Automated Heart and Lung Auscultation in Robotic Physical Examinations

Yifan Zhu[1], Alexander Smith[2], and Kris Hauser[1]

***Abstract*—This letter presents the first implementation of autonomous robotic auscultation of heart and lung sounds. To select auscultation locations that generate high-quality sounds, a Bayesian Optimization (BO) formulation leverages visual anatomical cues to predict where high-quality sounds might be located, while using auditory feedback to adapt to patient-specific anatomical qualities. Sound quality is estimated online using machine learning models trained on a database of heart and lung stethoscope recordings. Experiments on 4 human subjects show that our system autonomously captures heart and lung sounds of similar quality compared to tele-operation by a human trained in clinical auscultation. Surprisingly, one of the subjects exhibited a previously unknown cardiac pathology that was first identified using our robot, which demonstrates the potential utility of autonomous robotic auscultation for health screening.**

***Index Terms*—Telerobotics and teleoperation, planning under uncertainty, medical robots and systems**

I. Introduction

INFECTIOUS diseases pose substantial risks to healthcare providers and there have been worldwide initiatives to use robots and automation to help combat the COVID-19 pandemic [1]. In particular, performing physical examinations with robots has potential to reducing the risks to healthcare providers by minimizing person-to-person contact. Moreover, robotic health screening can also help promote preventative care and affordable routine checkups, particularly in rural, remote, and low-resource communities.

Motivated by the benefits of automating physical examinations, this letter explores the task of auscultation, i.e., listening to heart and lung sounds via a stethoscope, which is an important screening procedure that is performed at regular checkups and for patients exhibiting respiratory and cardiac symptoms. This letter demonstrates the first robotic system capable of performing automated heart and lung auscultation, based on the TRINA robot shown in Fig. 1. Prior tele-medicine robotic systems have been built for performing auscultation and echocardiography [2, 3] through tele-operation. Compared to tele-operation, automated medical exams offer several potential benefits including lower operation time, a shorter learning curve for physicians, and reduced cognitive load and tedium. Auscultation is, however, a challenging process to automate. The quality and interpretability of sounds produced by the anatomical structure of interest at the skin surface is highly variable and depends on the location, pressure, and steadiness of the stethoscope. If the stethoscope is placed at a poor location in relation to the patient's internal anatomical structures (e.g., over a rib or deep layers of fatty tissue), the sound may be attenuated or muffled in frequencies that are critical for diagnosis. A human doctor uses visual, tactile, and audio feedback as well as anatomical information and diagnostic expertise from prior medical training to localize informative listening locations.

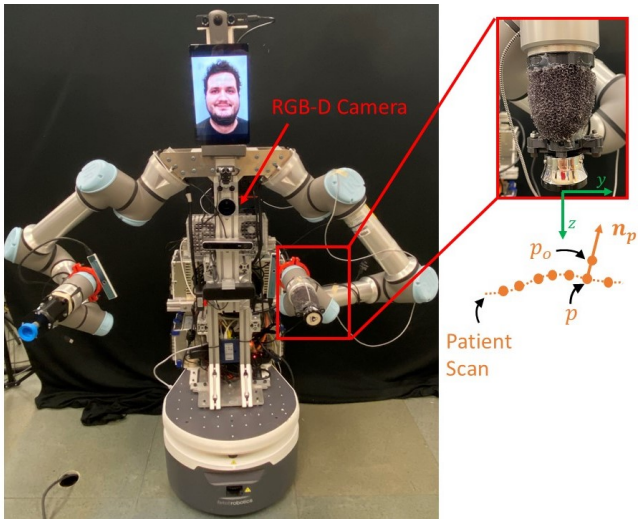


Fig. 1. The TRINA robot used in the experiments, with the zoomed-in view of the stethoscope on the right. The stethoscope frame is shown in green, and the patient scan point cloud in orange. [Best viewed in color.]

The goal of our system is to record high-quality sounds that a human doctor will listen to, so the robot's aim is to provide sounds of diagnostic utility rather than to perform diagnosis itself. First, our system captures a 3D point cloud scan of the patient, registers a human body model, estimates the locations of key anatomical landmarks, and produces a prior map of high-quality auscultation locations. It then adopts informative path planning using audial feedback to adaptively search over the region of interest for a high-quality auscultation location. Audial feedback relies on *sound quality estimators* trained on a database of heart and lung stethoscope recordings. To determine the optimal sensing location we formulate a Bayesian Optimization (BO) problem where the unknown sound quality field is estimated as a semi-parametric

Manuscript received: September 9, 2021; Accepted January, 11, 2022. This letter was recommended for publication by Associate Editor Editor Massimiliano Zecca and Editor Jessica Burgner-Kahrs upon evaluation of the Reviewers' comments. This work was supported in part by NSF under Grant 2025782. *(Corresponding author: Yifan Zhu.)*

[1]: Y. Zhu and K. Hauser are with the Departments of Computer Science, University of Illinois at Urbana-Champaign, IL, USA. {yifan16, kkhauser}@illinois.edu

[2]: A. Smith is with the Carle Illinois College of Medicine, IL, USA. ads10@illnois.edu

Digital Object Identifier (DOI): see top of this page.

residual Gaussian Process (SPAR-GP) model, with a prior map that depends on latent translation offset and sound quality scaling parameters.

Experiments on 4 healthy male human subjects demonstrate that our system performs heart and lung auscultation automatically, and locates sounds of similar diagnostic quality and execution time as tele-operation by a human trained in auscultation in medical school, and trained to use the robot by the investigators. The procedure is reliable and largely required no supervision, except for one subject where a perception error required anatomical landmarks to be input manually before ausculation. Although each subject reported good health and no respiratory or cardiac ailments, in a surprising discovery one subject exhibited a heart murmur that was identified during these experiments. The subject was thereafter referred to a physician for follow-up tests.

II. Related Work

A. Robotic Remote Auscultation and Sonography

Tele-nursing robots consisting of a mobile base and arms have demonstrated a variety of nursing tasks including handovers, vital signs monitoring, disinfection, and auscultation [3, 4, 5, 6]. Robotic technologies employed in the fight against COVID-19 and infectious diseases are reviewed in recent surveys [7, 8].

