Estimating the Motion of Drawers From Sound

Manuel Baum*[1,2] Amelie Froessl*[1] Aravind Battaje[1,2] Oliver Brock[1,2]

***Abstract*— Robots need to understand articulated objects, such as drawers. The state of articulated structures is commonly estimated using vision, but visual perception is limited when objects are occluded, have few salient features, or are not in the camera's field of view. Audio sensing does not face these challenges, since sound propagates in a fundamentally different way than light. Therefore we propose to fuse vision and audio sensing to overcome the challenges faced by vision alone. We estimate motion in several drawers and show that an audio-visual approach estimates drawer motion more reliably than only vision – even in settings where the purely visual approach completely breaks down. Additionally, we perform an in-depth analysis of the regularities that govern how motion in drawers shapes their sound.**

I. Introduction

Doors creak and make loud thuds, drawers rumble while opening, and even twisting lids off jars produces distinct sounds. Different articulated mechanisms produce different sounds. But the sound they make also varies with their motion—a quickly opened drawer sounds different from the same opened slowly. Thus, we can infer motion of articulated objects by their characteristic sounds.

In this paper, we introduce auditory perception of kinematic structures to robotics. We analyze the regularities that govern how moving drawers emit sound and find that drawers' velocity correlates positively with amplitude and pitch of sound. We show how to use these regularities in an audio-visual kinematic structure estimation system.

Our system enables a robot to perceive the kinematic structure of its environment in a wider range of experimental settings. This is because sound is complementary to vision—sound propagates in a fundamentally different way than light. It is less affected by occlusions and bad lighting conditions. It works even when there is no light at all. This allows us to use sound to estimate object motion in settings that are challenging to purely vision-based methods.

II. Related Work

We will first review approaches to robotic object perception that are based on sound in Section II-A. Then we review existing approaches to kinematic structure estimation in Section II-B. Our main evaluation criterion is going to be how robust these approaches are to adverse experimental conditions.

* These authors contributed equally.
[1] Robotics and Biology Laboratory, Technische Universität Berlin
[2] Science of Intelligence (SCIoI), Cluster of Excellence, Berlin, Germany
We gratefully acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

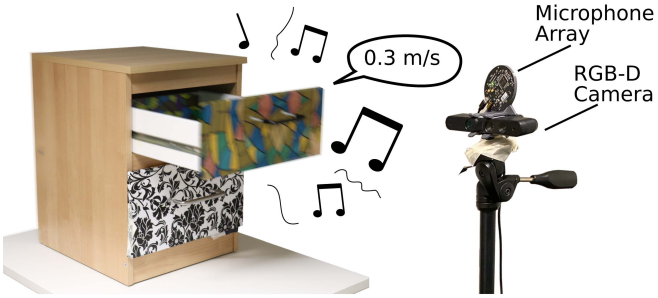

Fig. 1. We use audio-visual information to robustly estimate the kinematic state of a drawer. By fusing vision (RGB-D camera) and audio (two microphones), our method is able to track the state of the drawer even when we switch the light off or visually occlude the drawer. We evaluate the estimation using MOCAP tracking.

A. Estimating Object Properties From Sound

Sound can be used to recognize and classify objects [1], [2], [3], [4], [5], as well as to estimate the properties of objects such as their material [6], shape, height [7], as well as the contents' of a container [8]. The velocity of cars can also be estimated from sound [9]. These works show that audio is a useful sensor modality to make inference about objects, their motion and their properties. However, inferring motion of articulated objects from sound is rather unexplored. Our paper shows that sound can indeed provide useful information for kinematic structure estimation.

B. Sensor Modalities in Kinematic Structure Estimation

Articulated mechanisms are constrained in their motion according to their kinematic structure. This constrained motion becomes apparent in many sensor modalities, such as force feedback, vision, and sound. In the following, we explain the strength and weaknesses of approaches using different combinations of sensor modalities.

Force and proprioceptive sensing can be used when robots directly manipulate articulated objects [10], [11]. But this requires the robot to be physically in contact with the kinematic structure—remote estimation is not possible.

Approaches based on vision, estimate kinematic structures either from a single image [12], [13], [14], or integrate information over time [15], [16]. While visual estimation works well in favorable conditions, it fails in adverse settings, such as in bad lighting conditions or due to visual occlusions.

