2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2025)
October 19-25, 2025. Hangzhou, China

2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3315-4393-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/IROS60139.2025.11245966

TBAP: Tapping-Based Auditory Perception for Identifying Container Materials

Zehao Li, Shoujie Li, Hao Guo and Wenbo Ding

***Abstract*— In this study, in order to address the robotic auditory perception problem, we propose a novel framework for object material recognition of common containers, which combines deep learning with active auditory perception to achieve breakthrough results. We developed a modular robotic system for acoustic data acquisition that employs a hybrid mechanism of vertical translation and horizontal rotation that is capable of performing full-scale tapping in three dimensions. The system is capable of creating an acoustic dataset consisting of 50 containers made of five materials, which improves the data acquisition efficiency by 93.9% compared to manual operations. In addition, we propose an end-to-end transfer learning model, TBAP, which is trained on a crawler-generated pre-training dataset and 50 real scene samples, and achieves a recognition accuracy of 91.0% for unseen materials. To improve reliability, we design a dynamic confidence assessment mechanism that generates confidence indices through probability distribution analysis and feature stability assessment to support robust robot decision-making. Experimental results show that the framework greatly improves data acquisition efficiency while maintaining high recognition accuracy, providing a valuable tool for advancing acoustic perception research.**

I. INTRODUCTION

II. INTRODUCTION

The advancement of auditory perception technology enables deep interaction between intelligent agents and physical environments in embodied robotic perception research [1]. Unlike conventional visual perception, which struggles with environmental adaptability and interference [2], active auditory perception offers advantages such as non-contact measurement, visual interference resistance, and real-time responsiveness, making it ideal for object property recognition in complex scenarios [3]. Among various acoustic sensing techniques, impact acoustic analysis [4] stands out for its operational efficiency and cost-effectiveness, particularly in applications like industrial quality inspection and service robotics [5].

The design of acoustic data acquisition systems significantly affects recognition performance. Current systems primarily employ passive methods or robotic arm-based active systems [6], which are inefficient and lack scalability for large-scale datasets [7]. Additionally, existing datasets have limited material category coverage [8], restricting practical applications of recognition models. Containers, widely applicable and structurally suitable for auditory perception, provide an excellent sample type due to their ability to generate sound waves with simple contact.

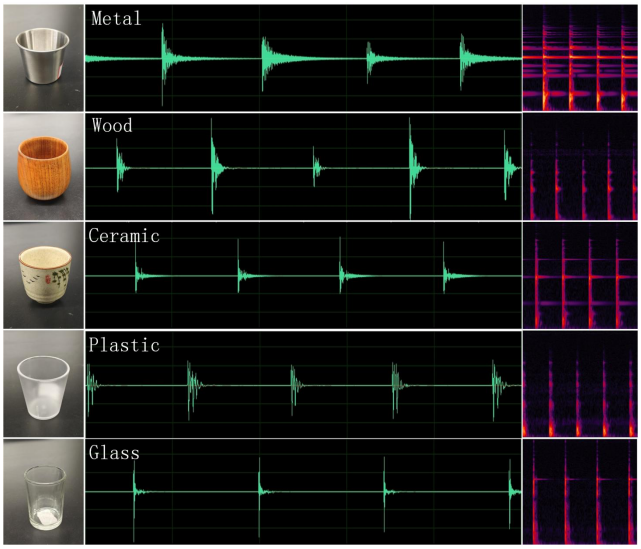

Fig. 1. Schematic diagrams of audio and frequency spectra of different materials.

To address these challenges, this paper introduces a novel data sampling system and recognition framework. The main contributions include:

- A motion structure combining a linear motor and rotating chassis for high-resolution data collection. A sliding mechanism with an electromagnet enhances object surface impact efficiency, improving manual data collection by up to 93.9%.
- An end-to-end material recognition algorithm utilizing a pretraining-finetuning transfer learning approach. An ASMR dataset and a curated dataset of 50 real-world objects support model training and evaluation, with a confidence index quantifying prediction reliability.
- Comprehensive experiments validate the system's advantages in sampling rate and recognition accuracy, with extensibility to industrial and household applications.