Giuliani et al. develop a robot and tele-operation user interface for echocardiography [2]. Several works have studied performing robot remote lung sonography using the MGIUS-R3 robotic tele-echography system produced by MGI Tech Co, Ltd. [9, 10, 11, 12]. Marthur et al. develop a semi-autonomous robotic system to perform trauma assessment remotely, where the locations on a patient to be assessed are identified automatically via perception and sonography is performed with tele-operation [13]. Instead of tele-operation, our proposed method performs auscultation fully automatically, which has the potential to reduce physician learning curve, cognitive load, and tedium. On a level of autonomy (LoA) scale from 0 (Full Manual) to 5 (Full Autonomy) [14], our system achieves a level of 4 (High-level Autonomy) because the robot autonomously performs the procedure while an operator monitors and intervenes when necessary.

B. Informative Path Planning

Our problem falls under the category of informative path planning (IPP), where the robot plans information-gathering paths given a probabilistic model of the quantity of interest. Sensor placement, active learning, and adaptive sampling problems can be seen as special instances of the IPP problem. IPP has been applied to a variety of robotics applications such as object tactile exploration [15], palpation-based tissue abnormalities detection [16, 17, 18], inspection and environment mapping [19, 20, 21].

Most related to ours are the works by Salman et al. [16] and Ayvali et al. [17]. Salman et al. propose both discrete and continuous trajectory search algorithms for robots to search for the boundaries of tissue abnormalities, i.e, high stiffness, through robotic palpation. Gaussian process (GP) regression

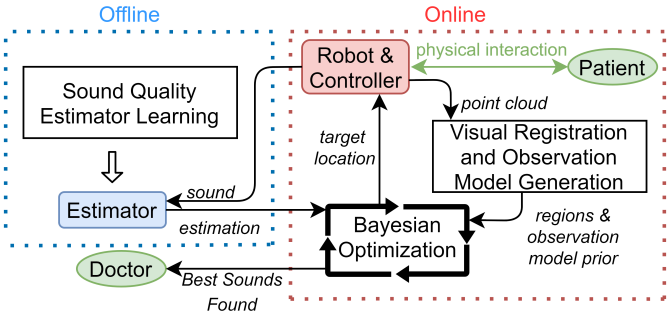


Fig. 2. The workflow of our method.

is used to model the distribution of stiffness. Four methods, including BO, active area search, active level sets estimation, and uncertainty sampling are compared, among which active area search had the best performance. Ayvali et al. also perform robot palpation to detect tissue abnormalities, using GP regression and BO. To inject user prior knowledge of the locations of tissue abnormalities, they add a decaying utility function whose value peaks at the user-provided locations to the acquisition function of BO. In contrast, our work adopts a discrete search BO to the auscultation setting, and introduces two technical contributions to make auscultation more practical. First, visual registration of anatomical landmarks eliminates the need for manual input to specify priors, except as a backup for registration failures, and second, the use of our SPAR-GP model leads to faster convergence than pure residual GP models.

Belonging to the family of semi-parametric regression methods, SPAR-GP performs function approximation with a combination of parametric functions and Gaussian process. In robotics, similar models have been applied to system identification of linear and nonlinear dynamical systems [22, 23, 24], but to our knowledge semi-parametric models have not been adopted in the informative path planning setting.

III. Method

The overview of our method is shown in Fig. 2. In the *offline phase*, we build a dataset of heart and lung stethoscope recordings of humans with labeled sound qualities and train heart and lung sound quality estimators (Section III-B). The goal of our system is to generate high (estimated) quality heart and/or lung recordings by placing a stethoscope at anatomically relevant portions of a patient's chest and back, while keeping the number of auscultation actions small. These estimators are the only feedback used by the system during the online exam; we do not assume human raters are available to provide ground truth.

During the *online phase*, the robot first performs visual registration of the patient to a reference human model with labeled *clinical auscultation locations*. An *observation model prior* is constructed based on the visual registration (Section III-C), which encodes anatomical prior information about expected sound quality. Finally we use BO (Section III-A) to select locations to auscultate to find the best sounds utilizing both the learned sound quality estimators and prior information from

visual cues. These targets are then used for motion control (Section III-D).

A. BO Formulation

We use BO to search adaptively for an auscultation location that yields a high quality sound within a specified anatomical region. We denote sound quality estimators for heart and lung sounds as $e_s(r)$, where s is an anatomical structure (heart or lung) and r is a stethoscope recording. The BO approach computes a probabilistic estimate of the unknown field $e_s(r(x))$ across the patient surface using a SPAR-GP model, which is a sum of the observation prior and the GP. An acquisition function is optimized to yield the new auscultation location. Given a new observation, the estimate is re-fit to the data and the process repeats until a termination criterion is met.

1) SPAR-GP: A GP models an unknown function f as a collection of random variables $f(x)$ which are jointly Gaussian when evaluated at locations x. A GP is fully specified by its mean function $m(\cdot)$, which we set to 0, and covariance function $k(\cdot,\cdot)$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

Given n existing observed function values $\bar{\mathbf{y}} = [y_1, \ldots, y_n]$ at $\bar{\mathbf{x}} = [x_1, \ldots, x_n]$, GP regression predicts the function values at new point x^* as a Gaussian distribution:

$$P(f(x^*)|\bar{\mathbf{x}}, \bar{\mathbf{y}}, x^*) \sim \mathcal{N}(\mathbf{k}\mathbf{K}^{-1}\bar{\mathbf{y}}, k(x^*, x^*) - \mathbf{k}\mathbf{K}^{-1}\mathbf{k}^T)$$

Here,

$$\mathbf{K} = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, k_n) \end{bmatrix} + \sigma^2\mathbf{I}$$

$$\mathbf{k} = \begin{bmatrix} k(x^*, x_1), & \cdots, & k(x^*, x_n) \end{bmatrix},$$

where σ is the standard deviation of noise at an observation.

For each anatomical location, the observation model is a sum of a parametric prior mean function $\mu_\theta(x)$ and a GP residual function $f_s(x)$: $e_s(r(x)) \approx \mu_\theta(x) + f_s(x)$. Since the GP models residuals with respect to the prior, we subtract the prior from the sound quality as the GP observations: $y_i = e_s(r(x_i)) - \mu_\theta(x_i)$. Here we further denote the history of the estimated sound qualities as $\bar{\mathbf{e}} = [e_1, \ldots, e_n]$.