More robust results can be achieved with multi-modality, i.e., fusing proprioception, force and vision [17] [18] [19]. Such fusion makes estimation robust to adverse lighting conditions [18] or missing physical contact [19]. But such approaches fail when visual and force-torque limitations

coincide, i.e. when a *remote* kinematic structure needs to be estimated under bad visual conditions. In contrast, acoustic sensing works in bad visual conditions and when the robot passively, remotely estimates a kinematic structure.

III. Estimating Motion From Sound

We aim to infer the motion of articulated mechanisms from sound. However, the relationship between sound and motion is complex, and varies between different instances of the same category of objects. Thus, we present a flexible learning approach that can obtain such a mapping from little data, to enable quick adaptation to novel objects.

Note, sound may be used to perceive different kinds of articulated mechanisms. We focus on drawers (prismatic joints) as a representative type of articulated object. Drawers are a common class of mechanisms and feature significant intra-class variance regarding their acoustic properties.

In Section III-A we explain how we estimate the speed of drawers from sound. For a full estimate of velocity, we need to augment this estimate of speed with a measure of direction. In Section III-B we estimate the direction of drawer movement using a microphone array. Section IV-B then explains how we fuse these estimates into an existing vision-based framework.

A. Estimating Speed of Kinematics Actuation

To estimate the speed of prismatic joints, we first divide the input sound signal into segments of length 0.064 seconds. Then we perform Fast Fourier Transform on each sample, only preserving positive frequencies, which yields a vector of length 512. We predict the drawer's speed from this feature vector using Principle Component Regression (PCR). For this, we perform Principal Component Analysis (PCA) and reduce the dimensionality of the feature vector to 50 by preserving only the projection onto the largest 50 principal components. This audio feature vector x_{audio} is the input to a linear regression model that predicts speed as a scalar. We train this model using supervised learning and minimize the least squares error. For training, we measure ground-truth speed of drawers in a motion capture system. The result is an estimator g which can estimate the absolute velocity v_{abs} of a drawer as in

$$v_{\text{abs}} = g(x_{\text{audio}}). \tag{1}$$

B. Estimating the Direction of Kinematics Actuation

We estimate the direction of motion in prismatic joints based on sound-source localization. We use the ODAS software [20], which localizes sound sources as coordinates on a unit sphere $s_t^m = (x_t, y_t, z_t)$ in a microphone array's coordinate system m. Since we will later fuse sound-based estimates with vision-based estimates, we transform the sound-source localization estimates to camera frame c by a known homogeneous transform T_m^c such that $s_t^c = T_m^c s_t^m$. This transformation is theoretically not correct, but it is sufficient in practice, because the microphone array and camera are spatially close, and because we are solely interested in the *relative* position of consecutive sound source localizations. To estimate the direction of motion, we compute the difference between consecutive sound source localization estimates, as in

$$\mathbf{d}_t = \frac{\mathbf{s}_t^c - \mathbf{s}_{t-1}^c}{\|\mathbf{s}_t^c - \mathbf{s}_{t-1}^c\|}. \tag{2}$$

We combine this direction of motion along with the speed estimate (Section III-A) to produce an estimate of velocity. We use this resulting quantity in a kinematics estimation framework described in the next section.

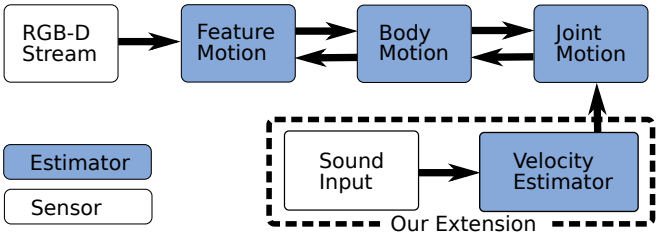

Fig. 2. The existing vision-based framework estimates kinematic joint motion from visual cues. It is based on several interconnected recursive estimators. Our work extends this model with additional velocity measurements derived from audio.

IV. Sound in Kinematic Structure Estimation

The full kinematic structure estimation problem requires to estimate the poses and velocities of rigid bodies, as well as of the kinematic joints that connect them. We do not try to solve this full problem just based on sound. Instead, our work builds on an existing vision-based framework for kinematic structure estimation, extending it by the aforementioned audio-based velocity estimator.

A. Vision-Based Interactive Perception Framework

The existing, vision-based framework [15] we extend is depicted in Figure 2. It consists of three interconnected recursive Bayesian estimators, implemented as Extended Kalman Filters (EKF). One filter tracks 3d feature points based on RGB-D input. It interacts with a second filter, which tracks rigid body hypotheses based on the first filter's feature points. A third filter estimates kinematic joints between these rigid body hypotheses.

