WAVE-TO-SOUND: REAL-TIME INTERACTIVE MUSIC CONTROL VIA ULTRASONIC HAND GESTURE CLASSIFICATION

Meongeun Kim
Graduate School of Metaverse
meongeun@kaist.ac.kr

Changhyeon Park
Graduate School of Culture Technology
sac7160@kaist.ac.kr

ABSTRACT

This project aims to develop an interactive system that utilizes ultrasonic audio data to classify hand gestures and apply them to real-time music control. Existing vision-based or sensor-based hand gesture recognition methods have limitations such as illumination, occlusion, and wearability. To solve these problems, this project adopts an approach that transmits ultrasonic waves to the hand and receives reflected signals in multichannel to classify gestures. Audio features such as raw waveform, time difference (TDoA), phase difference (IPD), and energy change were extracted from the received signals, and these features were trained using various deep learning models (CNN, CNN-RNN, etc.) to improve recognition performance. On the hardware side, we analyzed the differences in spatial information depending on the number of microphones and their placement configuration, and on the software side, we achieved recognition accuracy of up to 89% through feature fusion and network structure optimization. This system offers the possibility of intuitively controlling music with hand gestures without lighting or privacy issues.

1 Introduction

Hand Gesture Recognition is becoming a key technology for realizing user-friendly interfaces in various interaction systems such as augmented reality (AR), virtual reality (VR), smart home, and music control. Existing gesture recognition systems can be broadly categorized into vision-based and sensor-based approaches. Vision-based methods [5] have the advantage of enabling natural interaction by directly tracking hand movements with a camera, but they are vulnerable to various environmental constraints such as lighting changes, background complexity, occlusion, and privacy issues. Sensor-based methods [2], on the other hand, offer high accuracy and fast response times, but require users to wear equipment and suffer from sensor drift and calibration issues.

As an alternative to overcome the limitations of these existing methods, hand gesture recognition techniques utilizing ultrasonic sensing have been proposed in recent years. For example, BeamBand [3] analyzed ultrasonic signals reflected from the hand to recognize gestures, and Zhang et al. [6] proposed a method to estimate finger joint positions using ultrasonic transceivers. However, these studies still have limitations such as system complexity and low accuracy. In addition, existing work often classifies ultrasound signals as simply raw waveforms or does not utilize rich audio features.

In this project, we propose a novel ultrasonic-based gesture recognition system and its application to music effect modulation. In particular, by extracting various audio-based features such as TDoA, IPD, and energy curves as well as raw waveforms and integrating them into a multichannel deep learning model, we designed a structure that can simultaneously learn spatial and temporal information. In addition, we experimentally analyzed the differences in reflection patterns depending on the number and location of microphones, and optimized the hardware configuration based on this.

By realizing high-accuracy hand gesture recognition without wearing any devices on the user's hand, independent of lighting conditions and privacy issues, this research shows the potential for various applications such as real-time music control.

2 Related works

2.1 Hand Gesture Recognition

Hand gesture recognition is an emerging area of research based on a variety of input devices, including computer vision, wearable sensors, and ultrasound. Recently, approaches have been proposed that utilize lightweight deep learning models suitable for portable or real-time interfaces. For example, Sen et al. [5] implemented camera image-based hand gesture recognition using a model optimized by lightweight channel pruning on YOLOv5s, which significantly improved computational efficiency. On the other hand, Filipowska et al. [2] implemented a machine learning-based system to classify gestures by embedding pressure and bending sensors inside a wearable glove to precisely track finger movements. While this approach demonstrated high recognition accuracy, practical limitations exist, such as the user burden of wearing gloves and sensor calibration issues. Ultrasound-based approaches utilize the reflection patterns generated by hand gestures and have the advantage of recognizing gestures without visible light or wearing sensors. Zhang et al. [6] utilized