Although the anatomical reference provides sound quality peaks assuming an average human and perfect registration of anatomical landmarks, the robot's prior should capture the uncertainty in visual registration error and the effect a patient's body type has on the magnitude of the overall sound quality. Therefore, we make the prior μ_θ a parametric function of the latent variables θ representing the translation offset and sound quality scaling, and infer θ from observed sound qualities. In particular, a reference quality map $\mu_o(x)$ is first generated from visual registration and θ is initialized to θ_o, which includes zero translation offset and scaling of 1. The exact composition of the prior mean function is deferred to Section III-C. θ is inferred after each reading using the history $\bar{\mathbf{x}}$ and $\bar{\mathbf{e}}$ with maximum a posteriori (MAP) estimation. We use a likelihood function $\mathcal{L}(\theta|\bar{\mathbf{x}}, \bar{\mathbf{e}}) = \prod g(e_i|\mu_\theta(x_i), \sigma^2)$, where $g(\cdot|\mu_\theta(x_i), \sigma^2)$ is the probability density function of the Gaussian distribution $\mathcal{N}(\mu_\theta(x_i), \sigma^2)$. The prior of θ, $h(\theta)$, follows a multivariate Gaussian distribution $\mathcal{N}(\theta_o, \Sigma)$. We solve the MAP estimation problem by maximizing the posterior, using a standard numerical optimization solver:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(\theta|\bar{\mathbf{x}}, \bar{\mathbf{e}})h(\theta) \tag{1}$$

2) BO: Letting A be a region of interest on the patient surface near structure s, BO aims to solve:

$$\max_{x \in A} e_s(r(x)) \tag{2}$$

Since heart and lung sounds need to be listened separately and there are usually multiple parts of the anatomical structures that are of interest, we solve Eqn. 2 for each anatomical structure s separately, and for each s sequentially for their regions of interest. We also share the same GP across the entire chest or back of a patient for the same s.

In each iteration of BO, an acquisition function is used to determine the next location to observe the data. Two popular choices are expected improvement (EI) and upper confidence bound (UCB). Let the posterior mean and variance of the GP be $\mu_{\bar{y}}(x), \sigma^2_{\bar{y}}(x)$, then EI is defined as:

$$EI(x) = \mathbb{E}[\max(0, \mu_\theta(x) + \mu_{\bar{y}}(x) - e^*)], \tag{3}$$

where e^* is the best observed quality so far, and EI can be calculated in closed form [25]. UCB is defined as:

$$UCB(x) = \mu_\theta(x) + \mu_{\bar{y}}(x) + \beta\sigma_{\bar{y}}(x), \tag{4}$$

where β is an exploration weight that regulates how much bonus is given to uncertainty in the prediction.

The overall algorithm is outlined in Algorithm 1. For each iteration in each region, a point with the largest acquisition function value is first chosen to be observed, where $\xi(x)$ represents the acquisition function, and the quality is estimated. In particular, A contains sampled points on a patient surface, and the maximization is performed in a brute force fashion by calculating the acquisition function values across all points in A and selecting the maximum. In addition, the same point is observed at most once. Then we add the estimated qualities to the history and update the prior mean function by solving the MAP estimation, after which the residuals for GP are calculated, observations are added, and the GP is re-fit. For the termination criteria, we allow early termination if the estimated quality in a region is above an adequacy threshold for making a diagnosis. We also set the budget N_{max} to a reasonable number of auscultations per region to limit the maximum time of the procedure.

B. Sound Quality Estimator

1) Dataset: We first collect lung and heart stethoscope recordings of various qualities with TRINA from the researchers on this project. Each recording lasts 10 s, and we vary the locations on the subject's chest, whether or not the subject is wearing clothing, and whether the subject is taking a deep breath. 80 recordings are collected each for

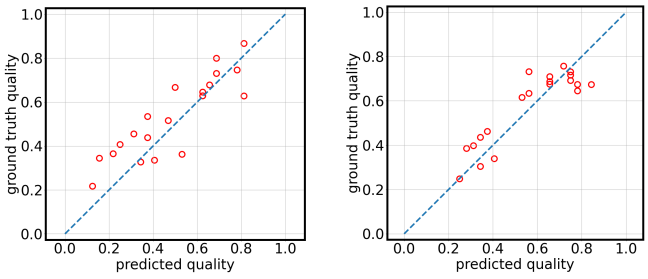
Algorithm 1: Bayesian Optimization Auscultation

Input: Structure s (heart/lung), prior μ_θ, regions A, max iterations N_{max};
Initialize $\bar{\mathbf{x}} = \{\}$, $\bar{\mathbf{y}} = \{\}$, $\bar{\mathbf{e}} = \{\}$;
for *all regions* A *near* s **do**
  for $k = 1, ..., N_{max}$ **do**
    $x_k \leftarrow \arg\max_{x \in A} \xi(x)$;
    if *region termination criteria met* **then**
      Go to next region;
    else
      Auscultate at x_k, obtain sound quality e_k;
      Set $\bar{\mathbf{x}} \leftarrow \bar{\mathbf{x}} \cup \{x_k\}$, $\bar{\mathbf{e}} \leftarrow \bar{\mathbf{e}} \cup \{e_k\}$;
      $\theta \leftarrow \text{argmax}\, \mathcal{L}(\theta|\bar{\mathbf{x}}, \bar{\mathbf{e}})h(\theta)$;
      Set $\bar{\mathbf{y}} \leftarrow \bar{\mathbf{y}} \cup \{e_k - \mu_\theta(x)\}$;
      Re-fit GP;
return recordings of max quality in each region;

heart and lung sounds. The ground truth qualities of the recordings were labeled by 4 medical school students who have undergone auscultation training in a Liaison Committee on Medical Education (LCME)-accredited medical program and have been applying their auscultation skills in clinical settings for at least one year. We note that prior studies suggest that the cardiac examination skills of trained medical school students are no worse and may even be better than those of experienced doctors [26]. Each label is a score between 0-1 with increments of 0.125 based on the rating guidelines provided by Grooby et al., where 0 means no detectable heart or lung signal, and 1 means clear heart or lung sounds with little to no noise [27]. A recording of poor quality either contains weak signals or has noise that obscures the signals. The noise typically comes from the stethoscope contacting a patient and heavy breathing sounds obfuscating heart sounds.

