

Recognition System Prototyping for Underwater Objects Based on Acoustic Data Processing

Chua-Chin Wang
Institute of IC Design
National Sun Yat-Sen University
Kaohsiung, Taiwan
ccwang@ee.nsysu.edu.tw

Abstract—This research proposes a robust prototyping system for underwater object recognition (UOR) leveraging deep learning and hardware acceleration. A novel recognition algorithm as well as hardware realization utilizing Mel-Frequency Cepstral Coefficients (MFCC) is demonstrated. MFCCs provide a hardware-efficient alternative that can be converted into 2D spectrograms, allowing resource sharing with image recognition hardware for low-power operation. A specialized AI Edge accelerator is proposed to minimize power consumption. The design optimizes the conventional CNN pipeline by integrating Convolution, ReLU, and Pooling into a single processing step, reducing pipeline stages and idle time. The experimental results of the proposed system outperform the state of art when multi-category object sounds are given.

Keywords—Sonar sensor node, MFCC, acoustic signature recognition, AI in water, prototyping

I. INTRODUCTION

Due to the challenges of observation in underwater environments, underwater acoustic signature recognition systems possess the capability to perform long-range, deep-sea detection under all conditions. One of the major methods for identifying low-frequency and low-SNR (Signal-to-Noise Ratio) sounds is through Gammatone Frequency Cepstral Coefficients (GFCCs), an auditory-perceptual feature representation widely used in underwater object recognition. GFCC features are obtained via a Gammatone filter bank followed by a Discrete Cosine Transform (DCT), offering superior capability in representing the spectral structure of signals. By contrast, another approach is based on Mel-Frequency Cepstral Coefficients (MFCC) extracted from acoustic signals as model inputs. The fundamental difference between GFCC and MFCC lies in their design principles: the GFCC filter response is based on a cochlear frequency decomposition model, whereas MFCC relies on the Mel scale. However, because GFCC can stably detect non-linearities and spectral envelopes, its features perform exceptionally well in describing acoustic signals, particularly in low-SNR scenarios. However, the implementation of the GFCC on hardware will demand too much power and cost due to the complexity [1][2][3].

The Mel scale is a non-linear frequency scale designed based on human auditory perception. Compared to linear spectra, the non-linear Mel spectrogram is better suited for deep learning models. When generating a Mel-spectrogram, the default setting typically defines the maximum intensity as the reference value (0 dB), with all other values converted into negative numbers relative to this peak.

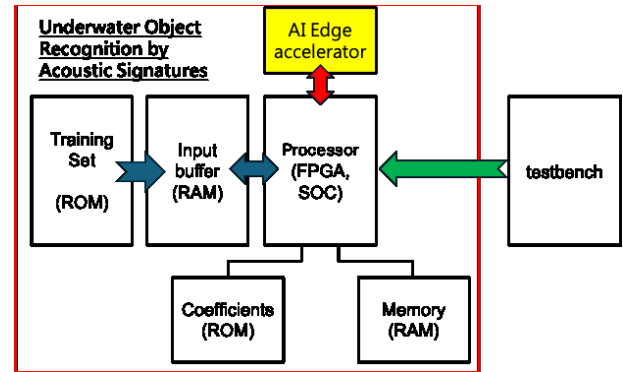


Fig. 1. Proposed prototype diagram of Underwater Object Recognition (UOR) by Acoustic Signatures

This study then explored deep learning based approaches, utilizing Mel-Frequency Cepstral Coefficients (MFCC) extracted from acoustic signals as model inputs to classify signatures from various sources, such as dolphins, fish, ships, whales, etc. Besides the developed algorithm, a prototype system is demonstrated for field experiments.

II. UNDERWATER OBJECT RECOGNITION BY ACOUSTIC SIGNATURES

Fig. 1 illustrates the proposed system of Underwater Object Recognition (UOR) by Acoustic Signatures, which contains 3 major blocks, namely Processor, AI Edge accelerator, and memories (RAMs, and ROMs). The entire hardware is design to carry out the signal processing flow shown in Fig. 2.



