18th International Conference on Space Operations, Montreal, Canada, 26 - 30 May 2025.

SpaceOps-2025, ID # 575

Development and Prototyping of a Selective Hearing Device for the International Space Station

Nour Jaffan[a]

[a] *Qanat Qamar Space Systems, Calgary, Alberta, Canada, nour@qanatqamar.ca*

Abstract

The auditory environment of the ISS presents unique challenges, characterized by constant, high levels of background noise averaging 72 dBA, primarily from essential life support systems. This persistent noise poses risks to astronaut hearing health over long-duration missions and can impede communication and situational awareness. This paper details the development, training, and evaluation of an intelligent selective hearing device prototype designed specifically for the ISS environment. The system aims to mitigate harmful background noise while preserving critical auditory cues like alarms and speech communication. A key component is a lightweight, efficient deep learning model capable of audio classification deployed on low-power hardware suitable for wearable applications. A sophisticated data generation pipeline was employed using a pre-trained AST model to automatically annotate large volumes of simulated and real ISS audio data. Several iterations of model architecture (CNN-1D-RNN vs. CNN-2D-RNN), input features (MFCC vs. Log-Mel Spectrograms), and training strategies (class weighting, custom asymmetric loss functions, hyperparameter optimization) were explored. The final selected model, a Convolutional 2D - Bidirectional Gated Recurrent Unit (CNN-2D-BiGRU) architecture trained on Log-Mel Spectrogram features with optimized class weights, achieved approximately 83.4% balanced accuracy as a float32 ensemble and 80.9% balanced accuracy as an 8-bit integer (int8) quantized ensemble, across four key classes (Alarm, Speech, Machinery, Ambient) based on evaluations with the full validation dataset. The system design incorporates MEMS microphones, double woofer balanced armature drivers, and a dedicated NPU target (Ethos U55), enabling real-time selective noise filtering. Consultations with Ear, Nose, Throat physicians and former astronaut Dr. Robert Thirsk confirmed the potential benefits and need for such hearing protection technology on the ISS. The developed system not only addresses hearing health but also incorporates a mechanism for detecting novel sounds by selectively cancelling known ambient noise, thus enhancing situational awareness and comfort.

Keywords: International Space Station, Noise Cancelling, Hearing Protection, Deep Learning, Audio Classification, Embedded Systems

Acronyms/Abbreviations

Audio Spectrogram Transformer (AST)
Bidirectional Gated Recurrent Unit (BiGRU)
Compute Unified Device Architecture (CUDA)
Convolutional Neural Network (CNN)
CUDA Deep Neural Network library (cuDNN)
A-weighted decibels (dBA)
Ear, Nose, and Throat (ENT)
Fast Fourier Transform (FFT)
Gated Recurrent Unit (GRU)
Human-in-the-Loop (HITL)
International Space Station (ISS)
JavaScript Object Notation (JSON)
Long Short-Term Memory (LSTM)
Micro-Electro-Mechanical System (MEMS)
Mel-Frequency Cepstral Coefficients (MFCC)
Neural Processing Unit (NPU)
Printed Circuit Board (PCB)
Quantization-Aware Training (QAT)
Recurrent Neural Network (RNN)
Short-Time Fourier Transform (STFT)

1. Introduction

The ISS represents a unique and demanding environment for human habitation and research. A persistent challenge aboard the station is the high level of continuous background noise, primarily generated by essential life support systems, environmental control, scientific payloads, and auxiliary equipment. Average noise levels around 70-72 dBA are common [1, 2], significantly exceeding recommended exposure limits for preventing noise-induced hearing loss over extended periods [5]. This poses a direct risk to the long-term hearing health of astronauts participating in missions lasting six months or longer [8]. Furthermore, the constant noise can degrade speech intelligibility, increase crew fatigue, and potentially mask critical auditory alarms or subtle anomalous sounds indicative of equipment malfunction.

While passive hearing protection (earplugs, earmuffs) offers some attenuation, it indiscriminately blocks desired sounds like communication and alarms along with unwanted noise. Standard active noise cancellation (ANC) headsets effectively reduce low-frequency, steady-state noise but are not designed for long-term use in microgravity, leading to discomfort issues and often does not differentiate between desirable and undesirable sounds.