We use intra-class correlation (ICC) [28] to measure the inter-rater reliability of the labels. We obtain the ICC estimates using a mean-rating, absolute-agreement, 2-way mixed-effects model, using the Pingouin package at https://pingouin-stats.org/. The calculated ICC estimates are 0.925 and 0.679, respectively for heart and lung, which correspond to excellent and moderate agreement [28]. The average variances for the 4 ratings of each stethoscope recording's quality are 0.0163 and 0.0400 for heart and lung respectively. The max variances are 0.0938 and 0.151 for heart and lung. The average of the 4 ratings of each recording is used as the quality label. We further augment the dataset with 5 synthetic recordings of pure noise with varying amplitudes and labeled with quality 0, which emulate erroneous conditions like placing a stethoscope on irrelevant parts of the body, rubbing the stethoscope, bumping the stethoscope, and failing to make contact.

2) Feature Extraction: We follow the method of Grooby et al. for heart and lung quality classification and modify it for regression. For each sound recording, we apply noise reduction [29], band-pass filter with cutoff frequency 50-250 Hz for heart recordings and 200-1000 Hz for lung recordings, and extract for all sound recordings the top features listed by Grooby et al. In particular, we use the top 5 features for heart sound and the top 2-6 features for lung sound (top 1 feature for lung is not used due to the lack of open-source code). The extracted features are then subsequently used for training the estimators.



(a) Heart (b) Lung

Fig. 3. Trained heart and lung sound quality estimator predictions on the testing set, where the horizontal axis is the predictions and the vertical axis the ground truth labels.

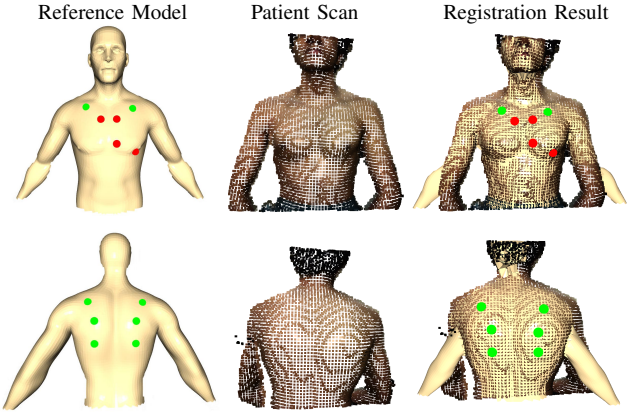


Fig. 4. Left to right: labeled reference models, subject ("patient") scan point clouds, and result of nonrigid registration. Auscultation locations for heart are labeled in red, and lungs in green.

3) Learning: We use the TPOT AutoML algorithm [30] to train the heart and lung sound quality estimators that take in the extracted features and predict the sound qualities. We split the dataset and use 25% as testing data. For the TPOT algorithm, we set generations = 100, population size = 100, and set the rest of the hyperparameters as default. We achieve a 0.0945 and 0.0733 mean absolute errors (MAE) on the testing set of heart and lung sounds respectively. The predictions on the testing set are shown in Fig. 3.

C. Visual Registration and Sound Quality Prior

Doctors are trained to auscultate typical points on a patient body as illustrated in Fig. 4, both for optimal sound quality and ability to diagnose abnormalities on specific structures of the heart and lung. The goal of visual registration is to locate these points on a patient, both to define regions for ausculation and to obtain a prior observation model. Note that we do not track the patient body after this initial registration is performed and assume that the patient body does not move significantly during auscultation. Small movements are accounted for by the motion controller (Sec. III-D)

1) Visual Registration: We first manually label a reference human mesh model with auscultation locations, which we register onto a 3D point cloud of a patient captured with a

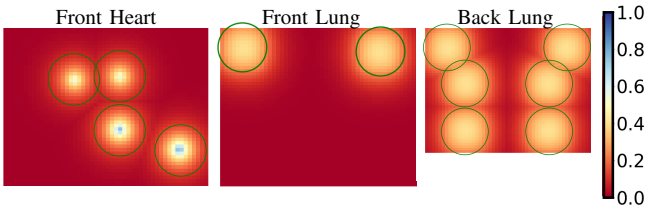


Fig. 5. Observation model priors for heart and lung on the front of the patient and lung on the back. Each region A is denoted with green circles. [Best viewed in color.]

RGB-D camera. In all of our experiments, we downsample the captured point cloud with a voxel size of 0.01 m, which strikes a balance between computational time and enough coverage by using a stethoscope whose diameter is approximately 0.02 m. We first perform rigid registration with RANSAC [31] followed by ICP [32] to provide a good initial alignment, then nonrigid registration with nonrigid ICP proposed by Amberg et al. [33], where a local affine deformation model is used. The initial and final results of one case of visual registration is shown in Fig. 4. The regions of interest on the patient surface are limited to a certain distance R from the estimated clinical auscultation locations (Fig. 5, green circles).

2) Point Cloud Projection and Sound Quality Prior: We project the point cloud onto the coronal plane of a patient, and BO only searches amongst locations on this projected plane. The third dimension (height) of the point cloud is associated with each projected point to de-project selected locations on the plane back into 3D space. As discussed in Section III-A, to obtain the prior of the observation model, we first generate a reference quality map $\mu_o(x)$, where we place a negative exponential function at each of the registered auscultation locations in the projected coronal plane. An example of the initial observation model prior is shown in Fig. 5. The latent parameters $\theta = [t_x, t_y, c]$ is initialized to $\theta_o = [0, 0, 1]$, where t_x and t_y are the translation offsets on the projected plane, and c the quality scaling term.