Fig. 2. Acoustic signal processing flow

A. Acoustic Data Acquisition for Underwater Objects

The primary obstacles in underwater acoustic recognition firstly are the difficulty of data acquisition and the imbalance of raw datasets, alongside inconsistent signal durations, which often lead to poor generalization in established models. It is well known that the publicly available databases are quite useful in the training or learning phase of AI-based recognition approaches, e.g., DOSITS (Discovery of Sound in the Sea) - Audio Gallery [7].

To obtain the requisite volume of underwater acoustic data, long-term deployment of underwater hydrophone systems is necessary. Referring to Fig. 3, our team has

deployed multiple sonar sensor nodes in several locations around Taiwan island over several years to record the sounds generated by various objects, including dolphins, whales, sciaenidae, waves, rains, boat, vessel, etc.

The acoustic signature dataset currently collected by our team consists of: biological species (21,750), natural background noise (300), and ships (37,649). All original audio files (.wav) were uniformly cropped to a duration of 2 seconds. For the two categories with larger sample sizes, we performed random selection after removing highly redundant audio segments. Conversely, for the category with a smaller sample size, data augmentation was implemented by adding white noise. Ultimately, all three categories were normalized to 3,000 samples each. Fig. 4 (a) and (b) are converted time-spectral samples of dolphin and ship.



Fig. 3. Acoustic data acquisition around Taiwan

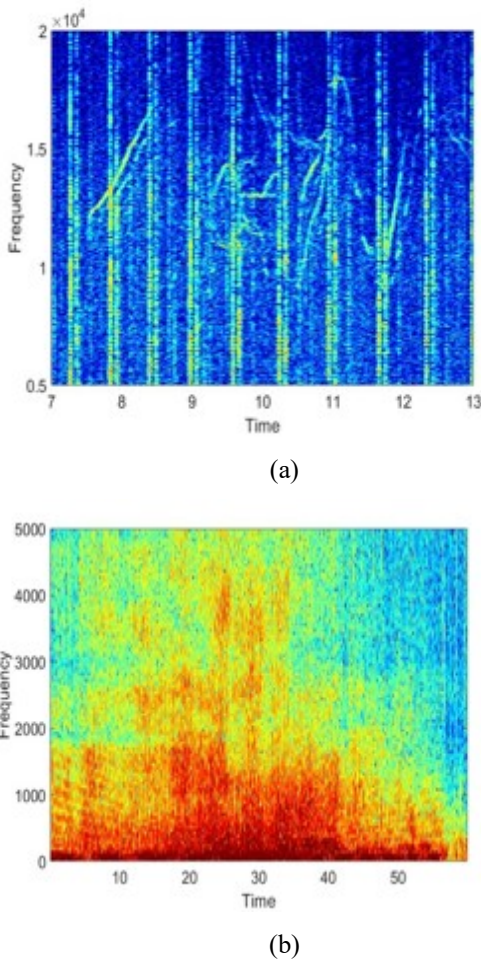


Fig. 4. Samples of time-frequency spectrals of (a) dolphin; (b) ship

Taking dolphin vocalizations as an example for AI learning to carry out underwater object recognition: their sounds are primarily concentrated around 2 kHz. When these are converted into Mel spectrograms, the extracted audio feature maps serve as the input for the AI model to be trained and recognized. Referring to Fig. 5 (a) and (b), Mel Spectrogram and MFCC of Dolphin's voice are shown, respectively.