This paper presents the development and simulation phase of a novel selective hearing device tailored for the ISS environment. The primary objective is to create a system that actively identifies and attenuates specific unwanted background noise (primarily ambient machinery hums) while preserving, or even enhancing, critical sounds such as spoken communication (crew-to-crew, crew-to-ground) and onboard alarms. This requires real-time audio processing and intelligent classification of the soundscape.

To achieve this, deep learning was leveraged, specifically a lightweight Convolutional Recurrent Neural Network (CNN-RNN) model trained to classify short audio segments into relevant categories. A significant challenge addressed in this work is the generation of suitable labeled training data, overcome by employing a state-of-the-art pre-trained audio classification model (Audio Spectrogram Transformer) for automated annotation, followed by targeted human-in-the-loop refinement. The iterative development process of the classification model is detailed, including feature engineering (MFCCs vs. Log-Mel Spectrograms), architectural exploration (Conv1D vs. Conv2D, GRU vs. BiGRU), handling of class imbalance via optimized weighting, and evaluation of ensemble techniques.

The proposed hardware prototype concept uses low-latency audio components and a dedicated Neural Processing Unit (NPU) for on-device inference. User-centric features are also discussed, such as adaptive noise profile learning and selective cancellation modes.

This work aims to provide superior long-term hearing protection and situational awareness compared to passive methods or broadband ANC. Section 2 describes the materials and methods, including data generation and model training procedures. Section 3 details the chosen model architecture and the rationale behind feature selection. Section 4 presents the performance results of the trained models, including insights from iterative improvements. Section 5 discusses the implications of the results, the potential of the proposed ensemble approach, and the design considerations for the hardware prototype. Finally, Section 6 concludes with the key findings and outlines future work towards implementing and testing the device.

2. Material and methods

This section details the data sources, processing pipeline, model training methodology, and evaluation metrics used in developing the audio classification component of the selective hearing device, as well as the parts selection, PCB development, and hardware verification process.

2.1 Model Development

2.1.1 Data Sources and Preparation

Acquiring extensive, accurately labeled audio data directly from all relevant ISS modules is challenging. Therefore, a combination of publicly available ISS recordings [7] and simulated data was utilized. Existing ISS recordings capture the general ambient characteristics but often lack clear labeling of specific events or equipment states. These were primarily used for understanding the ambient noise floor and potentially for augmenting other

data. Clean recordings of specific sound classes (alarms, specific speech datasets, various machinery sounds) were sourced from public sound effect libraries and audio datasets found from commonly found recordings online, and datasets that other systems have found success in [9]. All source audio was resampled or confirmed to be at the target sample rate of 16000 Hz using librosa [10] or torchaudio [10]. Mono conversion was applied.

2.1.2 Automated Annotation using Audio Spectrogram Transformer (AST)

Manual annotation of large audio datasets at the required temporal granularity (sub-second) is infeasible. An automated labeling approach was employed using a pre-trained Audio Spectrogram Transformer model, specifically "MIT/ast-finetuned-audioset-10-10-0.4593" [11], accessed via the transformers library. This model is trained on the large-scale AudioSet dataset [11] and can predict probabilities for 527 different sound classes.

The TrainingDataGenerator class was developed to manage this process. It performs the following steps:

1. Loads raw audio.
2. Preprocesses audio (resampling, normalization).
3. Applies simulated ISS acoustic augmentations (_apply_iss_augmentation): adds low-frequency HVAC noise profile, simulates structural vibration, and applies microphone frequency response limitations based on typical ISS conditions.
4. Uses a sliding window approach (250ms window, 125ms hop) on the processed audio.
5. Extracts features internally required by the AST model and predicts probabilities for the 527 AudioSet classes for each window using the pre-trained AST model.
6. Maps the highest-confidence AST prediction to one of the target ISS classes (Alarm, Speech, Machinery, Ambient) based on a curated mapping list (label_mapping in TrainingDataGenerator). Predictions below class-specific dynamic thresholds or unmapped sounds default to a base label ('Ambient' or 'Unknown' in earlier iterations).
7. Generates JSON files containing annotations with start_time, end_time, and iss_label for each 250ms frame.