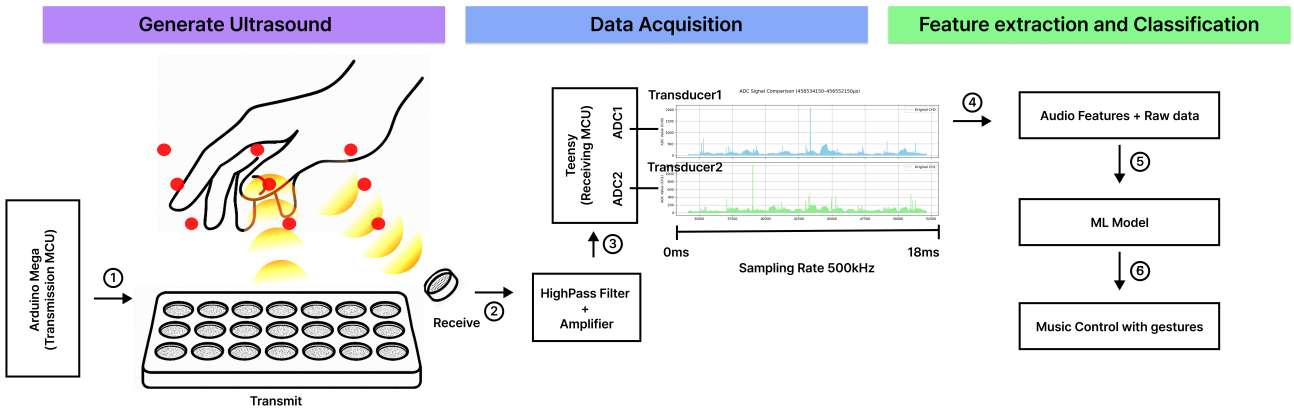

Figure 1. System Overview.

MEMS ultrasonic sensors to estimate the position of finger joints in 3D and proposed a system that accurately interprets echo-based time-of-flight (ToF) information. However, this approach is limited by the accuracy requirement of several centimeters and the difficulty of synchronization among multiple sensors

2.2 Audio-based Feature

Extracting meaningful features from audio-based signals is a key step in sound classification, source localization, and atypical gesture recognition, like in this study. In particular, in echo-reflection analysis using ultrasound, various representations have been studied that can reflect spatiotemporal features beyond the simple raw waveform. Liu et al. [4] proposed a method that utilizes STFT-based spectral analysis and inter-channel phase difference (IPD) and time difference (TDoA) information to accurately estimate the location of multiple sound sources. In this process, a residual attention mechanism was applied to maintain high accuracy even in noisy environments. Furthermore, in general sound source localization (SSL) research, a combination of features such as IPD, Generalized Cross-Correlation (GCC), Envelope Energy, and Spectral Flatness is often used to interpret complex acoustic scenes [1].

3 Method

The system aims to enable contactless interaction through ultrasound-based hand gesture recognition, enabling real-time music control. To this end, the overall pipeline consists of four main stages, as illustrated in Figure 1: (1) ultrasonic signal generation and echo reception, (2) data acquisition, (3) feature extraction and gesture classification, and (4) music control application. Users can intuitively manipulate music using simple hand gestures without the need for cameras or wearable sensors, while the system effectively captures hand movements by leveraging the spatiotemporal characteristics of reflected ultrasound signals.

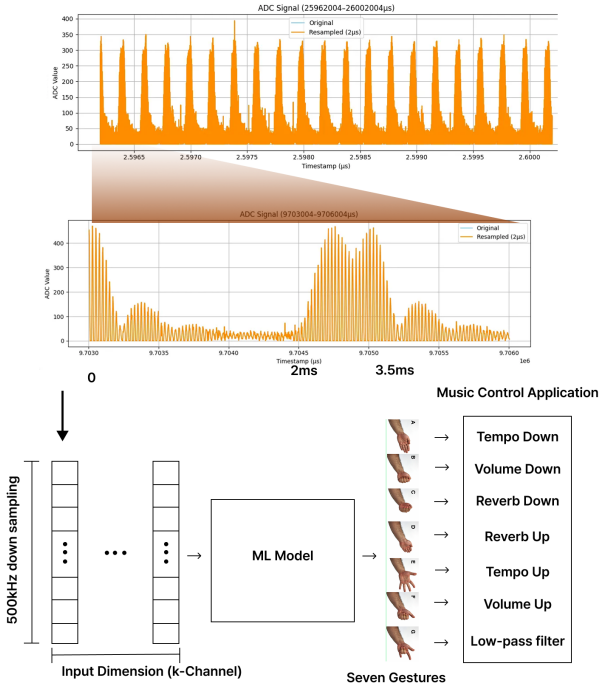

Figure 2. Figure captions should be placed below the figure.