D. Motion Control

The motion control system generates safe interaction between the stethoscope and the patient, keeping the interaction force at a desired auscultation force F_{aus}. According to both Nowak et al. [34] and our observations in experiments, auscultation force F_{aus} has a small effect on sound qualities. We set $F_{aus} = 7N$, which is both comfortable for human subjects and provides good sound qualities in practice. The robot starts at a manually defined home configuration q_{home}. For a location x selected by BO, we first get the point p on the patient point cloud where x is projected from. To auscultate at p, the robot first moves the stethoscope to an initial position p_o that is a distance $d_o = 0.08\,\text{m}$ from p in the direction of outward surface normal $\mathbf{n_p}$ at p to account for perception uncertainty, shown in Fig. 1. In particular, we set the target position of the stethoscope frame to be p_o, with the z-axis aligning with $-\mathbf{n_p}$ and solve for the target robot configuration q_0 with inverse kinematics (IK). The robot moves from q_{home} to q_o linearly in the Cartesian space, following maximum linear and angular velocities $v_{max} = 0.065\,\text{m/s}$ and $\omega_{max} = 1.0\,\text{rad/s}$. Starting from p_o, the robot moves the stethoscope in the $-\mathbf{n_p}$ direction at constant speed $s = 0.03\,\text{m/s}$, during which we also adopt impedance control for enhanced safety, where we utilize the wrist-mounted force-torque sensor on the arm as wrench feedback and control the stethoscope as a mass-spring-damper. The robot keeps moving the stethoscope until the external force in the $\mathbf{n_p}$ direction reaches F_{aus}, at which point we record the stethoscope audio for 10 s and use the learned estimators to estimate the sound quality. After auscultation is finished, the robot first moves back to q_o, waits for BO to give the next auscultation location, and moves to the next initial position with a linear Cartesian movement under velocity limits.

IV. Results

Throughout the experiments, we use the TRINA robot shown in Fig. 1. TRINA is a mobile bimanual manipulator with various visual sensors and swappable end-effectors. In this project, we use a Thinklabs digital stethoscope mounted on the left arm with firm foam as sound insulator. We use the Intel Realsense L515 RGB-D camera on TRINA to capture point clouds. For both the simulation and physical experiments, we use the squared exponential covariance function for GP with length scale $= 0.02$. We chose this length based on our observations of how sound qualities correlate across the surface of the body.

A. Simulation Experiment

First we compare the performance of different acquisition functions, and evaluate the efficacy of SPAR-GP. To do this, we compare the results of the BO algorithm under visual registration error and patient overall sound quality variations. We first generate a ground truth heart quality map across the patient chest in a similar fashion as the observation model generation in Section III-C, with the same shape as Fig. 5 but having the negative exponential functions placed at ground truth locations. Then for each setting of acquisition function and prior observation model, we generate a prior by randomly shifting and scaling the ground truth quality map and run the BO algorithm. In particular, the random shift (x,y sampled independently) is uniform between -0.02 m and 0.02 m and the random scale is uniform between 0.7 and 1.3. We run this 50 times and compare the average maximum observed qualities across different regions under a given budget N_{max}. Simulations are performed for the heart prior only, giving a total of 4 regions. We disable early termination (i.e., Line 6 of Alg. 1) in these experiments.

We set the GP observation noise $\alpha = 0.0417$, and region radius $R = 0.03\,\text{m}$. The covariance matrix Σ of the prior parameters θ is a diagonal matrix with entries $\sigma^2_{t_x} = \sigma^2_{t_y} = 1.33e{-}4$ (variance of the uniform random distribution of [-0.02,0.02]), and $\sigma^2_c = 0.03$ (variance of the uniform random distribution of [0.7,1.3]).

Results are summarized in Table I, where we compare 1) acquisition functions EI and UCB with different β parameters under different budgets for each of the regions; 2) BO with GP under Zero prior, Residual-GP with Fixed prior, and SPAR-GP

TABLE I
EFFECTS OF BO PARAMETERS ON MAX SOUND QUALITY IN SIMULATION

	$N_{max} = 3$			$N_{max} = 10$		
	Zero	Fixed	SPAR	Zero	Fixed	SPAR
EI	0.432	0.434	0.549	0.622	0.620	0.638
UCB, $\beta = 0.5$	0.370	0.428	0.550	0.629	0.624	0.660
UCB, $\beta = 1.0$	0.421	0.423	0.559	0.632	0.628	0.660
UCB, $\beta = 1.5$	0.455	0.420	0.544	0.631	0.624	0.657

TABLE II
STATISTICS OF THE 4 HUMAN SUBJECTS

	Subject 1	**Subject 2**	**Subject 3**	**Subject 4**
Weight (kg)	95.6	57.4	60.5	120
Height (m)	1.84	1.82	1.81	1.83
BMI	28.2	17.3	18.5	35.8
Age	33	22	21	28

(SPAR) prior. Numbers indicate the maximum sound quality over each region. SPAR-GP consistently outperforms BO with fixed prior and no prior, particularly when the budget of observations is small. In addition, there is no clear winner among the acquisition functions, so for the subsequent physical experiments we use EI because it is parameter-free.

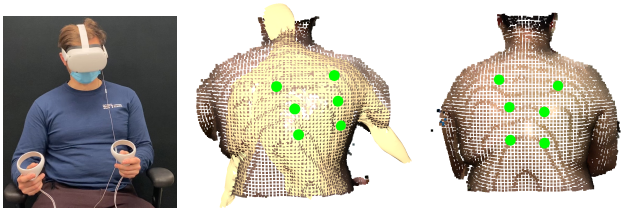
B. Physical Auscultation Experiments

We compare BO against two other baselines for a complete auscultation session that includes both heart and lung. In RO baseline, the robot automatically auscultate at the registered auscultation landmarks on the patient. In the DT baseline, a human tele-operates the robot to auscultate the patient, shown in Fig. 6. We perform each of the three methods once on 4 healthy male human subjects[1], whose statistics are listed in Table II.

1) Direct Teleoperation (DT): In the DT condition, an expert tele-operator performs auscultation at the specified regions, and adjusts the position of the stethoscope until he/she judges the sound quality to be good enough for making a diagnosis. The tele-operator is one of the medical school students that labeled the stethoscope recording dataset, and the tele-operator received tele-operation training on TRINA for 3 hrs prior to the experiments. Tele-operation on TRINA is achieved via the Oculus Quest VR headset and controllers (Fig. 7(a)), where the VR headset streams the stereo camera on TRINA's head, and the 6D pose of the stethoscope on TRINA is driven in velocity control mode to follow the velocity of one of the VR controllers while a clutch button is depressed.