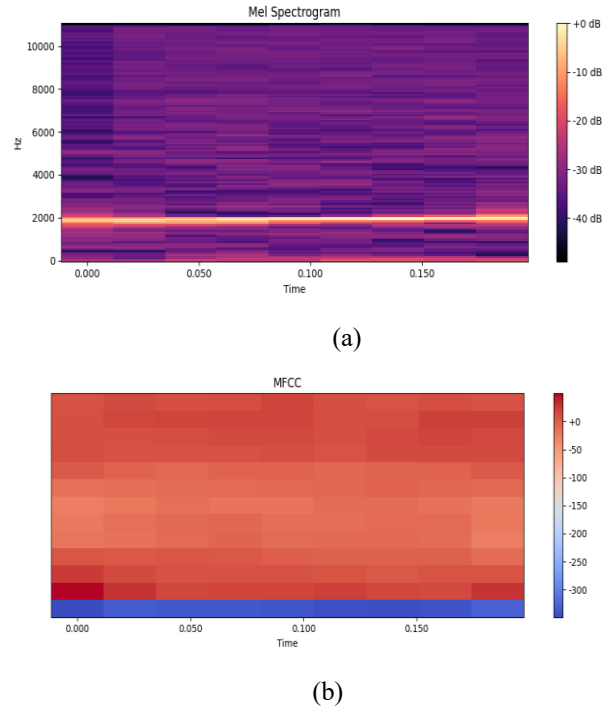


Fig. 5. Dolphin's voice (a) Mel Spectrogram; (b) MFCC

B. AI Training Model

Before the realization of the recognition system on hardware, it is needed to verify that the proposed AI training model is feasible. The primary advantage of utilizing the said MFCC approach is that, in addition to its effectiveness in acoustic feature extraction, it can be converted into a spectrogram format for training and recognition. This allows for the integration—or at least partial resource sharing—with the hardware used for image/video recognition if necessary [4][5][6], thereby achieving low power consumption.

As shown in the proposed AI network model in Fig. 6, the MFCC of the acoustic signal is used as the input to determine the final category of the sound. This model employs ResBlocks (Residual Blocks) for feature extractions, followed by Fully Connected (FC) layers for classification.

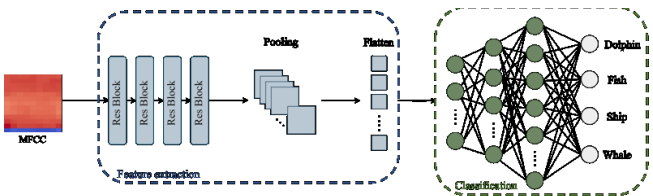


Fig. 6. AI network model for MFCC training and recognition

C. AI Accelerator Architecture in Edge Devices

The most common operation in all AI neural networks and machine learning is convolution, which is a derivative of the MAC (Multiply-Accumulate) operation—the most power-consuming component in such systems. To seriously reduce the overall power dissipation in the AI edges, we propose to re-organize the computation flow of the AI network and reduce the bit length without too much penalty on accuracy. Implementing a conventional 6-step CNN architecture directly, as shown in Fig. 7(a), would result in low processing speeds and high power consumption. Therefore, we propose a simplified architecture that integrates Conv+ReLU+Pool into a single processing step, as illustrated in Fig. 7 (b). This approach reduces the number of pipeline stages, thereby enhancing overall speed. Furthermore, the original floating-point operations will be quantized to int16; if the accuracy remains within an acceptable range, further reduction to int8 is also taken into consideration.

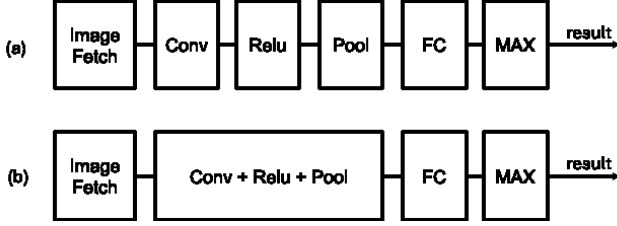


Fig. 7. (a) conventional CNN computation pipeline : (b) simplified CNN pipeline

Taking a 3×3 kernel as an example, a conventional MAC unit requires a total of nine multiply-accumulate operations to complete a single convolution window. After each window is processed, a reset is required. Since the system cannot output data or perform further calculations during a traditional reset period, a standard architecture would result in significant idle time. By contrast, the simplified CNN pipeline architecture will save a great portion of the idle time.