3.1 Ultrasonic Data Acquisition System

The system transmits ultrasonic waves sequentially for a total of 9 transmit directions, with a period of 2 ms (0.5 ms transmit + 1.5 ms receive) for each direction. By repeating this process nine times, a total of 18 ms of data is obtained for a single hand gesture. The transmit signal is generated by the PC and delivered to the transducer array via the transmit MCU, which switches directions and transmits ultrasonic waves in a predefined sequence.

The ultrasonic waves reflected from the user's hand are detected by the receiving transducers and are then digitized by the receiving MCU at high speed sampling at 2 μs intervals (500 kHz) through a high-pass filter and amplifier. The receiving microphone consists of two or more channels and is designed to capture spatial information, includ-

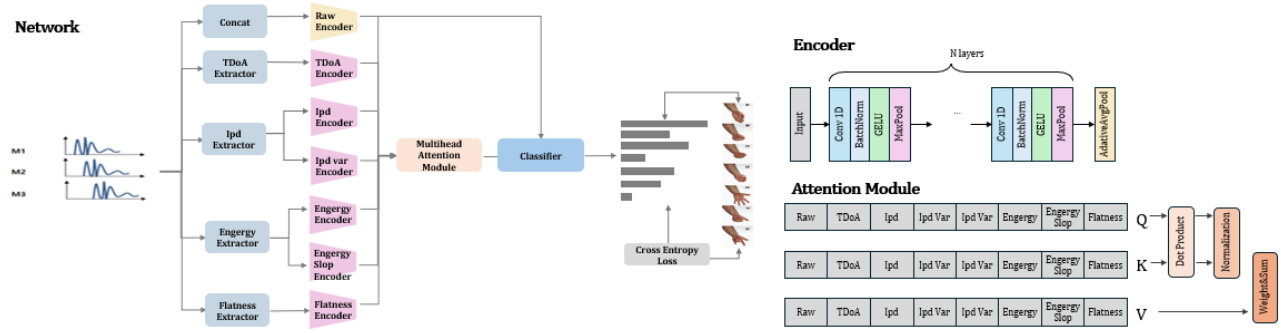

Figure 3. Model Framework

ing reflection time delay and phase differences, through position differences between channels.

The collected data is stored as a time series array of size (9000, k), where 9000 is the number of samples in 18 ms and k is the number of microphone channels. The ADC values are visually represented at the top of Figure 2, where the reflected echo pattern is clearly distinguishable depending on the gesture.

3.2 Feature Extraction

Given multi-channel ultrasonic signals $\mathbf{x}_i(t)$ from microphones M_i, where $i \in \{1, 2, 3\}$, we extract the following modality-specific features.

1. Raw Waveform. The raw time-domain signal from three microphones is defined as:

$$\mathbf{x}_{\text{raw}}(t) = [x_1(t), x_2(t), x_3(t)] \in R^3 \tag{1}$$

2. Time Difference of Arrival (TDoA). TDoA is computed as the pairwise sample-wise difference between channels:

$$\begin{aligned} \text{TDoA}_{i,j}(t) &= x_i(t) - x_j(t), \\ (i,j) &\in \{(1,2),(1,3),(2,3)\} \end{aligned} \tag{2}$$

The resulting 3-channel signal is:

$$\mathbf{x}_{\text{tdoa}}(t) = [\text{TDoA}_{1,2}(t), \text{TDoA}_{1,3}(t), \text{TDoA}_{2,3}(t)] \tag{3}$$

3. Interchannel Phase Difference (IPD). Let $Z_i(t, f)$ be the STFT of $x_i(t)$, then IPD is defined as:

$$\text{IPD}_{i,j}(f,t) = \angle Z_i(t,f) - \angle Z_j(t,f) \tag{4}$$

Temporal average over frequency yields:

$$\text{IPD}_{i,j}(t) = \frac{1}{F}\sum_{f=1}^{F} \text{IPD}_{i,j}(f,t) \tag{5}$$

The temporal variance of IPD is computed as:

$$\text{Var}_{\text{ipd}}^{i,j}(t) = \text{Var}_f\left(\text{IPD}_{i,j}(f,t)\right) \tag{6}$$