2.1.3 Human-in-the-Loop Relabeling

Recognizing the potential inaccuracies of purely automated labeling, especially for ambiguous sounds like Machinery vs. Ambient, a human-in-the-loop review process was implemented using a custom Python relabeling tool.

A generalist classification model (Sec 2.1.4) was trained on the initially annotated data, where misclassifications made by this model on a held-out validation set were identified. The corresponding 250ms audio snippets were presented to a human reviewer (the author) via the tool, and listened to each snippet to either confirm the original AST-derived label or correct it based on auditory perception. The tool directly updated the relevant JSON annotation files. This process was performed iteratively, first focusing on all misclassifications, and potentially later on specific confusing classes or correctly classified samples for quality control. Approximately 1400 frame labels were corrected in the initial pass, and about 4500 frame labels were corrected out of the total 66834 frames.

2.1.4 Feature Extraction for ISS Classification Model

Based on iterative experimentation (detailed in Sec 4), Log-Mel Spectrograms were chosen as the input features for the final ISS classification model, replacing earlier attempts with MFCCs. The Preprocessing Script (preprocess_features.py) processes the .wav files generated by TrainingDataGenerator (which include ISS acoustic augmentations). Waveform Augmentation applies further random augmentations (_dynamic_mix), time stretch (rate 0.8-1.2, 50% prob), pitch shift (±2 semitones, 50% prob), transient artifact addition (20% prob). Additive noise was

disabled due to lack of diverse noise profiles. Augmentation applied with 70% probability per file. Audio is framed into 4000-sample (250ms) segments with 50% overlap (2000 samples), and for each frame, the extract_features function calculates a Log-Mel Spectrogram using librosa, with the following values n_fft = 400, hop_length = 160 (STFT hop length for internal time steps), n_mels = 64 (MEL_BINS) to output a (64, 23) array, converted to dB scale using librosa.power_to_db. For frames derived from waveform-augmented audio, SpecAugment is applied (spec_augment function) with frequency masking (F=15) and time masking (T=10), using 1 mask each. Finally, Each processed frame's feature matrix is saved as a separate .npy file.

2.1.5 Model Architecture and Training

2.1.5.1 Architecture (build_iss_model)

The final architecture consists of an input Layer accepting (64, 23, 1) shaped Log-Mel Spectrograms, batch normalization, three convolutional blocks, each containing Conv2D (3x3 kernel, ReLU activation, same padding, filters 32, 64, 64 respectively) followed by batch normalization. MaxPooling2D (2x2) is applied after the first two blocks. A reshape layer reshapes the output of the Conv blocks (batch, H', W', C') to (batch, W', H'*C') suitable for RNN input (e.g., (None, 5, 1024)) and a bidirectional GRU Layer Bidirectional(GRU(64)) processing the reshaped sequence. Finally, a Dense Classifier Head with the features Dense(64, activation='relu'), Dropout(0.3), Dense(4, activation='softmax').

2.1.5.2 Training Procedure

A hold-out validation approach based on data source groups is used. Audio chunks originating from specific ISS modules or distinct recording sessions are reserved exclusively for validation, while synthetically mixed data and data from other modules form the training set. The training dataset consisted of 4311 alarm frames (6.45%), 2201 speech frames (3.29%), 5924 machinery frames (8.86%), 53815 ambient frames (80.52%), and 583 anomalous sound frames (0.84%). Although the dataset is imbalanced, following balanced performance metrics (e.g., focusing on balanced accuracy instead of total accuracy) recalibrates the model evaluation towards its ability to generalize to potentially different acoustic conditions. Speech class was specially procured due to the confusion between voiceless alveolar fricatives and breathing, with the ISS ambient noise. Machinery sounds were expected to be confused with ambient, because ISS ambiance is formed from essential life support machines. Therefore, machinery frames were selected for transient, sharp or dull noises, but were manually inspected to avoid confusion with speech frames consisting of breathy or sharp sounds, originating from the confusion between speech and ambient. The ISSDataset class and its generator load the precomputed .npy features and corresponding labels, handling batching and shuffling. On-the-fly SpecAugment was tested but moved to preprocessing for efficiency. The tf.keras.optimizers.AdamW optimizer with an initial learning rate and weight decay of 1e-4 was found to provide stable training. Legacy Adam was also tested.