Because each estimator implements a Bayes filter, they can interact by using such filters' common update functions – measurement and prediction updates. The belief of the rigid body estimator is used in the prediction step for feature motion, and conversely, it uses the belief over feature motion as a measurement. Likewise, the kinematic joint estimator is used in the prediction step of rigid body motion, and conversely, it uses the belief about rigid body motion as a measurement.

The kinematic joint tracker consists of multiple EKF filters, one filter for each pair of rigid bodies and each possible joint type (revolute, prismatic, rigid and disconnected). The state of a prismatic joint x^{joint} is defined by the joint's

orientation, azimuth and elevation (ϕ, θ), the actuation offset (q), and the velocity of the joint ($\dot{q}$).

$$x^{\text{joint}} = (\phi, \theta, q, \dot{q}) \tag{3}$$

In the *prediction step*, a state prediction is made for each filter based on the current joint-space velocity, performing velocity integration. Then, in the *update step*, the belief over the state is corrected using measurements from the rigid-body tracker. This vision-based measurement z^{vision} is the change in relative pose between the two rigid bodies connected by a joint. In the following section we will present an additional, audio-based measurement for velocity in joint-space for the prismatic joint filter. For further details on the visual part of the architecture we would like to refer the reader to [15].

B. Audio-Vision Fusion

The sound-based estimator we explained in Section III estimates the velocity of a kinematic joint. We add this estimator to the recursive filter framework as an additional measurement on a tracked kinematic joint's velocity.

To incorporate the audio-based velocity estimate, we need to transform the direction vector $\mathbf{d}_t$ from camera-frame to joint space. Thus, we project the audio-based direction estimate $\mathbf{d}_t$ onto the prismatic joint axis estimate of the kinematic joint filter, estimated from vision. We only retain the sign of this projection. Because the sign estimate is noisy, we filter it using a simple Hidden Markov Model (HMM). This HMM assumes the sign remains identical between time-steps with a probability of 0.7 and conversely assumes that it changes with probability 0.3. We multiply the maximum likelihood sign with the speed estimate v_{abs} to produce an estimate for the joint velocity, v.

We then fuse this joint velocity estimate v, with the current estimate of the joint parameters from vision. We use the following as the audio-based measurement model for the joint filter EKF mentioned in the previous section:

$$h^{\text{audio}} = \begin{bmatrix} 0 & 0 & 0 & \dot{q} \end{bmatrix} \tag{4}$$

where the joint velocity $\dot{q}$ from x^{joint} simply maps to v. The measurement noise is gaussian with standard deviation estimated as the regression model's average training error.

Finally, we perform a post-processing step to avoid drift due to velocity integration. As we only use audio to estimate joint velocity, small errors accumulate over time. To address this problem, we constrain the mean of the filter's joint position estimate to lie between the maximum and minimum joint limits that are updated in time-steps when vision-based measurements were available.

V. Experiments

Our analysis comprises two parts. In Section V-A we analyze the regularities that govern how motion in drawers shapes the sound they emit. Then, in Section V-B we evaluate the performance of our proposed approach.

Fig. 3. We use four drawers in our experiments. From left to right and from top to bottom these drawers are referred to as *IKEA small, IKEA large, cabinet top* and *cabinet bottom* throughout this paper. The drawers are of different size and material, and are mechanically mounted in different ways.

The data for the experiments was generated by a human experimenter actuating the different drawers shown in Figure 3. To record sensor input, we used an ASUS Xtion PRO LIVE camera as the RGB-D sensor, two ReSpeaker Mic Array v2.0 microphone arrays[1]. Ground truth data for the evaluation was recorded by tracking the drawers in a motion capturing system. This setup is also depicted in Figure 1.

A. Data Analysis – The Sound of Moving Drawers

We aim to show that audio holds intrinsic information about the motion of articulated objects in general, and drawers in particular. Thus, in this section we analyze how properties of sound relate to the motion of drawers.

As a first step of this analysis, in Figure 5 we visualize the spectogram of sound recorded while the drawer *cabinet top* was being repeatedly opened and closed. In this visualization it becomes apparent that we can distinguish at least three different types of events from sound: the drawer sounds differently when it is static, in motion, and when it hits a joint limit. The latter yields an especially salient event in the spectogram. This qualitative result suggests the velocity of a drawer indeed shapes the sound it emits, but to exploit this insight we need to inquire further *how* a drawer's sound depends on its actuation velocity.