5. Envelope Energy. The envelope energy is obtained using the Hilbert transform:

$$\begin{aligned} E_i(t) &= \left|\mathcal{H}[x_i(t)]\right|, \\ \mathbf{x}_{\text{energy}}(t) &= [E_1(t), E_2(t), E_3(t)] \end{aligned} \tag{7}$$

The energy slope is approximated using finite differences:

$$\mathbf{x}_{\text{energy\_slope}}(t) = \frac{d\,\mathbf{x}_{\text{energy}}(t)}{dt} \approx \mathbf{x}_{\text{energy}}(t) - \mathbf{x}_{\text{energy}}(t-1) \tag{8}$$

7. Spectral Flatness

Using the magnitude STFT $|Z_i(t, f)|$, spectral flatness is computed as:

$$\text{Flatness}_i(t) = \frac{\left(\prod_{f=1}^{F} |Z_i(t,f)|\right)^{\frac{1}{F}}}{\frac{1}{F}\sum_{f=1}^{F} |Z_i(t,f)|} \tag{9}$$

The final flatness vector is:

$$\mathbf{x}_{\text{flatness}}(t) = [\text{Flatness}_1(t), \text{Flatness}_2(t), \text{Flatness}_3(t)] \tag{10}$$

Combined Representation

Finally, we concatenate all features channel-wise:

$$\begin{aligned} \mathbf{x}_{\text{input}}(t) = [&\mathbf{x}_{\text{raw}}(t) \parallel \mathbf{x}_{\text{tdoa}}(t) \parallel \mathbf{x}_{\text{ipd}}(t) \parallel \mathbf{x}_{\text{ipd\_var}}(t) \\ &\parallel \mathbf{x}_{\text{energy}}(t) \parallel \mathbf{x}_{\text{energy\_slope}}(t) \parallel \mathbf{x}_{\text{flatness}}(t)] \end{aligned} \tag{11}$$

where $\parallel \cdot \parallel$ denotes channel-wise concatenation.

3.3 Deep Learning Model

To fully exploit the rich spatiotemporal information encoded in ultrasonic echoes, we designed a deep learning model based on a 1D Convolutional Neural Network (CNN) framework, as illustrated in Figure 3. Unlike classical classifiers that rely on flattened feature vectors and assume feature independence, our CNN-based model can hierarchically extract local temporal dependencies from sequential input data. This design choice is motivated by the nature of acoustic signals, where temporal dynamics such as echo delay patterns, inter-microphone phase shifts (IPD), and envelope variations encode subtle gesture cues that must be modeled jointly.

The model architecture consists of a stack of 1D convolutional layers with progressively increasing receptive fields, interleaved with Batch Normalization, GELU activations, and MaxPooling for regularization and temporal abstraction. These layers are followed by fully connected

Figure 4. Music Application Demo and Visualization

(FC) projection layers to map the extracted high-level features into class logits for gesture prediction. We empirically selected convolution kernel sizes and channel widths to balance model expressiveness and generalization capability under a limited dataset regime.

To accommodate multi-modal inputs (e.g., raw waveform, TDoA, IPD, energy features), we also designed modality-specific CNN encoders, each responsible for capturing domain-specific signal structure. The outputs from each encoder are subsequently fused using attention-based mechanisms to allow the model to dynamically weight the most informative modalities depending on the gesture instance.

The model was trained using a weighted cross-entropy loss to address class imbalance and CosineAnnealingLR was applied for learning rate scheduling to facilitate smoother convergence.

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N} w_{y_i} \cdot \log p_{i,y_i} \tag{12}$$

3.4 Music Control Application

The collected signals are converted into various audio features (raw waveform, TDoA, IPD, etc.) in the feature extraction stage and utilized for real-time music control through classification. As illustrated in Figure 4, each gesture is mapped to a specific audio control function—for instance, flexion increases the tempo, spiderman activates a low-pass filter, thumbs-up adjusts the volume upward, and wrist extension enhances the reverb effect.

This application allows users to intuitively manipulate music using only hand gestures without environmental constraints such as external lighting, viewing angle, or wearing equipment, and has a wide range of potential applications in performance, interactive art, and everyday environments.

4 Experiment

We conducted comprehensive experiments to evaluate the performance of our system in terms of microphone configuration, feature representation, and model architecture. Each experiment was aligned with a specific hypothesis to identify performance bottlenecks and optimization directions.