2) Bayesian Optimization (BO): For BO, we set the acquisition function to EI, $N_{max} = 4$, GP observation noise α = 0.0938 for heart and 0.151 for lung. Σ is a diagonal matrix with entries $\sigma^2_{t_x} = 2.12e{-}4$, $\sigma^2_{t_y} = 9.77e{-}5$, and $\sigma^2_c = 0.03$. In addition, We terminate early if sound quality exceeds 0.5. This threshold was judged by the 4 raters as the minimum quality for making a confident diagnosis.

[1]This human subjects study was fully reviewed and approved by the University of Illinois at Urbana-Champaign Institutional Review Board, with IRB#21849.



(a) (b) (c)

Fig. 6. (a) Tele-operator performing auscultation using VR headset and controllers. (b) Automatic registration failed on the back of subject 4 due to difference in their body type from the reference model. (c) Manual registration on the back of subject 4.

The values for α correspond to the maximum inter-rater variance for each recording in the quality estimator dataset. The values for R and Σ were determined empirically as a function of estimated registration error. We estimate error using a public dataset of complete human scans [35]. We select 30 male human mesh models that cover a wide range of body types, with maximum, minimum, and average body mass index (BMI) 34.9, 17.8, and 26.5, respectively. We manually label the nipples, segment the mesh to emulate a RGB-D scan, and apply visual registration. The registration MAE error of the nipples is 0.0173 m, and the maximum error is 0.0389 m, which we set as R.

3) Results: Table III breaks down the time spent on visual registration (Reg.), the total amount of time spent during the auscultation session (Total), the total number of auscultations (No.), the average of the maximum rated quality in each region (Avg. Max), and the minimum of the maximum rated quality across all regions (Min Max). Note that sound qualities are average ratings given by the the human raters, not values from the estimator. One example of the final auscultated locations for the heart and lung, and the posterior sound quality estimates are shown in Fig. 7.

TABLE III
EXPERIMENTAL RESULTS ON HUMAN SUBJECTS

Subject	**Method**	**Reg. (s)**	**Total (s)**	**No.**	**Avg. Max**	**Min Max**
1	BO	171	538	22	**0.727**	**0.438**
	RO	171	370	12	0.648	0.312
	DT	0	656	19	0.651	0.281
2	BO	183	496	18	0.716	**0.531**
	RO	183	434	12	0.651	0.312
	DT	0	559	14	**0.804**	**0.531**
3	BO	177	1155	37	0.552	**0.375**
	RO	177	442	12	0.484	0
	DT	0	613	16	**0.576**	0.25
4*	BO	148	694	21	0.635	0.469
	RO	148	495	12	0.599	0.375
	DT	0	468	14	**0.711**	**0.625**

*: Manual registration was performed on the back of the subject.

We use a linear mixed effects model to evaluate the methods, where we set the best estimated sound quality found in each region to be the dependent variable, methods to be fixed effects, and structure s (heart/lung), auscultation region, and subject to be random effects. The results are presented in Table IV. While the estimated fixed effects coefficients show

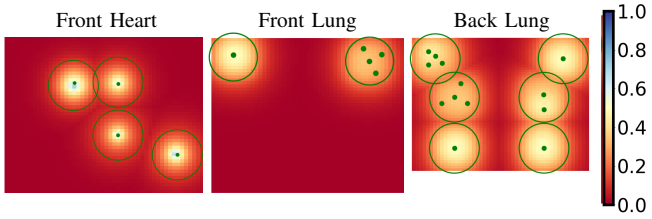


Fig. 7. Posterior sound quality predictions for heart, front lung, and back lung on Subject 1. Observed locations are marked as green dots.

TABLE IV
MIXED EFFECTS MODEL RESULTS (β:FIXED EFFECTS COEFFICIENT, CI: CONFIDENCE INTERVAL, s:ANATOMICAL STRUCTURE)

	Fixed Effects		Variances of Random Effects			
Method	β	**95% CI**	s	**Region**	**Subject**	**Residual**
BO	0.662	[0.556,0.768]				
RO	0.598	[0.492,0.704]	3.40e-12	1.77e-3	6.03e-3	1.96e-2
DT	0.703	[0.597,0.859]				

separations among the methods, they are not strong compared to the confidence intervals. Therefore, we perform a post-hoc pairwise testing using the Tukey's test [36]. The p-values for the pairs BO-DT, BO-RO, and RO-DT are 0.326, 0.0700, and 0.00110, respectively. In addition, all three random effects have a small contribution to the overall variances, where the residual variance dominate.

To further understand diagnostic utility, we ask the raters to answer a series of yes-no questions regarding the recordings with the best estimated quality in each region. For heart sounds, the key features doctors look for are the first and second heart sound (S1 and S2), while for lung sounds, doctors look for inspiration and expiration. Therefore, for heart, *Q1* and *Q2* are whether the raters are able to hear S1 and S2, respectively. For lung, we ask whether they can hear inspiration (*Q1*) and expiration (*Q2*). Then for both heart and lung recordings, *Q3* asks whether any sounds not originating from the target anatomy could be heard, and *Q4* asks whether continuous noise or static obfuscated the heart/lung sounds. *Q5* asks whether they can make a diagnosis based on the recording. The results are shown in Table V.

TABLE V
% OF YES RESPONSES FOR DIAGNOSTIC UTILITY OF RECORDINGS HIGHER IS BETTER FOR Q1, Q2, AND Q5

	Heart					Lung				
Method	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*
BO	**100**	**98.4**	**12.5**	**26.6**	**87.5**	93.0	86.7	9.38	**23.4**	74.2
RO	87.5	92.2	25.0	31.3	78.1	89.1	74.2	11.8	24.2	66.4
DT	90.6	95.3	32.8	46.9	85.9	**96.9**	**89.8**	**6.25**	31.3	**85.2**

C. Discussion

Although direct teleoperation (DT) produces better sounds than registration-only automated auscultation (RO) ($p = 0.001$), there is no statistically significant difference between DT and Bayesian optimization (BO) ($p = 0.326$). BO obtained higher minimum sound quality on patients 1 and 3 (Table III), and better diagnostic values for heart sounds (*Q5*, Table V). BO was also almost always able to generate S1 and S2 heart sounds (*Q1* and *Q2*, Table V). This could indicate greater consistency than a human operator, but more data is needed to test this hypothesis. The evidence suggests that BO achieves higher qualities than RO but is somewhat weak due to the small sample size ($p = 0.070$). BO does achieve a larger minimum quality and better diagnostic values from Table III and Table V. As we expected, adaptive sensing helps overcome uncertainty in visual registration and anatomical differences. Overall diagnostic value (*Q5*, Table V) roughly matches the results from the mixed effects model, where DT and BO perform comparable on heart sounds but DT outperforms BO on lung sounds, and both outperform RO.

