Dual-Layer Composite Model and IoT-Based Indoor Sound Visualization System for the Deaf or Hard of Hearing Community

Student: Yi-Chien Chen, Yu-Hsuan Wu, Hung-Ai Huang

Motivation

In the past, equipment for monitoring sound events was often costly, and the systems lacked real-time capability, making them difficult to apply in daily life. We proposed a system with real-time capability that offers accurate prediction of 13 important categories and the ability to predict up to 521 categories overall. The results are visualized on smartphones and smartwatches, enhancing user accessibility and interaction.

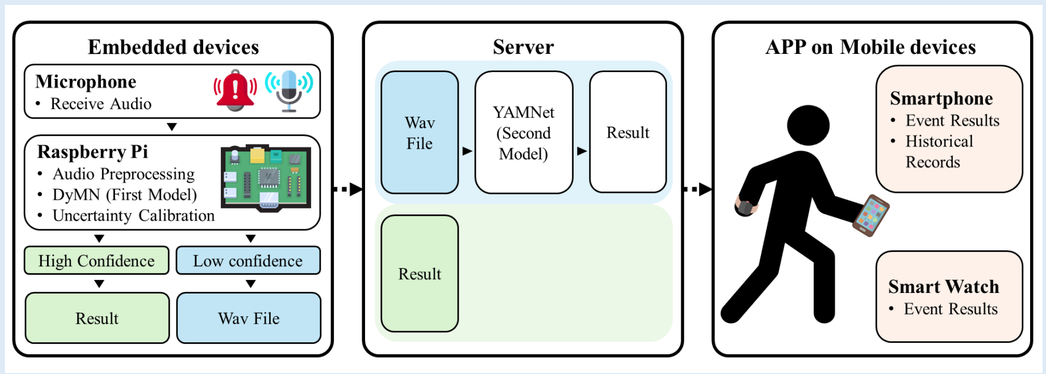

Fig. 1. Overall system architecture diagram.

Methods

Fig. 1 shows the overall system architecture. The microphone detects environmental sounds in real-time, processed by a trained sound classification model on the Raspberry Pi, using the Scaling-Binning Calibrator for improved calibration. Based on the model's confidence, results are either sent directly to the client or forwarded to the server's larger model, YAMNet, for further classification.

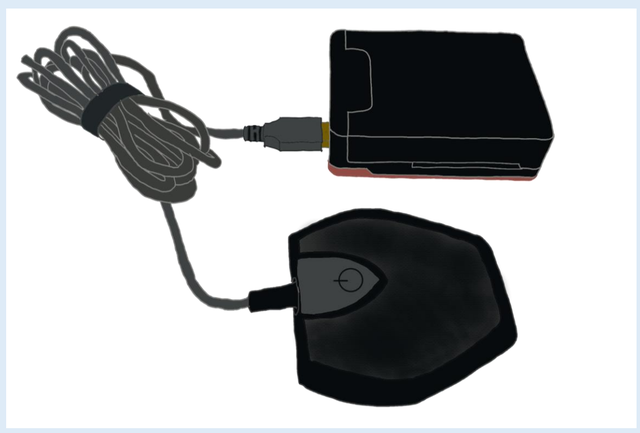

Fig. 2. Hardware Device.

Fig. 2 is hardware device in our system. We selected the Raspberry Pi 5 for its enhanced CPU capabilities, ensuring faster and smoother real-time sound detection. The MIC-G11 microphone, utilizing USB 3.0 for connectivity, captures sounds from all directions up to 8 meters away without angle adjustments, making it ideal for extensive and accurate sound detection.