D. Processing Element (PE) in AI Accelerator

Based on our previous AI Accelerator designs [5][6], we adopt a modified version as shown in Fig. 8, where an 8×32 PE (Processing Element) array architecture consisting of 8-row output window parallelism and 32-column output channel parallelism. The outputs are shifted to the left, and PEs within the same column share a common set of kernel values. To minimize the hardware cost and power consumption of the array, the activation functions, accumulation of pixels at the same position across different input channels, and quantization are performed only when data is written to the output SRAM. By not embedding individual activation functions and accumulators within every PE, any redundant synthesis of these components in each unit are no longer needed. The PE details are given in Fig. 9 [5][6].

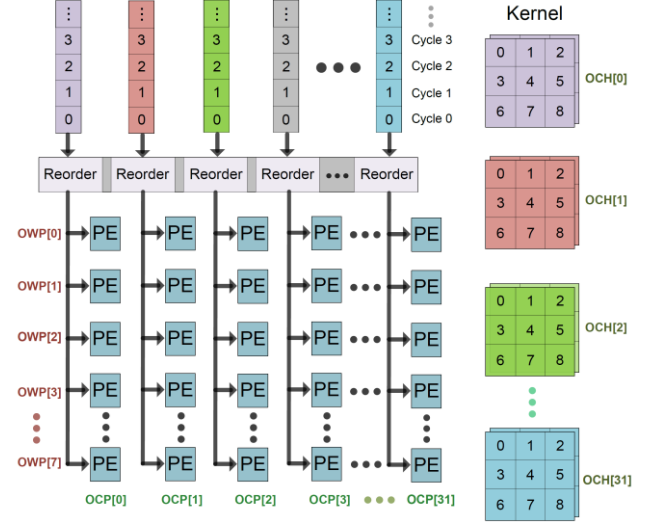


Fig. 8. PE arrangement in AI accelerator

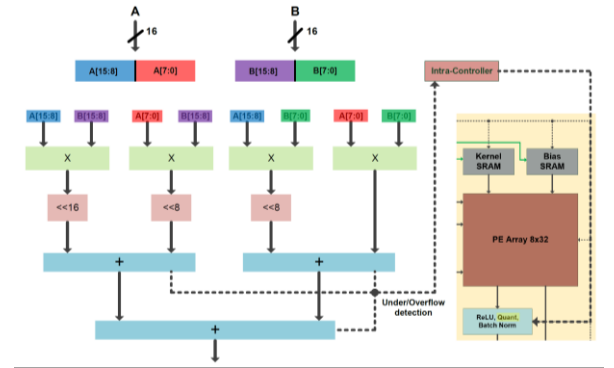


Fig. 9. Signal flow in PE

III. VERIFICATION AND SYSTEM EXPERIMENT

A. Conversion of Underwater Sound into Freq-Spectrum

Kv260 FPGA board with DPU (deep learning unit) is used to carry out the conversion of collected underwater sounds. We execute an application written in C on the system, in which the audio streaming components are encapsulated as a GStreamer plugin for processing. The application is designed to ingest audio data, convert the audio into spectrograms, perform model inference via the DPU, and output the final results. Fig. 10 shows some samples of the conversion.

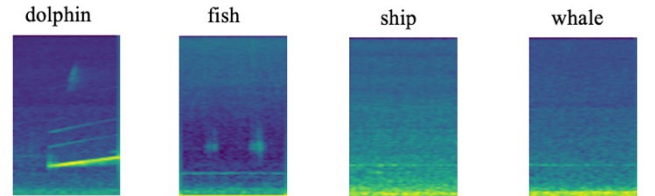


Fig. 10. Samples of converted audio signals

B. Model Training

The training of a CNN model involves iterative processes using a training set, where parameters are updated through an optimizer. Once training is complete, an independent test set

is used to verify the recognition accuracy. The reason for separating the training and test sets is to ensure the generalization of the recognition accuracy and to prevent the parameters from overfitting to the training data. Referring to Fig. 11, two processes are required.