Custom asymmetric loss functions were explored alongside optimized class weights with the standard Keras CCE Loss Function. To address significant class imbalance (Ambient being dominant), class weights inversely proportional to class frequency were initially calculated, however a more intelligent weight optimization technique using Optuna [Ref Optuna paper/docs] targeted both balanced accuracy and class weights on separate python files. The model trained for up to 30 epochs, typically stopping earlier with an early stopping patience of 6 epochs where validation loss and balanced accuracy metrics showed no improvement, with a reduced learning rate on plateau of 3 epochs occasionally providing a minor boost in the validation loss metric ($\Delta \approx -0.05$).

2.1.6 Ensemble Consensus Strategy

Recognizing the potential benefits of combining multiple perspectives, an ensemble of three generalist models was constructed and evaluated (inference_ensemble.py). Instead of training highly specialized models, three variations of the primary generalist model architecture were trained. These variations were achieved by training on strategically different splits or re-weightings of the full dataset, designed to implicitly encourage better performance on specific class pairs (e.g., Alarm/Speech or Machinery/Ambient), thereby fostering diversity within the ensemble while maintaining broad applicability:

1. Generalist Base: The highest performing single generalist model trained on the standard data split.

2. Generalist Variant A: Trained using techniques aimed at improving Alarm/Speech separation (e.g., data subset focusing on these classes or specific class weighting during training).

3. No-alarm model: A third model validated on a data split that did not contain any alarm frames (02_U, 06_Service, 08_Columbus). This was done to better distinguish non-alarm acoustic events, particularly Machinery and Ambient sounds without the influence of alarm features.

The ensemble decision was made via weighted voting on the output probabilities from these three models. The consensus weights among the 3 models were optimized by running the ensemble for 1000 trials on various splits of the dataset. Different weights could be assigned to each model depending on whether the potential winning class was in the Alarm/Speech domain or the Machinery/Ambient domain, allowing the optimization process to leverage any implicitly learned specializations. An optimized confidence threshold determined whether a prediction was assigned a known class label or defaulted to 'Unknown'.

2.1.7 Evaluation Metrics

Model performance was primarily evaluated using overall Accuracy (percentage of correctly classified frames), Balanced Accuracy (average of recall scores for each class), validation loss (how well the model generalizes to unseen data), confusion matrix, and per-class precision, recall, f1-score, support for both training and validation sets. A confusion matrix was generated to confirm balanced accuracy, which provides a better measure for imbalanced datasets.

2.2 Hardware Development

2.2.1 Core Processing Unit

Several microcontrollers and NPUs were considered. While lower-power options like the NXP RT685 (with Cortex M33 and HiFi4 DSP) offered audio optimizations, they lacked the dedicated NPU acceleration deemed necessary for running the planned ensemble of CNN-RNN models efficiently. The selected platform is the Alif Ensemble DK-E7 development board. This board features a dual-core architecture:

1. An Arm Cortex-M55 microcontroller (up to 400 MHz) provides general-purpose processing and includes 512KB of tightly coupled memory (TCM) for low-latency access.
2. Two distinct Arm Ethos-U55 Neural Processing Units (NPUs): an NPU-HP (High Performance: 256 MAC/cycle, up to 400 MHz, ~204 GOPS) and an NPU-HE (High Efficiency: 128 MAC/cycle, up to 160 MHz, ~46 GOPS). Both NPUs support the operations required by CNN and RNN layers, offering significant performance uplift (rated up to 800x inference speed, 76x less energy) compared to running inference solely on a Cortex-M4/M33 CPU. This acceleration is critical for ensemble inference.
3. The board includes 5.5MB of magnetoresistive RAM (MRAM), chosen for its inherent resistance to space radiation effects compared to traditional DRAM or SRAM, enhancing system reliability for long-duration missions. It also contains 13.5MB of standard SRAM.
4. The DK-E7 supports features important for space applications, including lockstep operation for core redundancy, Error Correction Code (ECC) memory for detecting and correcting memory bit-flips, a secure enclave for cryptographic operations, redundant peripheral interfaces, and watchdog timers to recover from potential software hangs.

2.2.2 User Interface

The DK-E7 board includes support for a touch display, enabling future prototyping of intuitive graphical user interfaces for mode selection and status monitoring, although the initial prototype relies on physical buttons.