The paper is structured as follows: Section 2 reviews related work, Section 3 introduces the data acquisition sys-

*This work was supported in part by National Key R&D Program of China (No. 2024YFB3816000), Shenzhen Key Laboratory of Ubiquitous Data Enabling (No. ZDSYS20220527171406015), Guangdong Innovative and Entrepreneurial Research Team Program (2021ZT09L197), and Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (No. SZPR2023005).(Corresponding author: Shoujie Li and Hao Guo, sa170090@mail.ustc.edu.cn)

Z.Li, S.Li, H.Guo and W.Ding is with the Shenzhen International Graduate School, Tsinghua University, Shenzhen, China .(Email:lizehao22@tsinghua.edu.com)

This paper has supplementary downloadable material at:https://github.com/QuQuBSM/TBAP-Tapping-Based-Auditory-Perception-

tem, Section 4 details the TBAP model and confidence loss function, Section 5 presents experimental results, and Section 6 concludes with key findings and discussions.

III. RELATED WORKS

In recent years, deep learning methods [9] have emerged as revolutionary tools for acoustic feature extraction [10], with their core advantage lying in the ability to automatically learn acoustic patterns in a data-driven manner. By constructing end-to-end neural network models, an intelligent mapping can be achieved from raw acoustic signals to material classification, enabling key tasks such as impact localization [11], acoustic feature extraction [12], and material discrimination [13]. Existing studies indicate that convolutional neural network (CNN)-based acoustic classification models [14] have achieved recognition accuracies exceeding 80% in laboratory settings. This paper will next provide a brief review of related work in the areas of material classification and auditory perception.

A. Machine Learning for Material Classification

Object material recognition plays a crucial role in robotic embodied perception and holds great promise for its future development.For example, Li et al. [15] proposed a hybrid tactile sensor that combines triboelectricity and electromagnetic induction. It can identify different objects with high precision and enhances the object recognition ability of robots in complex environments. Zhu et al. [16] designed a soft modular glove based on the multifunctionality of materials, which features multi-modal perception and enhanced tactile feedback. It achieves two-way multi-modal communication through a simple design.

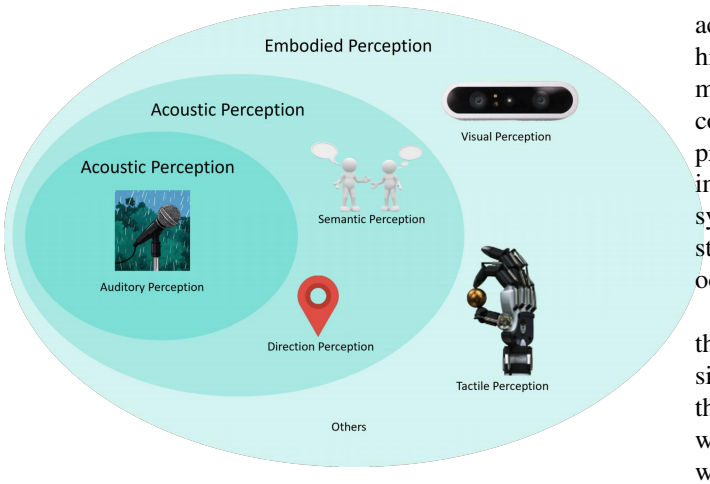

Fig. 2. Categorization of sound perception in embodied perception.

B. Auditory Perception

Experiments have shown that sound possesses inherent intuitiveness and accuracy [17]. Hearing is one of the most important information modalities for humans, enabling tasks such as localization [18], [19], communication [20], and even movement [21]. Inspired by this, deep neural networks for sound perception were used to solve problems including natural sound recognition [22], service robot [23] and flexible sensors [24]. Thus, deep learning has strong potential in the field of auditory perception.

In summary, there are two major challenges : First, auditory features are highly susceptible to environmental noise, leading to a significant decline in the generalization performance of the models [25]. Thus, it is crucial to stably and efficiently acquire high-quality datasets. Second, traditional machine learning architectures struggle to provide reliable discrimination, limiting their ability to distinguish complex materials. This necessitates the design of unique end-to-end methods to address this issue [26].

IV. AUDIO DATA ACQUISITION SYSTEM

To accomplish the task of object material recognition based on deep learning and auditory active perception, an automated data collection system must meet the following key requirements: high precision, efficiency and speed, user-friendliness, and multi-dimensional capabilities. To address these challenges, this work introduces a unique data collection framework designed to enable the rapid acquisition of impact audio data from common containers in everyday scenarios. The proposed system is capable of quickly and reliably collecting audio data generated by different objects under tapping actions, ensuring both the diversity and consistency of the collected data. This system provides sufficient and high-quality training data for deep learning models, while simultaneously improving data collection efficiency and reducing labor costs.

A. Hardware Design

The system adopts a composite motion architecture characterized by ”vertical translation and horizontal rotation.” A high-precision linear motor drives the solenoid valve striking mechanism along the Z-axis, while a 360° rotating turntable, controlled by a servo motor at the base, facilitates comprehensive three-dimensional spatial coverage. The system integrates acoustic sensors, motion control modules, and data synchronization units, enabling automated path planning for striking operations in scenarios with complex geometries and occlusions.