Two of the most important and distinct properties of sound are its loudness and pitch. In Figure 4 we visualize how these properties change with varying speeds of the four drawers. Loudness is measured as the mean amplitude of the sound spectrum and pitch is measured as the most dominant frequency in the sound spectrum at that point in time. We measure reference speed in a motion capturing system. The data shows that both, loudness and pitch *coarsely* increase with higher speeds in each drawer. There are also data-points

[1]For technical reasons these microphone arrays could not record audio and perform sound source localization at the same time.

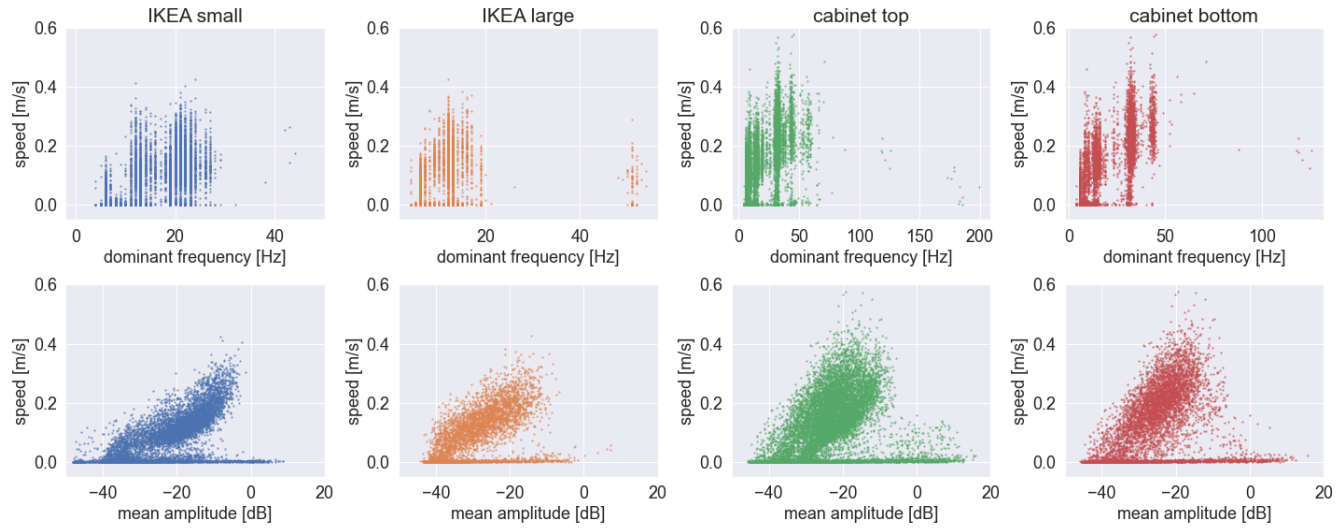

Fig. 4. The pitch (dominant frequency) and loudness (mean amplitude) of sound plotted against the speed of four different drawers while they are being actuated. Data-pairs of mean amplitude plotted against sound seem to follow a bimodal data-distribution. While higher speeds generally lead to louder sound, there are also data-points with high amplitude at zero speed. This is likely because the drawers make a loud sound when they are hitting their joint-limits, yet at that instant they stop moving. For *cabinet top* and *cabinet bottom* pitch also coarsely increases with speed, which for *IKEA large* and *IKEA small* such a relation is less visible.

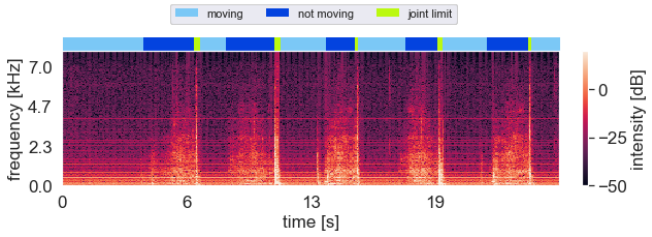

Fig. 5. Sections of the sound spectrogram correspond to distinct phases of an articulated drawer's motion. The spectrogram is extracted from a recording session of the *cabinet top* drawer. Sections with higher intensity, thus higher amplitude, correspond to states where the joint was in motion. In contrast, sections with lower intensities and lower amplitude correspond to no motion in the joint. The short streaks of high intensity correspond to events when the drawer hits its joint limit and emits a loud *bang* noise.

with high amplitude at zero speed. This is likely because the drawers make a loud sound when they are hitting their joint-limits, yet at that instant they also stop moving. Figure 4 also reveals that the regularities that connect sound to velocity are different for each drawer, which makes generalization between drawers challenging.