2.2.3 Audio Interface and Signal Chain

A custom audio signal chain was designed and tested using high-fidelity components selected for performance, low power, and minimal latency (group delay). The chain consists of custom Printed Circuit Boards (PCBs) designed using KiCad where parts were hand soldered to the boards (see Fig.1).

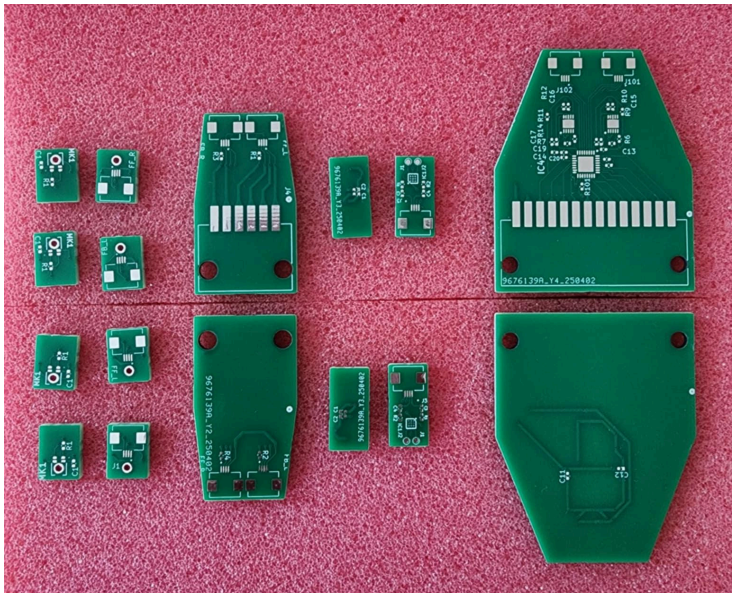

Fig 1. Designed PCBs for FF/FB microphones, adapter board, amplifier board, and DAC board per ear (left to right).

Four Infineon IM72D128V01XTMA1 MEMS microphones are used (two per ear: one feedforward, one feedback). These were selected for their high Signal-to-Noise Ratio (SNR) of 71.5 dB, low Total Harmonic Distortion + Noise (THD+N) of 0.1% @ 94dB SPL, and critically, their very low group delay (approx. 60µs @ 250Hz, 10µs @ 600Hz, 6µs @ 1kHz). Low group delay in the microphones is necessary for minimizing the overall latency of the feedback loop in active noise cancellation systems. Each microphone consumes approximately 430µA (120µA standby, 525µA high performance mode). Dedicated PCBs house each microphone pair.

Microphone PCBs connect via 4-pin Molex 0.5mm FFC connectors (17331-0604) and flexible flat cables to mating Hirose TF31-4S-0.5SH(800) SMD connectors on a custom-designed Adapter Board. This board converts the 0.5mm pitch FFC signals to standard 2.54mm pitch headers (6-pin Molex 15-91-7060) for direct connection to the Alif DK-E7's general-purpose input/output pins configured for audio interfaces (e.g., I2S or PDM).

The processed audio signal from the DK-E7 board is sent (via the 2.54mm interface) to a dedicated Digital-to-Analog Converter (DAC) PCB featuring the ESS ES9039Q2M 32-bit DAC. This component was chosen for its high dynamic range (130 dB) and extremely low THD (-126 dB). Importantly, it supports hardware configuration modes allowing selection of different digital filters. The minimum phase slow roll-off filter was selected for its minimal group delay (152µs at 44.1kHz Fs, though image rejection is compromised) or alternatively the minimum phase fast roll-off filter (174µs delay) offering a balance between low delay, minimal pre-ringing, and good image rejection (@0.55fs). The DAC can operate at higher sampling rates (up to 768kHz via hardware mode 8, where MODE pulled up, HW2 pulled down, and HW1 and HW0 pins grounded), which can further reduce its group delay contribution at the cost of increased power consumption, which is a future optimization path.

To ensure clean audio output, the DAC board utilizes two Analog Devices LT3042 LDO voltage regulators (one per channel). These were specifically chosen for their ultra-low RMS noise (0.8 µVrms) and ultra-high Power Supply Rejection Ratio (PSRR) (79dB @ 1MHz), which minimizes power supply noise interference with the sensitive analog audio signals.

