identifies the direction of the sound target from the diffraction point by analyzing the magnitudes of diffraction and reflection sounds $M_i^d(\omega)$ and $M_i^r(\omega)$. The loss of high-frequency signal components is assumed to be less with a microphone closer to LOS (Microphone 2 in this case) as

there is no loss with a microphone on the LOS to the sound target. This assumption, in fact, has been found to be valid by the work of Medwin a quarter-century ago [**?**] shown in Figure **??**. The magnitude of diffraction sound drops when the "degree of NLOS" represented by the orientation angle is increased. This makes the proposed approach define the diffraction loss as

$$L_i = \int \left[M_i^r(\omega) - M_i^d(\omega) \right] \mathrm{d}\omega \geq 0, \forall i \in \{1, 2\} \tag{6}$$

and associate it with the degree of NLOS. The work of Medwin also shows that the diffraction loss is approximately proportional to the degree of NLOS. The sound direction from the diffraction point is given by

$$\theta_d = \theta_1 + \frac{\theta_2 - \theta_1}{L_1 - L_2} L_1. \tag{7}$$

Estimation of Sound Direction from Reflection Signals: Figure **??** shows the proposed approach for estimation of sound direction from reflection signals. Reflection makes the sound propagation and the subsequent target estimation complicated, but if the wall is smooth and yields specular reflection, the sound direction can be estimated easily by introducing a virtual target [**?**], which is located symmetrically to the real target relative to the wall of reflection. Let the position of the virtual target be $\left[\hat{x}^t, \hat{y}^2\right]$. The measured TDOA can be associated with the position of the virtual target as

$$\begin{cases} \left(\hat{x}^t\right)^2 + \left(\hat{y}^t\right)^2 = \left(r + c\Delta t_d\right)^2 \\ \left[\left(\hat{x}^t\right)^2 - d_{12}\right]^2 + \left(\hat{y}^t\right)^2 = r^2 \end{cases}, \tag{8}$$

where d_{12} is a distance between Microphones 1 and 2. Since r is unknown unlike the diffraction, the two equations with three unknowns, $\hat{x}^t$, $\hat{y}^t$ and r, introduce the relationship between $\hat{x}^t$ and $\hat{y}^t$ through the elimination of r. Derivation attempted as a preliminary study for this project yields the relationship as

$$\left(\hat{y}^t\right)^2 = \left(\frac{d_{12}^2}{c^2\Delta t^2} - 1\right)\left(\hat{x}^t\right)^2 - d_{12}\left(\frac{d_{12}^2}{c^2\Delta t^2} - 1\right)\hat{x}^t + \left[\frac{\left(d_{12}^2 + c^2\Delta t_d^2\right)}{4c^2\Delta t^2} - d_{12}^2\right]. \tag{9}$$

The further mathematical manipulation shows that this equation asymptotically yields the sound direction as

$$\theta_r = \pi - \hat{\theta}_r = \lim_{r\to\infty} \tan^{-1}\frac{\hat{y}^t}{\hat{x}^t} = \cos^{-1}\frac{c\Delta t_d}{d_{12}}. \tag{10}$$

Construction of Joint Observation Likelihood through Data Fusion: While the sound can be better identified in direction rather than distance, it is also possible to make an estimate on how far the sound target is. The proposed approach makes the estimate by utilizing any available information including the magnitude, sound patterns stored in a database, or sound characteristics in a knowledge base and constructs an observation likelihood for each of the diffraction and reflection signals by modeling uncertainties. For the jth pair of microphones, the diffraction and reflection likelihoods are then combined to create an auditory joint observation likelihood via the canonical data fusion formula:

$$l_j^a\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) = l_j^d\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) l_j^r\left(\mathbf{x}_k^t | {}^s\tilde{\mathbf{z}}_k^t, \bar{\mathbf{x}}_k^s, \bar{\mathbf{m}}_k\right) \tag{11}$$

where $l_j^d(\cdot)$ and $l_j^r(\cdot)$ are the diffraction and reflection observation likelihood. Figure **??** illustrates the diffraction and reflection observation likelihoods as well as the joint observation likelihood where the observation likelihood is represented by an ellipsoid indicating a probability distribution with a covariance. The diffraction and reflection likelihoods are shown to have high eccentricity due to more accuracy in direction than in distance. Since the difference of the diffraction and reflection likelihoods in orientation may not be significant, the resulting auditory joint likelihood could also given by an ellipsoid with high eccentricity, but the proposed approach, utilizing the diffraction and reflection physics of sound, could estimate the location of the sound target.

IV. CONCLUSIONS

APPENDIX

Appendixes should appear before the acknowledgment.