This analysis shows that sound carries information useful to estimate motion in kinematic structures in general, and particularly motion in drawers. Sound is a rich source of information, and even when we substantially simplify the signal to just pitch and loudness, we still see a strong relation between sound and motion. Next we will evaluate if this information is sufficient to track the state of an articulated drawer over time.

B. Audio-Visual Drawer Motion Estimation

As described in Section IV-B, we track the actuation state of drawers using a system of interconnected Bayes filters, augmented by an audio-based velocity estimator. We first evaluate the estimator's performance in Section V-B.1 and then evaluate the performance of the full recursive estimation system, using that estimator, in Section V-B.2.

1) Instantaneous Velocity Estimation from Audio: The data analysis in the previous section showed that both, loudness and pitch of sound carry information about the speed of a drawer. Now we assess the performance of different regression models that aim to estimate speed from both these features, as well as from a more powerful feature representation obtained by PCA. In Figure 6 we plot the Root Mean Squared Error (RMSE) of different regression models, separately for each of the drawers. We compute RMSE for test-data splits using 7-fold cross-validation, preserving the data's temporal order. The results show that amplitude and frequency are useful features to predict speed from sound, especially in combination. However there is useful information in the sound signal beyond these two features. The PCR model (Section III-A) outperforms the other estimators, including polynomial regression of degree $p = 3$ on the combined feature vector of frequency and amplitude.

One strength of the PCR model is that its simplicity allows us to investigate what it learned. The model first performs PCA for dimensionality reduction, then linear regression on the reduced, 50-dimensional feature vector. Figure 7 plots the 20 largest eigenvalues on a logarithmic scale, which shows that the majority of variation in the data is captured in the first few principal components. In Figure 7 we also plot the eigenvectors to the four largest principal components. This reveals a relevant insight. The eigenvector that corresponds to the largest eigenvalue has an almost constant value for all frequencies. This indicates that it captures the overall loudness of sound, irrespective of frequency. The eigenvector

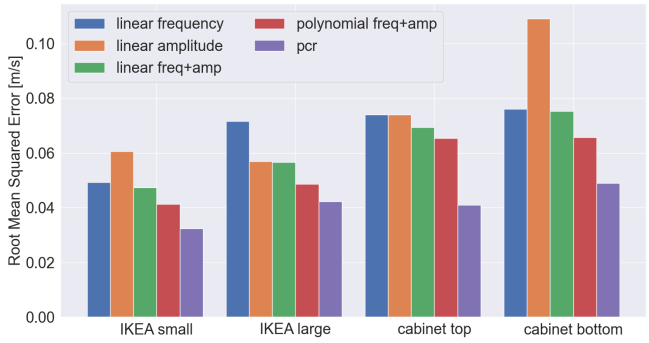

Fig. 6. We tested five different models to estimate drawer speed from sound. Here we analyze their Root Mean Squared Error (RMSE). The PCR model outperforms the other approaches. This indicates that also other features than loudness and pitch of sound are helpful to estimate motion in a drawer.

to the second largest eigenvalue has a largely linear, inclining shape. This indicates that higher frequencies produce larger responses, similar to the relationship of dominant frequency in Figure 4. The PCR model thus operates on similar features as the other models, but further principal components allow for more fine-grained estimation results. We can assess which eigenvectors contribute most strongly to speed estimates on average, by plotting the product of eigenvalues and corresponding linear regression weights (Figure 7, middle). This reveals that only the four first principal components significantly influence estimation results, but the strongest contribution by far comes from the first principal component, which we interpret as *loudness* of the sound.

Along with speed it is also necessary to know the direction of motion. In Section III-B we presented an engineered estimator for this task. Its classification performance is shown in Table V-B.1. The classification rates are not very high, due to the noisy input from sound-source localization. However, our framework that integrates this estimator can still utilize the estimator's noisy predictions. It could still be worthwhile to denoise this input further. Next we evaluate the performance of the full estimation system.