Using BO, the robot takes less execution time than DT on subjects 1 and 2, and similar amount on subject 4. However, on subject 3, BO spent much more time auscultating, mainly due to the poor lung qualities observed on the patient. In fact, the robot ended up using the entire budget N_{max} on all lung locations. Future implementations could skip between regions, and return to past regions if there is sufficient time and need. Moreover, we limited the movement speed of the robot empirically to avoid intimidating the subjects, but further refinement of the motion controller could improve transit times.

For subject 4, the visual registration for the back exam was identified as having poor alignment, shown in Fig. 6. The tele-operator decided to reject the automatic registration, and instead manually labeled the auscultation points during the experiment, as shown in Figs. 7(b) and 7(c). This deviation from fully-automatic procedure is indicated by an asterisk in Table III. The poor result from the visual registration is likely due to the difference between the body type of the subject (BMI = 35.8) and the reference human model. In future work, visual registration can be improved in various ways, such as adopting multiple reference models for different body types. We also note that raters noticed a moderate amount of noise and static (*Q4*, Table V), and we assume this mainly comes from the robot motors because these sounds are not present during manual auscultation. DT was slightly higher than the automated methods, likely due to human tremor.

Finally, a surprising finding during these studies was that both the tele-operator and the raters detected an incidental cardiac pathology on subject 2. During DT, which were performed before BO and RO, the heart sound was described as a systolic III/IV high pitched crescendo-decrescendo murmur localized to the right-upper sternal border. Upon listening with BO and RO, the murmur was no longer present and the auscultation could be described as a normal S1 and S2 with physiologic S2 splitting distinctly noted. This murmur's transient nature suggests a few possible diagnoses, including aortic stenosis, bicuspid aortic valve, or an innocent murmur. Subject 2 was recommended to see a physician for further cardiac workup. This is promising since it shows that robotic auscultation can record cardiopulmonary sounds useful in diagnosis.

V. CONCLUSION AND FUTURE WORK

This letter proposed a method that enables a robotic nursing assistant to automatically perform auscultation. To choose

sensing locations, we show that using a Bayesian optimization that leverages visual prior information of the clinical auscultation locations outperforms choosing locations according to the prior only, and our system is capable of locating sounds of comparable quality with tele-operation by a human with clinical auscultation expertise.

While the results are promising, our experimental study only includes 4 young human subjects. In future work a larger-scale study is needed to evaluate its diagnostic capability on populations likely to exhibit abnormalities. Furthermore, the system is significantly slower than human doctors, and we intend to accelerate it by making movement faster, and reducing visual registration times, and compare it to manual auscultation. We would also like to improve the robustness of visual registration by using reference human models with varying body types. Finally, we would like to integrate our method with diagnostic algorithms to pre-screen for possible abnormalities.

REFERENCES

[1] A. Khamis, J. Meng, J. Wang, A. T. Azar, E. Prestes, H. Li, I. A. Hameed, Á. Takács, I. J. Rudas, and T. Haidegger, "Robotics and intelligent systems against a pandemic," *Acta Polytechnica Hungarica*, vol. 18, no. 5, pp. 13–35, 2021.

[2] M. Giuliani, D. Szczęśniak-Stańczyk, N. Mirnig, G. Stollnberger, M. Szyszko, B. Stańczyk, and M. Tscheligi, "User-centred design and evaluation of a tele-operated echocardiography robot," *Health and Technology*, vol. 10, no. 3, pp. 649–665, 2020.

[3] G. Yang, H. Lv, Z. Zhang, L. Yang, J. Deng, S. You, J. Du, and H. Yang, "Keep Healthcare Workers Safe: Application of Teleoperated Robot in Isolation Ward for COVID-19 Prevention and Control," *Chinese J. Mechanical Engineering*, vol. 33, no. 1, 2020.

[4] Z. Li, P. Moran, Q. Dong, R. J. Shaw, and K. Hauser, "Development of a tele-nursing mobile manipulator for remote care-giving in quarantine areas," in *IEEE Int. Conf. Robotics and Automation*, 2017.

[5] K. Arent, M. Cholewiński, W Domski, M Drwięga, J. Jakubiak, M. Janiak, B. Kreczmer, A. Kurnicki, B. Stańczyk, D Szczęśniak-Stańczyk, *et al.*, "Selected topics in design and application of a robot for remote medical examination with the use of ultrasonography and ascultation from the perspective of the remedi project," *J. Automation Mobile Robotics and Intel. Sys.*, vol. 11, no. 2, pp. 82–94, 2017.

[6] R. Kim, J. Schloen, N. Campbell, S. Horton, V. Zderic, I. Efimov, D. Lee, and C. H. Park, "Robot-Assisted Semi-Autonomous Ultrasound Imaging with Tactile Sensing and Convolutional Neural-Networks," *IEEE T. Medical Robotics and Bionics*, vol. 3, no. 1, pp. 96–105, 2021.

[7] X. V. Wang and L. Wang, "A literature survey of the robotic technologies during the COVID-19 pandemic," *J. Manufacturing Systems*, no. February, 2021.

[8] A. Di Lallo, R. Murphy, A. Krieger, J. Zhu, R. H. Taylor, and H. Su, "Medical Robots for Infectious Diseases: Lessons and Challenges from the COVID-19 Pandemic," *IEEE Robotics and Automation Magazine*, 2021.

[9] K. D. Evans, Q. Yang, Y. Liu, R. Ye, and C. Peng, "Sonography of the Lungs: Diagnosis and Surveillance of Patients With COVID-19," *J. Diagnostic Medical Sonography*, vol. 36, no. 4, pp. 370–376, 2020.