- Forward Propagation : the neural network layers to compute the predicted output
- Backward Propagation: updating parameters using optimization algorithms to minimize the error

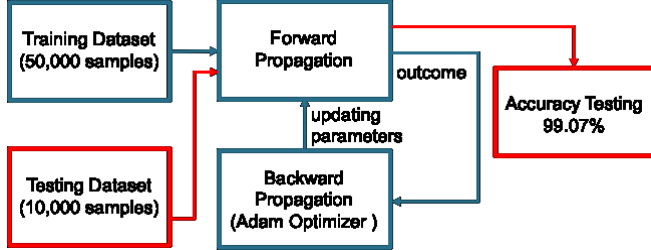


Fig. 11. Model training procedure

Notably, the forward propagation path in Fig. 11 is elaborated in Fig. 12, where even a simple testing for MNIST recognition is proved to attain over 99% mAP. The parameters of each stage is listed as follows.

- Convolution: $16 \times 3 \times 3$ filter, padding = 1, stride = 1
- Activation Function: Relu
- Pooling: 2×2 maxpooling
- Fully Connected: (3136, 10)

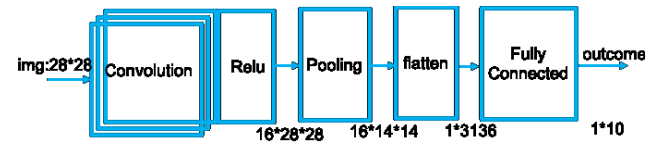


Fig. 12. Forward propagation flow

C. Acoustic pattern recognition experiment

Based on the said AI model and deep learning, this research extracts Mel-spectrograms from acoustic signals as model inputs to classify acoustic signatures from various biological sources, natural noise, and ships. The balanced dataset used for model development consists of 3,000 samples for each category: Biological (3,000), Natural Noise (3,000), and Ships (3,000). The parameters in this recognition experiment, as shown in Fig. 13, is listed as follows.

- Training Dataset: 7,200 samples
- Test Dataset: 1,800 samples
- Epochs: 100
- Training loss: 0.0029
- Validation accuracy: 100%
- Optimizer: Adam optimizer (learning rate = 0.001)
- Loss Function: Cross-Entropy Loss
- Accuracy: 98.67%

Referring to the performance comparison in Table I, this investigation taking advantage of novel AI edge model and training approach to attain the best accuracy in the recognition of multi-category objects.

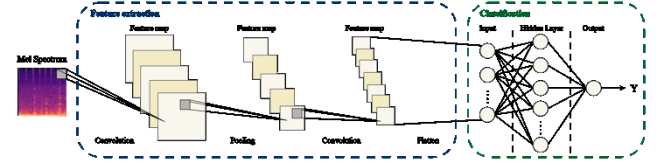


Fig. 13. Acoustic pattern recognition experiment setup

TABLE I. PERFORMANCE COMPARISON

	Features	Accuracy	Dataset
[8] 2025	Ship only	99.8%	ShipsEar
[9] 2024	Ship only	99.3%	ShipsEar
[10] 2025	Mammals only	95%	Self-labeled
[11] 2024	Ship only	95%	ShipsEar
[12] 2026	Propeller only	94.7%	DeepShip
ours	Ships + Naturals + Mammals	98.67%	DOSITS+ Self-labled

IV. CONCLUSION

This research successfully demonstrates a high-accuracy prototyping system for underwater acoustic recognition. By leveraging MFCC feature extraction and a ResBlock-based deep learning model, the system achieved a 98.67% classification accuracy across biological mammals, ship, and environmental noise categories. To overcome hardware constraints, a specialized AI Edge accelerator was designed to integrate convolution steps and utilize int16/int8 quantization, significantly reducing power consumption and hardware cost.