TABLE I
SIGN OF VELOCITY – CLASSIFICATION RATE PER DRAWER

Drawer	Classification rate
IKEA small	58.1 %
IKEA large	56.8 %
cabinet top	59.6 %
cabinet bottom	63.4 %

2) Filtering Drawer Motion With Sound Input: In Section IV-A we explained a vision-based framework [15] to estimate kinematic structures. In this section we compare the performance of that purely vision based approach to our suggested audio-visual approach presented in Section IV-B. Figure 8 shows an example sequence where a drawer is being actuated repeatedly, together with the joint position estimate of both models. The estimation quality is representative of both model's average estimation performance.

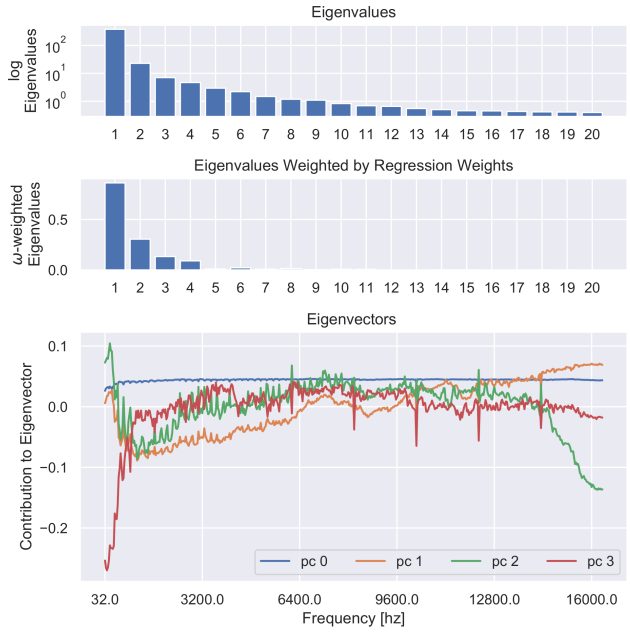

Fig. 7. Top: we plot the 20 largest eigenvalues on a log-scale. This show that the majority of variance is captured by the largest components, but that smaller components also expain some variance. Middle: the eigenvalues multiplied by their linear regression weights show that only variation in the four first principal components contributes significantly to velocity estimation. Bottom: We plot the eigenvectors that correspond to the 4 largest eigenvalues. Plotted in frequency-space, the eigenvectors are jagged on a short time-scale, but follow long range trends. The eigenvector that corresponds to the largest eigenvalue has an almost constant value for all frequencies. This implies that it captures the overall loudness of sound. The eigenvector to the second largest eigenvalue has a largely linear, inclining shape. This indicates that the second-largest degree of variation corresponds to the overall pitch of sound.

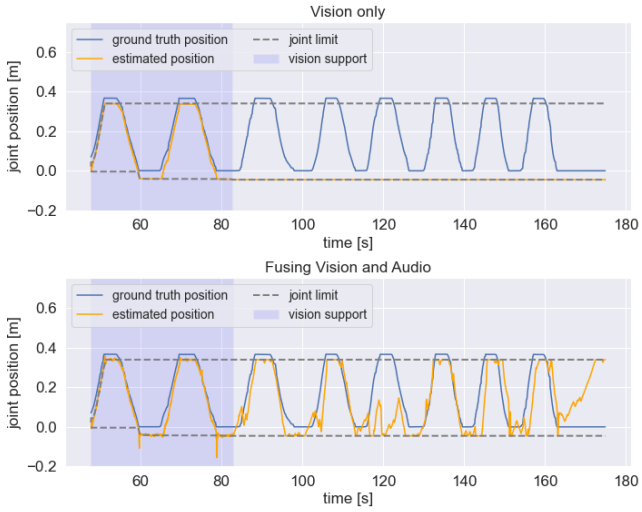

Fig. 8. We plot the belief about the configuration of the prismatic joint in a drawer. The top plot shows a purely visual approach [15] and the bottom plot shows our approach that extends it by fusion with auditory perception. The shaded light blue region indicates an initial phase in the experiment when the light in the experimental compartment was on, but after this phase it was turned off. Purely visual perception breaks down when the light is turned off, while our proposed audio-visual approach is largely able to maintain good estimates.

In the beginning of the experiment there are no occlusions and lighting conditions are ideal. In this initial phase, both approaches can properly estimate the actuation state of the drawer while it is being opened and closed. However after about $t = 40$ seconds we turn off the light in the experimental compartment. Then, the vision-only estimate can no longer perform state updates, since it does not receive measurements of visual features. Contrarily, the audio-visual model continues to maintain estimates that are largely correct, except for a short time-span at around $t = 120s$, when the velocity sign estimates are erroneous. But critically, the audio-visual approach is functional in a condition where the purely visual approach completely breaks down. This shows that multi-modal perception approaches can overcome challenges that are fundamental limitations to single sensor modalities. Multi-modality is then not only a means to reject noise, it opens up categorically new application scenarios.