As is shown in Fig.3, for the linear motor, we employed the DKC-Y110 motor from KHMOS, which features a significant advantage: The motor can be precisely controlled through digital control. Regarding solenoid valve selection, we opted for the commonly used KK-0520B solenoid valve, which operates at a rated voltage of 24V. This choice ensures compatibility with the linear motor, simplifying power delivery. Additionally, the solenoid valve control is automated to toggle between high and low levels using an ESP32 development board, which greatly facilitates the realization of an intelligent data acquisition system.This allows for multi-angle, locatable sound field signal acquisition. The linear motor is dynamically coordinated with a programmable stepper motor to precisely control speed and travel range, while the audio acquisition system maintains spatiotemporal

data consistency by synchronizing the striking mechanism's timing with the audio sampling process.

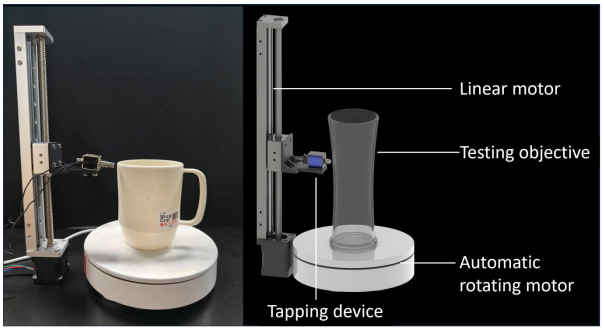

Fig. 3. In our design, the linear motor primarily facilitates the automated vertical movement of the solenoid valve, while the rotating disk at the base ensures that all sides of the object have an equal opportunity to be impacted.

B. Design of Tapping Device

The design of the striking device is a critical component for interacting with target objects, and its performance directly affects the validity of the collected audio data. We first need to demonstrate the effectiveness of this device. The device is primarily driven by a solenoid valve controlled by a linear motor. Its structure consists of an electromagnet with a metal striking head attached to the front. On one hand, it strives to ensure that the material and dimensions closely resemble those of a robot's finger. On the other hand, when the metal striking head collides with containers of various qualities, the resulting sound is relatively clear, which aids in determining the characteristics of the object. The motor is positioned at an appropriate striking distance (typically 3 to 5 cm) to ensure that the sounds generated by striking the object can be clearly collected, while avoiding interference from mechanical vibrations triggered by being too close.

The effectiveness of the proposed method can be demonstrated by the following evidence. Each impact i is constructed as a smooth Gaussian approximation of a Hertzian half-sine impulse, where the Gaussian profile preserves the essential characteristics of elastic impact dynamics while ensuring differentiability. The approximation is parameterized by impact time t_i and time constant τ_i, where τ_i physically corresponds to the characteristic duration of the Hertzian contact.

$$C_i(t; t_i, \tau_i) = \exp\left(-\frac{6}{\tau_i^2}\left(t - t_i - \frac{\tau_i}{2}\right)^2\right) \tag{1}$$

The coefficient 6 ensures the Gaussian's full width at half maximum (FWHM) matches the half-sine duration τ_i, maintaining energy equivalence with the classical impact model. This parameterization allows the Gaussian to capture both the temporal localization (t_i) and impact intensity ($1/\tau_i$) observed in physical impacts.

Multiple impacts with different parameters are scaled by individual magnitudes in vector $\mathbf{m}$ and summed to compose a force profile $\mathbf{F}$.

$$\mathbf{F}(t; \mathbf{m}, \mathbf{t}, \boldsymbol{\tau}) = \mathbf{m}^T \mathbf{C}(t; \mathbf{t}, \boldsymbol{\tau}) \tag{2}$$

By observing the tapping point, it is evident that the electromagnet used for striking exhibits excellent output force stability across numerous repeated experiments, owing to its constant power supply. This stability significantly ensures the quality of the collected data and enhances the reliability of the sampled data. Furthermore, the electromagnet can be mounted on various mechanical structures, highlighting the versatility and potential of our approach for a wide range of application scenarios.

V. TAPPING-BASED AUDITORY PERCEPTION MODEL FRAMEWORK

For deep learning (DL) models, achieving efficient and fast classification is undoubtedly crucial. However, many existing works face two major challenges: insufficient data [27] and high cost of reliable estimation [7].