ACKNOWLEDGMENT

This work was partially funded by National Science and Technology Council (NSTC), Taiwan, under grant NSTC 115-2224-E-110-001, 114-2218-E-110-009, 115-2811-E-110-001, and 112-2221-E-110-063-MY3.

REFERENCES

- [1] S. Feng, S. Ma, X. Zhu, and M. Yan, "Artificial Intelligence-Based Underwater Acoustic Target Recognition: A Survey," *Remote Sensing*, no. 16, vol. 3333, 2024.
- [2] J. Xu, Y. Xie, and W. Wang, "Underwater Acoustic Target Recognition based on Smoothness-inducing Regularization and Spectrogram-based Data Augmentation," *Pre-print submission to Elsevier*, Apr. 2024.
- [3] Z. Xiao, L. Yang, P. S. Bithas, I. S. Ansari, X. Li and M.-S. Alouini "Performance analysis of mixed underwater acoustic/optical relaying systems," *IEEE Trans. Wireless Commun.*, vol. 23, no. 9, pp. 11357–11371, Sept. 2024.

- [4] C.-H. Yeh, C.-H. Lin, L.-W. Kang, C.-H. Huang, M.-H. Lin, C.-Y. Chang, and C.-C. Wang, "Lightweight deep neural network for joint learning of underwater object detection and color conversion," *IEEE Trans. on Neural Networks and Learning Systems*, vol. 33, no. 11, pp. 6129-6143, Nov. 2022.
- [5] C.-C. Wang, R. G. B. Sangalang, C.-P. Kuo, H.-C. Wu, Y. Hsu, S.-F. Hsiao, and C.-H. Yeh, "A 40.96-GOPS 196.8-mW digital logic accelerator used in DNN for underwater object recognition," *IEEE Trans. on Circuits and Systems I: Regular Papers*, vol. 69, no. 12, pp. 4860-4871, Dec. 2022.
- [6] C.-C. Wang, S.-H. Luo, H.-C. Wu, R. G. B. Sangalang, C.-P. Jou, H. Hsia, and L.-C. Cho, "A power-efficient 0.5668 TOPS/W digital logic accelerator implemented using 40-nm CMOS process for underwater object recognition usage," *IEEE Access*, vol. 13, no. 2, pp. 28123 - 28134, Feb. 2025.
- [7] DOSITS (Discovery of Sound in the Sea) - Audio Gallery, <https://dosits.org/galleries/audio-gallery/>.
- [8] J. Li, R. Zhang, and H. Wang, "Advanced feature fusion and residual attention networks for robust underwater acoustic target recognition," *IEEE Access*, pp. 14210-14225, Jan. 2025.
- [9] X. Chen, Y. Wu, D. Liu, and L. Zhao, "MGCL-Delta-NCA-ResNet18: A hybrid deep learning model for low-SNR underwater sound classification," *IEEE Sensors Journal*, pp. 8741-8756, Mar. 2024.
- [10] S. R. Thompson, M. Garcia, and A. Martinez, "Automated identification of marine mammal vocalizations using MFCC and multi-scale dilated causal convolutional networks," *Frontiers in Marine Science*, pp. 1102-1118, May 2025.
- [11] Y. Wang, A. Sun, and J. Tan, "A novel lightweight CNN architecture for real-time underwater target recognition based on MFCC feature maps," *Sensors*, pp. 2145-2162, Nov. 2024.
- [12] D. H. Kim, S. J. Lee, and Y. K. Park, "Deep learning-based classification of ship propeller noise using transfer learning on Mel-frequency spectrograms," *Journal of Marine Science and Engineering*, pp. 452-468, Feb. 2026.