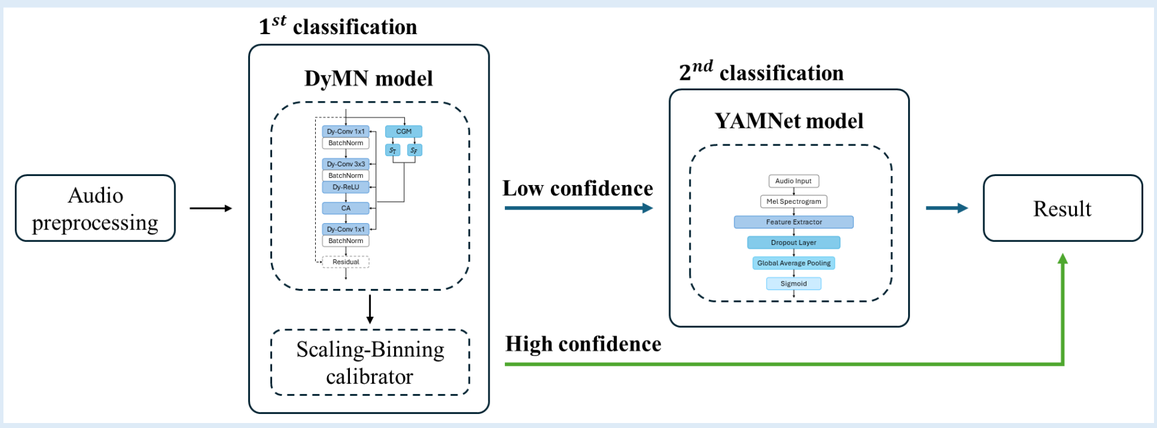

Fig. 3. Software Architecture

Fig 3 is software architecture, includes signal preprocessing, a small model (DyMN with uncertainty-calibration), and the large model YAMNet, along with the final classification results for the user. We enhance the small model's edge device performance using a Transformer-to-CNN knowledge distillation technique and employ a scaling-binning calibrator for precise error estimation with few samples. If the small model's confidence is low, audio is sent to YAMNet on the server, which classifies up to 521 sound types for thorough recognition. The small and large models collaborate to ensure precise, efficient sound detection, with the results relayed to the user via the server.

Table 1. Categories used for the small model

ESC50			ReaLISED
dog	crying_baby	washing_machine	speech
cat	laughing	vacuum_cleaner	walking
water_drops	door_wood_knock	clock_alarm	
pouring_water	glass_breaking		

We processed specific audio categories from ESC50 and ReaLISED (see Table 1) into a dataset of 100 five-second samples per category, formatted at 16 kHz, 32-bit mono. Features were calculated using the small model's method, and the model was trained, validated, and tested in an 8:1:1 ratio. The final model achieved an 87.9% accuracy on the test set.

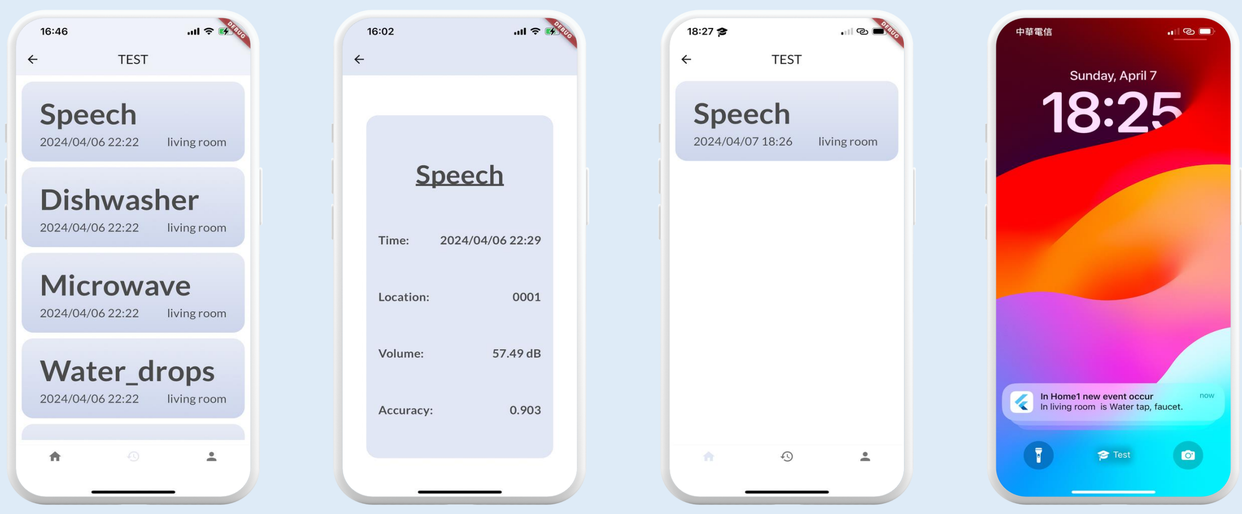

Fig. 4. User Interface

We used Flask to manage data transmissions between the Raspberry Pi and server, and from the server to users. Real-time data is uploaded by the Raspberry Pi in .json format to a local network, with classification results stored on the server, and audio files processed further by the large model.

Results and Discussion

Fig. 4 displays our user interface, which provides real-time sound notifications and access to event details and historical events. The comparison of our system with past related research is shown in Table 2, which has great advantages.

Table 2. Comparison with other related systems

Work	Hardware Device	Model	Speed	Accuracy	User Interface	Categories	Dataset
[1]	SurfacePro 3	VGG16 (CNN)	Real-time	85.9%	Yes	19	Limited
[2]	Raspberry Pi + Thingy 52	CNN	Real-time	94.64%	No	16	Limited
[3]	Raspberry Pi 4B	YAMNet + CNN-1D	Real-time	96.66%	No	10	Limited
our	Raspberry 5 + Microphone	dymn + YAMNet	Real-time	87.9%	Yes	13 + 521	Flexible

Conclusion

This research introduces a home sound visualization system aimed at helping deaf or hard of hearing individuals. It not only breaks through the limitations on sound categories but also enhances user accessibility and interaction.