Specifically, collecting highly reliable and reconstructable acoustic information corresponding to tapping materials and timing is undoubtedly time-consuming and costly [28]. Moreover, the domain gap between high-quality tapping audio data and the data generated from tapping in real-world robotic interaction scenarios can lead to prediction instability. In severe cases, this may result in significant performance degradation or even errors.

To address these issues, we propose a novel framework that comprehensively considers real-world factors and deep learning methodologies. At the same time, we introduce a Confidence Index, a quantitative metric designed to evaluate the reliability of deep learning model predictions in object material recognition tasks.

A. Data Preprocessing

CNN-based approaches are prevalent in acoustic tasks, and our work builds on an improved CNN framework for inferring material properties. This end-to-end framework includes data preprocessing, model pre-training, parameter shifting, fine-tuning, and resultant output, relying exclusively on multi-layer CNNs for prediction. In addition, we employ a pre-training approach to address the limitations of the dataset, which enhances the generalization ability of the model.

For all data, we performed preprocessing. Our preprocessing tools mainly include two methods, STFT as well as MFCC. The raw audio waveforms were converted to spectrograms using the Short Time Fourier Transform (STFT) to achieve compatibility with CNN. For the vibration data, we extracted features using Mel Frequency Cepstral Coefficients (MFCC), a method that captures frequencies associated with material properties. Through this part of the work we achieve an improvement in the model perception capability.

B. Pre-Training Method

To overcome the limitations of dataset size and the diversity of tapping interaction sounds in real-world scenarios, we autonomously acquired an ASMR dataset via web crawling. Autonomous Sensory Meridian Response (ASMR) refers to a pleasurable tingling sensation triggered by sensory stimuli,

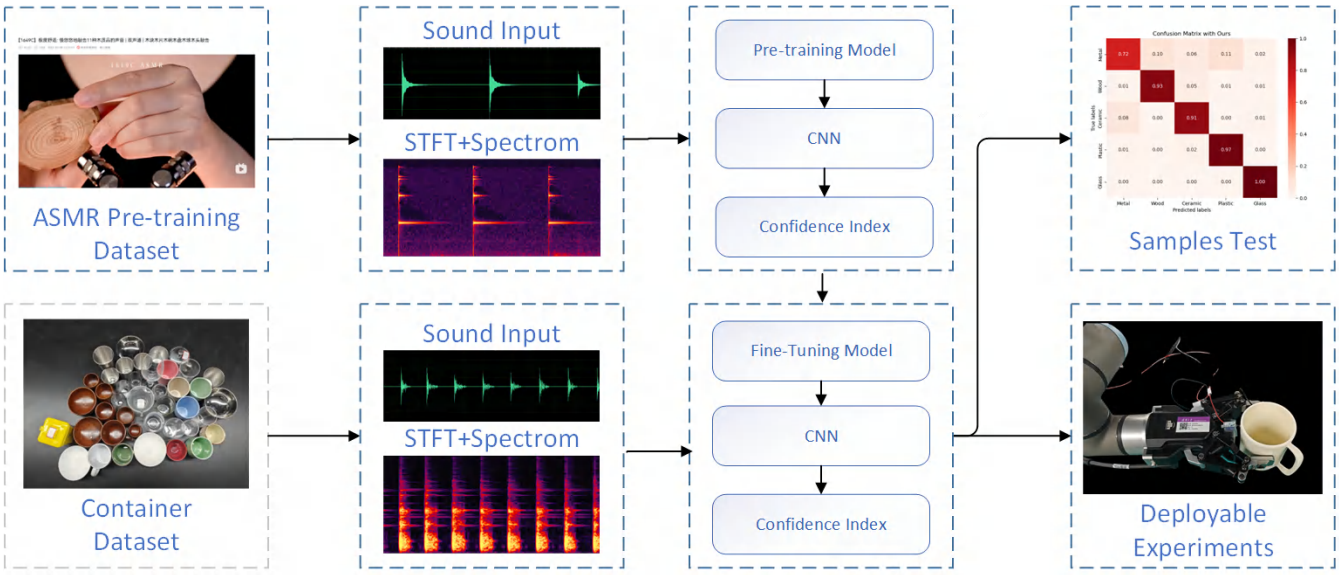

Fig. 4. Framework of TBAP Model: We employed a similar network architecture during the pre-training phase; however, only a subset of the parameters was transferred to the subsequent model.