4.1 Effect of Microphone Configuration

To test Hypothesis 1 (H1)—that increasing the number of microphones and optimizing their spatial arrangement improves recognition accuracy—we evaluated seven different configurations ranging from single-microphone setups to all three microphones combined (Mic 0–1–2). As shown in Table 1, the full three-microphone setup achieved the highest classification accuracy of 84%, outperforming all dual and single configurations. This result highlights the importance of spatial diversity in capturing high-resolution echo profiles and resolving gesture ambiguities. Notably, while Mic 1 alone performed relatively well in recognizing wrist extension gestures (95%), it lacked generalizability across other classes, confirming that single-microphone configurations are insufficient for robust recognition.

Among the dual-microphone setups, the Mic 0–1 configuration stood out, achieving a high accuracy of 82%, outperforming other dual combinations such as Mic 0–2 and Mic 1–2. This result suggests that the relative placement and angular coverage of Mic 0 and Mic 1 creates an optimal spatial baseline for discriminating hand motion trajectories. Overall, the results support both aspects of the hypothesis: microphone quantity enhances signal diversity, while strategic spatial placement maximizes directional sensitivity and gesture discriminability.

Table 1. Classification accuracy per gesture by microphone configuration. 3-mic setup achieves best overall accuracy.

Placement	Extension	Flexion	Spiderman	Stretch	Thumbs_up	Wrist_ext.	Wrist_flex.	Accuracy
Mic 0	0.58	0.52	0.49	0.62	0.72	0.83	0.88	0.66
Mic 1	0.69	0.53	0.50	0.55	0.59	**0.95**	0.74	0.65
Mic 2	0.52	0.47	0.56	0.43	0.70	0.67	0.85	0.60
Mic 0-1	0.72	**0.73**	**0.82**	**0.75**	**0.90**	0.94	0.97	**0.83**
Mic 1-2	0.72	0.68	0.63	0.65	0.69	0.87	0.91	0.73
Mic 0-2	0.71	0.62	0.60	0.60	0.75	0.90	0.95	0.72
Mic 0-1-2	**0.78**	0.80	0.81	0.71	0.87	0.92	**0.97**	0.84

4.2 Feature Representation and Fusion

For Hypothesis 2 (H2), we hypothesized that multi-modal audio features contain complementary information and that combining them would enhance performance. The results in Table 2 support this claim. Using only the raw waveform yielded a strong baseline (F1-score = 0.83), but when combined with time difference of arrival (TDoA), interchannel phase difference (IPD), energy slope, and spectral flatness, performance improved to an F1-score of 0.85.

Interestingly, we observed that early fusion using uniform encoders degraded performance, likely due to the incompatibility of temporal and statistical features when processed jointly without disentanglement. In contrast, late fusion with separate encoders preserved the modality-specific characteristics and yielded better generalization. Moreover, residual connections within the raw encoder allowed the model to retain high-frequency local patterns while integrating spatial cues.

4.3 Model Performance Comparison

To validate Hypothesis 3 (H3)—that CNN-based models are better suited for ultrasonic gesture classification than traditional or sequential models—we compared the proposed 1D CNN model with several machine learning and

Table 2. Performance comparison of different feature fusion strategies and encoders. Raw-based late fusion achieves the best overall performance.

Setting	Fusion Type	Used Features	Precision	Recall	F1-score
Uniform Encoder	Early Fusion	Raw	**0.84**	0.83	0.83
+ Multi-feature Fusion	Early Fusion	Raw, TDoA, IPD, Stats	0.77	0.77	0.77
Separate Encoder	Late Fusion	Raw, TDoA, IPD, IPD_VAR, Flatness, Energy_Slope	0.82	0.82	0.82
+ Raw-based Fusion	Late Fusion	Raw + [TDoA, IPD, IPD_VAR, Flatness, Energy_Slope]	**0.86**	**0.86**	**0.85**

deep learning baselines, as shown in Table 3. Through extensive hyperparameter tuning—including adjustments to the number of convolutional layers, learning rate, dropout rate, and hidden unit size—the CNN model achieved a validation accuracy of 89%, outperforming classical methods (e.g., Logistic Regression at 67%, SVM at 65%) and sequential models (GRU at 77%, Transformer at 78%).