The results in Figure 8 are repeatable and similar for the 4 different drawers we used in our experiments. In order to obtain numerical results we recorded 7 to 9 sessions per drawer. We computed the RMSE between the mean belief of our approach and the ground truth joint actuation state of the drawers, averaged over all time-steps and all recording sessions. The RMSE is plotted in Figure 9. We show the RMSE for the full time-series (*total*), but also separately for the phase of the task when the light is turned on (*light on*) and when it is turned off (*light off*). For all four drawers we can see the same ordering of estimation performance. Compared to the vision-only approach [15], our audio-visual approach performs better, except for phases of the task where the visual conditions are ideal (*light on*). We suspect that, also in this condition, the audio-visual approach could reach the same performance as the purely visual one if we properly tuned the measurement noise for each of the sensor models.

This analysis confirms our hypothesis that sound is a powerful sensor modality that augments and complements vision-based kinematics estimation. Although the sound-based approach is less precise than the visual one, it can be applied in situations that are challenging for purely vision-based perception. Using sound in kinematics estimation enables perception in categorically different, more challenging conditions. Further, the audio-visual fusion approach is precise and robust, which provides further evidence that multimodality is a key ingredient for robust perception.

C. Limitations

While our work confirms that we can track the state of a drawer using audio, this work is still subject to limitations.

The approach assumes that a single, well trackable sound-source is in the scene. We informally observed that the system is robust against light background noise, like audible footsteps, but a principled analysis of its sensitivity to noise is still pending. Denoising the input could increase robustness.

Related, the system cannot disambiguate sound from different kinematic joints. This would be necessary to solve complex kinematic problems, such as mechanical puzzles [21]. We could enable multi-joint tracking by performing probabilistic data association [22].

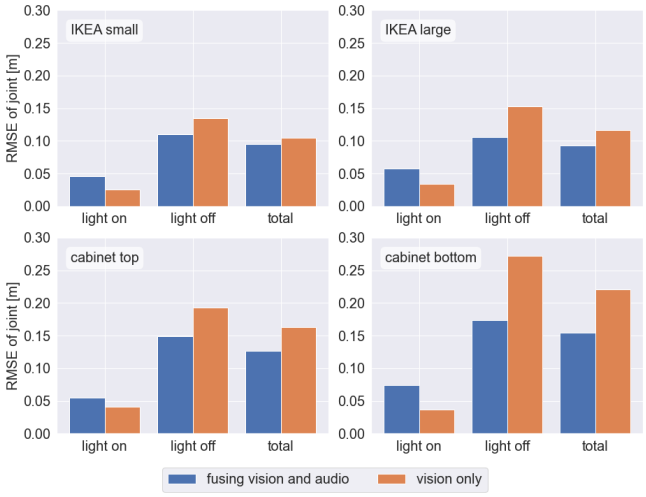

Fig. 9. RMSE of the joint position estimate calculated for all sessions per drawer. By fusing audio with vision measurements (blue), we are able to decrease the error when there is no vision support.

Further, the system currently needs to be bootstrapped by vision. We only track the articulation state of a known prismatic joint that the system could first visually perceive. We expect that it will be difficult to completely alleviate this restriction, but by actively moving the directional microphone we may be able to extract more spatial information.

Another challenge is generalization between different kinematic structures, which we do not address in this work. Assessing and improving the system's generalization capabilities would be an important next step.

VI. Conclusion

We analyzed the regularities that govern how the speed of a drawer shapes its sound and used these regularities in a model that tracks the motion of drawers. Our analysis shows that drawers sound louder and higher in pitch when they are actuated with higher velocity. We trained a regression model to predict motion from sound, and our analysis shows that it is also tuned to use these features. We then fused this sound-based motion estimate into an existing vision-based kinematics estimation framework. The resulting system bolsters the vision-based estimation framework with increased robustness to conditions with deteriorated visual inputs. This shows that sound is a promising sensor modality that can complement visual estimation of motion in articulated objects. Such complementarity of sensor modalities is a key ingredient to make robotic perception systems robust. It will be interesting to explore the limits of sound-based kinematics estimation in future work.