TABLE I
PARAMETERS USED FOR PRE-TRAINED MODULE AND FINE-TUNING MODULE

Period	Parameter	Value
Pre-trained	Sample Size	62769
	Learning Rate	0.0001
	Training Epoch	100000
	Batch Size	26
	Weight Decay	0.0001
Fine-tuning	Sample Size	3050
	Learning Rate	0.00001
	Training Epoch	100000
	Batch Size	26
	Weight Decay	0.00001

such as auditory and tactile inputs, and this study focused on auditory interactions with materials like wood, plastic, glass, metal, and ceramics.

The dataset was manually screened to ensure consistency with real-world sampling conditions. After preprocessing, 6.16 GB of audio data was segmented into 62,769 labeled samples, enhancing model generalization due to its real-world origin. Pre-training utilized a mean square error (MSE) loss function to minimize the variance of cosine similarity predictions, with the Adam optimizer ensuring efficient convergence and stable learning.

Most model parameters were retained post pre-training, transferring essential feature representations to improve generalization and reduce overfitting during fine-tuning. Hyperparameters were carefully adjusted to address domain differences between pre-training and task-specific datasets, achieving a balance between robustness and sensitivity for optimal downstream performance.

C. Confidence Index Design

The confidence index (conf) is a quantitative indicator used to measure the reliability of the prediction results of deep learning models in object material recognition. In embodied intelligence systems, robots execute corresponding tasks based on the prediction results of the models. For example, in smart home scenarios, they classify and sort items, and in industrial production, they sort the materials of components. To ensure practical applicability, the confidence index is explicitly constrained within the range $[0, 1]$, aligning with its role as a probability-like measure. The accuracy and interpretability of the predictions are of crucial importance.

The design of this loss function innovatively integrates the confidence index (conf) with classification predictions, forming a dual-path optimization mechanism. The core formula of the total loss is:

$$L_{\text{total}} = \alpha \cdot \frac{1}{n} \sum \left((\text{conf} \cdot \hat{y} - y)^2 \right) + L_{\text{samples}},$$

where the predicted probability $\hat{y} \in (0, 1)$ is produced by the model through the Sigmoid function, and the confidence index $\text{conf} \in [0, 1]$ is constructed as a learnable parameter dynamically associated with each sample. The derivation of this formulation stems from the need to disentangle prediction reliability (via conf) from the raw classification output ($\hat{y}$), thereby enabling a more interpretable and adaptive loss function design.

The novelty of this design lies in the following aspects:

(1) The product term $\text{conf} \cdot \hat{y}$ introduces a confidence calibration mechanism. When the model prediction is accurate, minimizing the mean squared error term forces conf towards $y/\hat{y}$. Since the Sigmoid output $\hat{y}$ represents the model's estimated probability of the positive class, this mechanism drives conf to approach 1 for correct predictions ($\hat{y} > 0.5$ and $y = 1$, or $\hat{y} < 0.5$ and $y = 0$), while suppressing conf for

incorrect predictions. This dynamic ensures that conf reflects the model's confidence in its predictions.

(2) The separation of the confidence pathway from the sample pathway enables a dual-feedback loop, regulated by the α coefficient. This dual-pathway design reduces the pressure of loss gradients for incorrect predictions by allowing lower conf values. Additionally, it enhances the model's self-assessment capability, as conf evolves into a meaningful, interpretable dynamic indicator rather than a static artifact. The primary purpose of this dual-path loss design is to improve both the model's prediction accuracy and its ability to quantify uncertainty, leading to more robust and reliable decision-making in real-world applications.

VI. EXPERIMENTS

In this section, we present comprehensive experiments to evaluate the feasibility and superiority of our proposed system. The section begins with a detailed description of the experimental setup and parameter configurations. Subsequently, we discuss studies on the loss function, feasibility evaluation, and comparisons with state-of-the-art (SOTA) methods, all aimed at validating the effectiveness of our approach.

Fig. 5. Real objects used in our dataset. Objects contain the following materials: wood, ceramics, glass, plastic, steel.

A. Implementation Details and Evaluation Metrics

First, the test samples used in this study are introduced. For sample selection, 50 commonly encountered containers of varying sizes and shapes were chosen, making this dataset the largest of its kind in terms of similar datasets and reflecting the focus of this work on everyday scenarios. The samples include cups, bowls, and various challenging containers, with each object differing in size. These 50 containers cover five material categories: metal, wood, ceramic, plastic, and glass. The selection of these five categories was informed by previous studies such as Sonicsense [14] and other related works in this field.The other parameters of the data acquisition system are detailed in Table II. A key feature of this method is its low demand for computational resources. The model is implemented using PyTorch on