[10] S. J. Adams, B. Burbridge, L. Chatterson, V. McKinney, P. Babyn, and I. Mendez, "Telerobotic ultrasound to provide obstetrical ultrasound services remotely during the COVID-19 pandemic," *J. Telemedicine and Telecare*, 2020.

[11] R. Z. Yu, Y. Q. Li, C. Z. Peng, R. Z. Ye, and Q. He, "Role of 5G-powered remote robotic ultrasound during the COVID-19 outbreak: Insights from two cases," *European Rev. Medical and Pharmacological Sciences*, vol. 24, no. 14, pp. 7796–7800, 2020.

[12] J. Wang, C. Peng, Y. Zhao, R. Ye, J. Hong, H. Huang, and L. Chen, "Application of a Robotic Tele-Echography System for COVID-19 Pneumonia," *J. Ultrasound in Medicine*, vol. 40, pp. 385–390, 2021.

[13] B. Mathur, A. Topiwala, S. Schaffer, M. Kam, H. Saeidi, T. Fleiter, and A. Krieger, "A semi-autonomous robotic system for remote trauma assessment," *IEEE Int. Conf. Bioinformatics and Bioengineering*, pp. 649–656, 2019.

[14] T. Haidegger, "Autonomy for Surgical Robots: Concepts and Paradigms," *IEEE T. Medical Robotics and Bionics*, vol. 1, no. 2, pp. 65–76, 2019.

[15] D. Driess, D. Hennes, and M. Toussaint, "Active multi-contact continuous tactile exploration with gaussian process differential entropy," in *IEEE Int'l Conf. Robotics and Automation*, 2019.

[16] H. Salman, E. Ayvali, R. A. Srivatsan, Y. Ma, N. Zevallos, R. Yasin, L. Wang, N. Simaan, and H. Choset, "Trajectory-Optimized Sensing for Active Search of Tissue Abnormalities in Robotic Surgery," *IEEE Int. Conf. Robotics and Automation*, pp. 5356–5363, 2018.

[17] E. Ayvali, A. Ansari, L. Wang, N. Simaan, and H. Choset, "Utility-Guided Palpation for Locating Tissue Abnormalities," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 864–871, 2017.

[18] R. E. Goldman, A. Bajo, and N. Simaan, "Algorithms for autonomous exploration and estimation in compliant environments," *Robotica*, vol. 31, no. 1, pp. 71–87, Jan. 2013.

[19] G. A. Hollinger, B. Englot, F. S. Hover, U. Mitra, and G. S. Sukhatme, "Active planning for underwater inspection and the benefit of adaptivity," *Int. J. Robotics Res.*, vol. 32, no. 1, pp. 3–18, 2013.

[20] G. Hollinger and G. Sukhatme, "Sampling-based Motion Planning for Robotic Information Gathering," *Robotics: Science and Systems*, vol. 3, no. 5, 2013.

[21] J. Binney and G. S. Sukhatme, "Branch and bound for informative path planning," *IEEE Int'l Conf. on Robotics and Automation*, pp. 2147–2154, 2012.

[22] T. Wu and J. Movellan, "Semi-parametric Gaussian process for robot system identification," *IEEE Int'l Conf. on Intelligent Robots and Systems*, pp. 725–731, 2012.

[23] J. Ko, D. J. Klein, D. Fox, and D. Haehnel, "Gaussian processes and reinforcement learning for identification and control of an autonomous blimp," *IEEE Int'l Conf. on Robotics and Automation*, no. April, pp. 742–747, 2007.

[24] R. Camoriano, S. Traversaro, L. Rosasco, G. Metta, and F. Nori, "Incremental semiparametric inverse dynamics learning," *IEEE Int. Conf. Robotics and Automation*, vol. 2016-June, pp. 544–550, 2016.

[25] E. Brochu, V. M. Cora, and N. de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning," *ArXiv*, 2010.

[26] J. M. Vukanovic-Criley, S. Criley, C. M. Warde, J. R. Boker, L. Guevara-Matheus, W. H. Churchill, W. P. Nelson, and J. M. Criley, "Competency in cardiac examination skills in medical students, trainees, physicians, and faculty: A multicenter study," *Archives of internal medicine*, vol. 166, no. 6, pp. 610–616, 2006.

[27] E. Grooby, J. He, J. Kiewsky, D. Fattahi, L. Zhou, A. King, A. Ramanathan, A. Malhotra, G. A. Dumont, and F. Marzbanrad, "Neonatal Heart and Lung Sound Quality Assessment for Robust Heart and Breathing Rate Estimation for telehealth Applications," *IEEE J. Biomedical and Health Informatics*, vol. 2194, no. c, pp. 1–12, 2020.

[28] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *J. Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[29] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, 2020.

[30] T. T. Le, W. Fu, and J. H. Moore, "Scaling tree-based automated machine learning to biomedical big data with a feature set selector," *Bioinformatics*, vol. 36, no. 1, pp. 250–256, 2020.

[31] M. A. Fischler and R. C. Bolles, "Random Sample Paradigm for Model Consensus: A Apphcatlons to Image Fitting with Analysis and Automated Cartography," *Graphics and Image Processing*, vol. 24, no. 6, pp. 381–395, 1981.

[32] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Sensor Fusion IV: Control Paradigms and Data Structures*, P. S. Schenker, Ed., vol. 1611, Apr. 1992, pp. 586–606.

[33] B. Amberg, S. Romdhani, and T. Vetter, "Optimal Step Nonrigid ICP Algorithms for Surface Registration," in *IEEE Conf. Computer Vision and Pattern Recognition*, vol. 8, IEEE, Jun. 2007, pp. 1–8.

[34] K. M. Nowak and L. Nowak, "On the relation between pressure applied to the chest piece of a stethoscope and parameters of the transmitted bioacoustic signals," in *Int. Congress on Acoustics*, 2016.

[35] Y. Yang, Y. Yu, Y. Zhou, S. Du, J. Davis, and R. Yang, "Semantic Parametric Reshaping of Human Body Models," *Int. Conf. 3D Vision*, pp. 41–48, 2015.

[36] W. Haynes, "Tukey's Test," in *Encyclopedia of Systems Biology*, New York, NY: Springer New York, 2013, pp. 2303–2304.