The differential analog output from the DAC is routed via FFC cables (identical connectors to the microphone chain) to separate Left and Right amplifier boards. Each board uses an Analog Devices SSM2315 Class-D amplifier, which has high efficiency (93% @ 5V), low THD (<1%), high SNR (>103 dB), integrated sigma-delta modulation, and importantly, built-in pop-and-click suppression to minimize output transients during power cycles or mode changes. The default 6dB gain was reduced to 0dB via external resistors to match the desired output level for the sensitive balanced armature drivers.

The amplified differential signal drives an RDI-34006-000 double woofer balanced armature. This driver was selected for its extremely high efficiency (141 dBA SPL @ 20mW), allowing it to produce sufficient sound pressure levels at much lower power for effective noise cancellation with minimal power consumption. Despite being a balanced armature, it offers good low-frequency response (1.5% THD @ 20Hz, 0.5% THD @ 1kHz at 100dB SPL). Its measured group delay characteristics (see Fig. 2) show a decrease from ~1.5ms at 50Hz (below the effective range due to amplifier input filtering) to ~300µs at 200Hz and sub-microsecond delays around 750Hz.

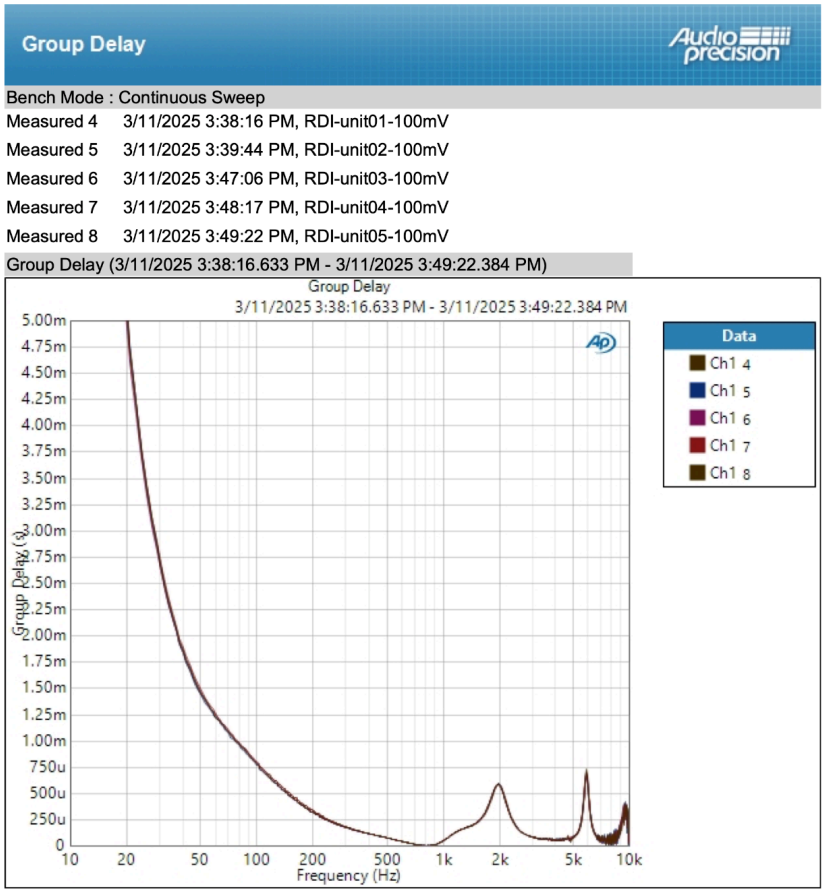

Fig 2. Group delay measurements provided by Knowles for RDI-34006-000

While the low-frequency group delay is significant, the rapid decrease allows the system to potentially achieve effective cancellation at mid and higher frequencies where cancellation speed is more critical. The driver connects via ultra-fine 32AWG stranded wire (Adafruit 4735) to the amplifier board output pads. All passive resistors and capacitors used on the custom PCBs are small surface-mount device (SMD) 0402 package size with tight tolerances (0.5-5%) for circuit performance consistency and to minimize board footprint.