TABLE II
PARAMETERS USED FOR AUDIO DATA ACQUISITION SYSTEM

Attribute	Parameter	Value
Data Acquisition System	Frequency	44100Hz
	Sound Channel	Mono
	Tapping Gap	0.2s
	Horizontal rotation speed	$\frac{1}{30}Ht$
	Vertical motion speed	$6^\circ/s$

a 12th-generation Intel Core i5-12500 server with a base clock speed of 3 GHz. Experiments are conducted on a dataset comprising both pretraining data and data collected using the proposed hardware system. Each dataset instance lasts for 1 minute with a sampling frequency of 5 Hz, corresponding to 5 taps per second. Temporal information is associated with the spatial distribution along the longitudinal axis of the container. Preprocessed graphical images are stored in a dataset labeled as Pre-Train and are split into training, validation, and test sets in a ratio of 8:1:1. The network architecture is designed in an end-to-end manner and optimized using the Adam optimizer. To quantitatively evaluate the reconstruction performance, several standard evaluation metrics are employed, including Peak Signal-to-Noise Ratio (PSNR), accuracy, and F1 score.

B. Performance Analysis

To demonstrate the effectiveness of training, the model was continuously evaluated on the validation set every 1,000 time steps during training.

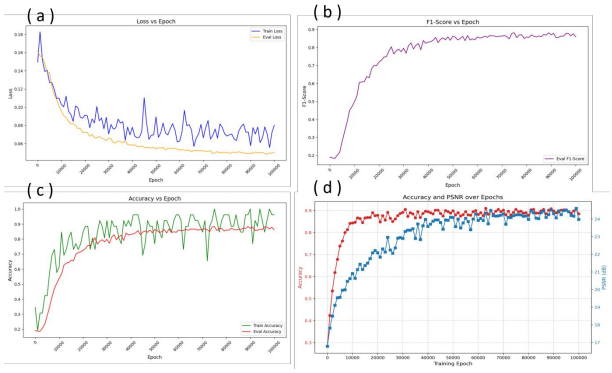

Fig. 6. Results of our model training.(a) Loss (b) F1-score (c) Acc (d) PSNR.

As shown in Fig. 6, both training loss (Train Loss) and evaluation loss (Eval Loss) decrease and converge, with no significant overfitting observed. Training accuracy (Train Accuracy) and evaluation accuracy (Eval Accuracy) stabilize at high levels, indicating reliable performance on validation data. The evaluation F1-Score (Eval F1-Score) improves rapidly and stabilizes around 0.9, demonstrating excellent predictive accuracy and balance. While PSNR shows slight fluctuations in later stages, both classification and reconstruction performance have reached high levels.

The curves indicate that the training accuracy increased rapidly during the initial phase and gradually converged to a

high, stable value. This suggests that as training progresses, the model becomes increasingly robust to noise interference, fully demonstrating the effectiveness of the training process, specifically its ability to perceive audio in complex environments.

C. Comparison with SOTA

For better illustration, a practical comparison was made using SOTA methods including Sonicsense [14] and Realimpact [27]. All of the above methods are primarily based on tapping audio to recognize objects, and are highly valuable for reacting to the state-of-the-art of our research methods. Referring to the above confusion matrix comparison Fig.7, it can be observed from the figures that our model achieves significant improvement in addressing the classification problem of wooden products, and its overall performance surpasses that of the industry-leading models. Meanwhile TableIII below demonstrates that our method performs well on various metrics and has significant positive value for active acoustic perception.

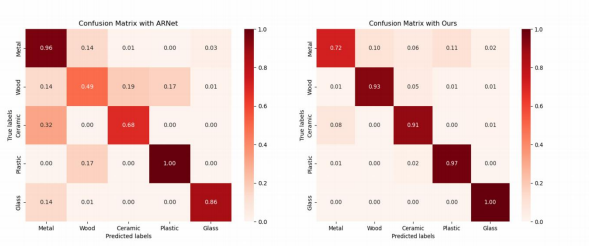

Fig. 7. Confusion matrix of classification results: (A) Our model (B) SOTA model [14].

TABLE III
COMPARISON AMONG CUSTOMIZED SOTA METHODS AND OUR PROPOSED METHOD

Method	Accuracy	Discrimination Times	F1-score
Random	20%	1	0.1
Realimpact [27]	75.1%	1	0.60
Sonicsense [14]	79.8%	24	0.52
Ours	**91.0%**	**1**	**0.85**

D. Ablation Experiment

In this section, we validate the effectiveness of pre-training design through ablation experiments.

Firstly, as shown in Fig 8, pre-training demonstrates the following two main effects: 1. After repeated experimental validation, we conclude that pre-training contributes more than 20% to improving model accuracy; 2. Pre-training can effectively overcome the overfitting problem of the model.