References

[1] D. Gandhi, A. Gupta, and L. Pinto, "Swoosh! rattle! thump! - actions that sound," in *Robotics: Science and Systems XVI*, Robotics: Science and Systems Foundation, 7 2020.

[2] J. Sinapov and A. Stoytchev, "Grounded object individuation by a humanoid robot," in *2013 IEEE International Conference on Robotics and Automation*, pp. 4981–4988, 2013.

[3] J. Sinapov, M. Wiemer, and A. Stoytchev, "Interactive learning of the acoustic properties of household objects," in *2009 IEEE International Conference on Robotics and Automation*, pp. 2518–2524, 2009.

[4] J. Sinapov and A. Stoytchev, "Object category recognition by a humanoid robot using behavior-grounded relational learning," in *2011 IEEE International Conference on Robotics and Automation*, pp. 184–190, 5 2011.

[5] E. Torres-Jara, L. Natale, and P. Fitzpatrick, "Tapping into touch," in *Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, vol. 123, pp. 79–86, Lund University Cognitive Studies, 2005.

[6] G. Zöller, V. Wall, and O. Brock, "Acoustic Sensing for Soft Pneumatic Actuators," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Madrid, Spain), pp. 6986–6991, Oct. 2018.

[7] Z. Zhang, Q. Li, Z. Huang, J. Wu, J. Tenenbaum, and B. Freeman, "Shape and material from sound," in *Advances in Neural Information Processing Systems*, vol. 30, p. 11, Curran Associates, Inc., 2017.

[8] J. Sinapov, C. Schenck, and A. Stoytchev, "Learning relational object categories using behavioral exploration and multimodal perception," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5691–5698, 2014.

[9] S. Djukanović, J. Matas, and T. Virtanen, "Acoustic vehicle speed estimation from single sensor measurements," *IEEE Sensors Journal*, vol. 21, no. 20, pp. 23317–23324, 2021.

[10] Y. Karayiannidis, C. Smith, F. E. Vina, P. Ögren, and D. Kragic, "Model-free robot manipulation of doors and drawers by means of fixed-grasps," in *2013 IEEE International Conference on Robotics and Automation*, pp. 4485–4492, 2013.

[11] A. Jain and C. C. Kemp, "Pulling open doors and drawers: Coordinating an omni-directional base and a compliant arm with equilibrium point control," in *2010 IEEE International Conference on Robotics and Automation*, pp. 1807–1814, 2010.

[12] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3703–3712, 2020.

[13] X. Wang, B. Zhou, Y. Shi, X. Chen, Q. Zhao, and K. Xu, "Shape2motion: Joint analysis of motion parts and attributes from 3d shapes," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8868–8876, 2019.

[14] H. Abdul-Rashid, M. Freeman, B. Abbatematteo, G. Konidaris, and D. Ritchie, "Learning to infer kinematic hierarchies for novel object instances," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8461–8467, 2022.

[15] R. Martín-Martín and O. Brock, "Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2494–2501, 2014.

[16] J. Sturm, C. Stachniss, and W. Burgard, "A Probabilistic Framework for Learning Kinematic Models of Articulated Objects," *Journal of Artificial Intelligence Research*, vol. 41, pp. 477–526, Aug. 2011.

[17] K. Hausman, S. Niekum, S. Osentoski, and G. S. Sukhatme, "Active articulation model estimation through interactive perception," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3305–3312, 2015.

[18] R. Martín-Martín and O. Brock, "Cross-modal interpretation of multimodal sensor streams in interactive perception based on coupled recursion," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3289–3295, 2017.

[19] M. A. Lee, B. Yi, R. Martín-Martín, S. Savarese, and J. Bohg, "Multimodal sensor fusion with differentiable filters," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10444–10451, 2020.

[20] F. Grondin, D. Létourneau, C. Godin, J.-S. Lauzon, J. Vincent, S. Michaud, S. Faucher, and F. Michaud, "ODAS: Open embedded audition system," *Frontiers in Robotics and AI*, p. 125, 2019.

[21] M. Baum, M. Bernstein, R. Martín-Martín, S. Höfer, J. Kulick, M. Toussaint, A. Kacelnik, and O. Brock, "Opening a lockbox through physical exploration," in *Proceedings of the IEEE International Conference on Humanoid Robots (Humanoids)*, 2017.

[22] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems Magazine*, vol. 29, no. 6, pp. 82–100, 2009.