Table 3. Validation accuracy of traditional machine learning and deep learning models. Our model outperforms all baselines.

	Machine Learning								Deep Learning		
	SGD	SVM	KNN	RF	LogReg	MLP	LDA	GNB	GRU	Transf.	Ours
Accuracy	0.63	0.65	0.34	0.62	**0.67**	0.63	0.62	0.57	0.77	0.78	**0.82**

These results illustrate the CNN's strength in extracting localized patterns from audio signals, while avoiding the training instability often associated with recurrent architectures. The superior performance of the CNN can also be attributed to its ability to generalize across modalities, as its convolutional layers learn hierarchical representations from multi-modal audio inputs.

In a direct comparison with the state-of-the-art BeamBand system, Table 4 demonstrates that our model outperforms BeamBand across most gesture classes in terms of precision. Particularly, it achieves a substantial margin in detecting complex gestures like "thumbs up" and "wrist flexion", indicating improved robustness and temporal discrimination.

Table 4. Per-class performance comparison between BeamBand and our model. Our model shows superior precision in most gesture classes.

		Extension	Flexion	Spiderman	Stretch	Thumbs_up	Wrist_ext.	Wrist_flex.	Accuracy
BeamBand	Precision	**0.84**	**0.83**	0.82	0.71	0.79	0.89	0.83	0.81
Ours	Precision	**0.84**	**0.83**	**0.87**	**0.81**	**0.92**	**0.99**	**0.96**	**0.89**
	Recall	0.84	0.83	0.87	0.91	0.90	0.96	0.98	–
	F1-score	0.84	0.83	0.78	0.86	0.91	0.98	0.97	–

5 Conclusion

In this work, we present Wave-to-Sound, an ultrasonic hand gesture recognition system capable of real-time music control without relying on cameras or wearable sensors. By integrating rich acoustic features—such as TDoA, IPD, and energy dynamics—with a multi-modal CNN framework, we achieved high gesture classification performance (up to 89%) under varying hardware and signal conditions.

Our contributions are threefold:

- **Hardware-Software Co-Design:** We implement an full pipeline from ultrasonic signal processing to gesture-driven music modulation.
- **Multi-Modal Feature Learning:** We show that combining raw and derived audio features enables more robust and expressive gesture representation.
- **Real-Time and Intuitive Interaction:** Our 1D CNN-based architecture balances accuracy and computational efficiency, enabling real-time and intuitive music manipulation without the need for external lighting, cameras, or wearable devices.

These results open possibilities for novel interfaces in interactive media, performance, and ambient computing where non-contact and privacy-respecting input modalities are essential.

6 Author Contributions

Meongeun Kim extracted audio features from the raw data and implemented entire deep and machine learning models. And she conducted experiments for performance improvement and implemented music modulation application. Changhyeon Park came up with an idea. And he implemented the ultrasonic generation/data acquisition system and collected datasets. They wrote the report together.

7 References

[1] Dhwani Desai and Ninad Mehendale. A review on sound source localization systems. *Archives of Computational Methods in Engineering*, 29(7):4631–4642, 2022.

[2] Anna Filipowska, Wojciech Filipowski, Paweł Raif, Marcin Pieniażek, Julia Bodak, Piotr Ferst, Kamil Pilarski, Szymon Sieciński, Rafał Jan Doniec, Julia Mieszczanin, et al. Machine learning-based gesture recognition glove: Design and implementation. *Sensors (Basel, Switzerland)*, 24(18):6157, 2024.

[3] Yasha Iravantchi, Mayank Goel, and Chris Harrison. Beamband: Hand gesture sensing with ultrasonic beamforming. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–10, 2019.

[4] Mengran Liu, Hanghai Feng, Qiang Zeng, Chuanqi Gong, Chao Zhou, and Zeming Jian. Multi-source localization method based on enhanced stft and residual attention mechanism. *Sensor Review*, 2025.

[5] Abir Sen, Tapas Kumar Mishra, and Ratnakar Dash. Novel human machine interface via robust hand gesture recognition system using channel pruned yolov5s model. *arXiv preprint arXiv:2407.02585*, 2024.

[6] Qiang Zhang, Yuanqiao Lin, Yubin Lin, and Szymon Rusinkiewicz. Hand pose estimation with mems-ultrasonic sensors. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023.