E. Confidence Index and Sampling Rate Analysis

In order to verify that our hardware system has an excellent sampling rate, we compared the difference between the time required for a human to manually take a total of 300 sample points per container and our system. The

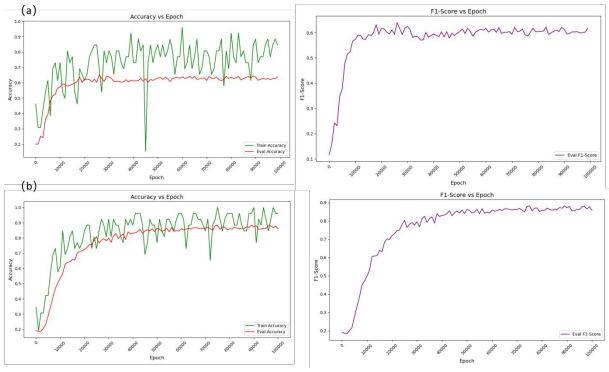

Fig. 8. Experiment Results with out Pre-traing(a) and with Pre-training(b)

comparison results are as follows through 50 repetitions of the experiment. Our system achieves about 93.9% rate improvement (comparing with manual operation), which significantly improves the data collection efficiency.

Meanwhile, our method ensures the effectiveness of the confidence index playing a role in the loss function by setting the weight of the confidence index to 0.2. And by restoring the data acquisition process, our confidence index has an important positive significance in evaluating the quality of a single knock.

F. Deployable Experiments

To further evaluate the effectiveness of our approach in real-world scenarios, we designed an experiment where a mechanical gripper was used to hold the container while a solenoid valve, fixed to the gripper, performed the knocking action, as illustrated in Fig.8. Experimental results demonstrate that for all 50 samples, the proposed interaction method achieves a stable accuracy rate exceeding 85% in single-instance evaluations, highlighting its practicality and reliability. By collecting acoustic signals from different positions, the model is able to accurately identify the material of the object. However, due to the uneven distribution of acoustic signals from various positions within the dataset, the confidence index may vary accordingly.

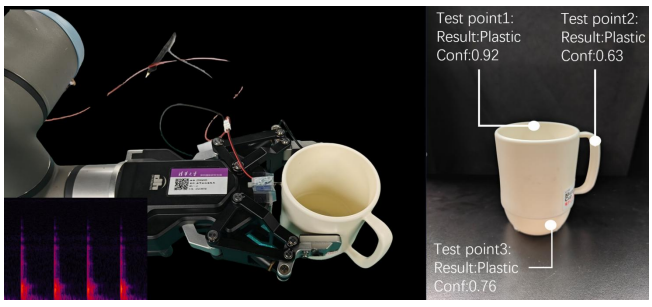

Fig. 9. This process involves active acoustic perception and material identification in a real robotic work scenario, where a robotic arm controls a mechanical gripper.

VII. CONCLUSIONS

This paper proposes an innovative and practically valuable method for object material recognition based on deep learning and auditory active perception. The research has constructed an efficient automatic data acquisition system, which, through the optimization of both hardware and software, has collected a large-scale, high-quality audio dataset, effectively addressing the issues of low efficiency and inconsistent data quality inherent in traditional methods. The CNN-based pre-trained model is designed rationally, enabling effective capture of audio signal features and significantly enhancing the capability for material information extraction and learning. The optimized training process has resulted in a model that surpasses traditional methods in both accuracy and stability. Additionally, the research innovatively constructs a confidence index that comprehensively evaluates prediction reliability from multiple factors, providing support for robotic decision-making. Finally, the usability and robustness of the framework have been validated through experiments. In the future, we plan to further explore the principles and applications of active acoustic perception, optimize model performance, enhance the system's adaptability in complex and dynamic scenarios, and attempt to integrate this method with other sensing technologies to promote the diversification and flexibility of intelligent systems in practical applications.

REFERENCES

[1] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[2] M. Du, O. Y. Lee, S. Nair, and C. Finn, "Play it by ear: Learning skills amidst occlusion through audio-visual imitation learning," *arXiv preprint arXiv:2205.14850*, 2022.

[3] C. Chen, S. Majumder, A.-H. Ziad, R. Gao, S. Kumar Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," in *ICLR*, 2021.

[4] D. Gandhi, A. Gupta, and L. Pinto, "Swoosh! rattle! thump!–actions that sound," *arXiv preprint arXiv:2007.01851*, 2020.

[5] H. Liang, S. Li, X. Ma, N. Hendrich, T. Gerkmann, F. Sun, and J. Zhang, "Making sense of audio vibration for liquid height estimation in robotic pouring," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 5333–5339.

[6] C. Schenck, J. Sinapov, D. Johnston, and A. Stoytchev, "Which object fits best? solving matrix completion tasks with a humanoid robot," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 3, pp. 226–240, 2014.

[7] C. Heggan, S. Budgett, T. Hospedales, and M. Yaghoobi, "Metaaudio: A few-shot audio classification benchmark," in *International Conference on Artificial Neural Networks*. Springer, 2022, pp. 219–230.

[8] V. Dean, S. Tulsiani, and A. Gupta, "See, hear, explore: Curiosity via audio-visual association," *Advances in neural information processing systems*, vol. 33, pp. 14 961–14 972, 2020.

[9] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4316–4336, 2020.

[10] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1070–1077.

[11] Q. Li and Q. Lu, "Impact localization and identification under a constrained optimization scheme," *Journal of Sound and Vibration*, vol. 366, pp. 133–148, 2016.

[12] D. Bonet-Solà and R. M. Alsina-Pagès, "A comparative survey of feature extraction and machine learning methods in diverse acoustic environments," *Sensors*, vol. 21, no. 4, p. 1274, 2021.

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

[14] J. Liu and B. Chen, "Sonicsense: Object perception from in-hand acoustic vibration," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: https://openreview.net/forum?id=CpXiqz6qf4

[15] N. Li, Z. Yin, W. Zhang, C. Xing, T. Peng, B. Meng, J. Yang, and Z. Peng, "A triboelectric-inductive hybrid tactile sensor for highly accurate object recognition," *Nano Energy*, vol. 96, p. 107063, 2022.

[16] M. Zhu, Z. Sun, and C. Lee, "Soft modular glove with multimodal sensing and augmented haptic feedback enabled by materials' multifunctionalities," *ACS nano*, vol. 16, no. 9, pp. 14 097–14 110, 2022.

[17] R. Wang, C. Jung, and Y. Kim, "Seeing through sounds: Mapping auditory dimensions to data and charts for people with visual impairments," vol. 41, pp. 71–83, 2022.

[18] Z. Chen, S. Qian, and A. Owens, "Sound localization from motion: Jointly learning sound direction and camera rotation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7897–7908.

[19] C. Dodsworth, L. J. Norman, and L. Thaler, "Navigation and perception of spatial layout in virtual echo-acoustic space," vol. 197, pp. 104 185–104 185, 2020.

[20] M. Magee, C. Lewis, G. Noffs, H. Reece, J. C. S. Chan, C. J. Zaga, C. Paynter, O. Birchall, S. R. Azocar, A. Ediriweera, K. Kenyon, M. W. Caverle, B. G. Schultz, and A. P. Vogel, "Effects of face masks on acoustic analysis and speech perception: Implications for peri-pandemic protocols," vol. 148, pp. 3562–3568, 2020.

[21] D. ASHMEAD, D. DAVIS, and A. NORTHINGTON, "Contribution of listeners' approaching motion to auditory distance perception." vol. 105, pp. 108 080–108 080, 2023.

[22] Q. Shi, J. Fan, Z. Wang, and Z. Zhang, "Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain," vol. 130, pp. 108 837–108 837, 2022.

[23] P. Agarwal, S. Swami, and S. K. Malhotra, "Artificial intelligence adoption in the post covid-19 new-normal and role of smart technologies in transforming business: a review," vol. 15, pp. 506–529, 2022.

[24] Y. Gao, C. Yan, H. Huang, T. Yang, G. Tian, D. Xiong, N. Chen, X. Chu, S. Zhong, W. Deng, Y. Fang, and W. Yang, "Microchannel-confined mxene based flexible piezoresistive multifunctional microforce sensor," vol. 30, 2020.

[25] B. Shi, M. Sun, K. C. Puvvada, C.-C. Kao, S. Matsoukas, and C. Wang, "Few-shot acoustic event detection via meta learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 76–80.

[26] A. Parnami and M. Lee, "Learning from few examples: A summary of approaches to few-shot learning," *arXiv preprint arXiv:2203.04291*, 2022.

[27] S. Clarke, R. Gao, M. Wang, M. Rau, J. Xu, J.-H. Wang, D. L. James, and J. Wu, "Realimpact: A dataset of impact sound fields for real objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1516–1525.

[28] Y. Wang, N. J. Bryan, J. Salamon, M. Cartwright, and J. P. Bello, "Who calls the shots? rethinking few-shot learning for audio," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 36–40.