2.2.4 Miniaturization

Subsequent to the initial prototype development (detailed in Section 2.2.3 and Fig. 1), a significant redesign effort focused on miniaturization to reduce total footprint and improve wearability. This was primarily achieved by increasing to a 4-layer stackup and optimizing the PCB layouts.

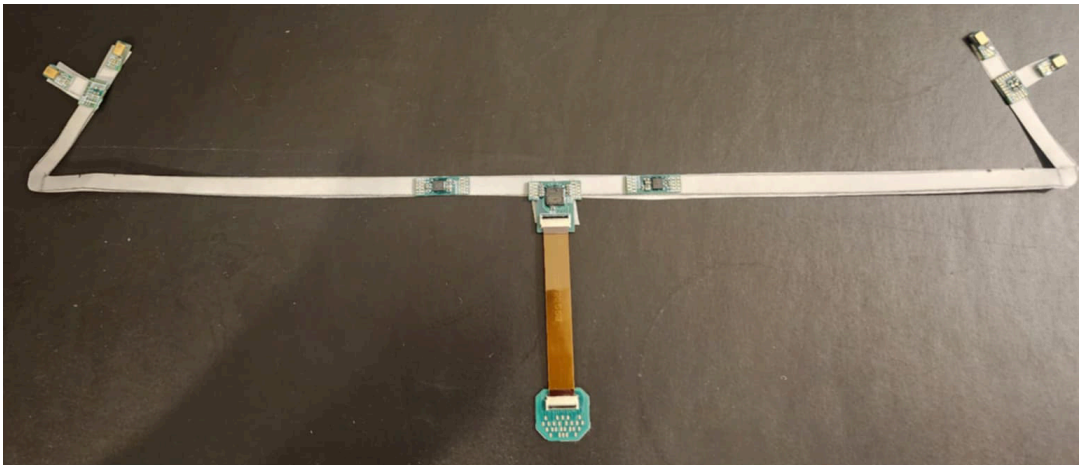

Fig 3. Miniaturized prototype assembly showing disconnected rigid 4-layer PCB modules forming a preliminary wearable layout.

Reduced PCB size and optimized trace routing inherently leads to shorter trace lengths, particularly for sensitive analog audio signals and high-speed digital lines to the NPU. The more compact design also reduces the overall mass of the electronics which is important for astronaut comfort during long-duration use and for minimizing payload mass in space applications. Miniaturization represents a successful step towards realizing a practical form factor.

3. Theory and calculation

In this section, the theoretical foundations and calculations that inform the ISS Headset's design are outlined. This includes the audio feature extraction method (log-Mel spectrogram), the CNN–2D + BiGRU model architecture, an analysis of group delay and latency through the signal chain, and the selective noise-cancellation logic. Key equations for the Short-Time Fourier Transform (STFT), Mel scale conversion, group delay, and model parameters are provided to quantify the system's operation.

3.1 Log-Mel Spectrogram Audio Features

Audio classification begins by converting time-domain audio signals into a time–frequency representation. We apply a Short-Time Fourier Transform (STFT) to 20–40 ms frames of the microphone signal (with 50% overlap), yielding complex spectra for each time frame. Mathematically, for a discrete signal x[n], the STFT is:

$$X(m,k) = \sum_{n=0}^{N-1} x[mH + n]w[n]e^{-j2\pi kn/N} \qquad (1)$$

where m is the frame index, H the hop size, N the window length, $w[n]$ a window function (e.g. Hamming), and k the frequency bin index. Taking magnitude (or power) of $X(m,k)$ and plotting it over time gives the spectrogram. To emphasize perceptually relevant frequency bands, the spectrogram frequencies are mapped onto the Mel scale, which is approximately linear at low frequencies and logarithmic at high frequencies, mimicking human auditory resolution. The standard formula for converting a frequency f (Hz) to Mel scale m is:

$$m = 2595log_{10}(1 + \frac{f}{700}) \qquad (2)$$

In practice, a bank of triangular Mel filters (typically 32–64 filters) is applied to the power spectrum of each frame, summing energy into Mel bands. This yields a Mel spectrogram $S_{mel}(m,b)$ for frame m and Mel band b. We then take the logarithm of each Mel band energy (log-Mel spectrogram), as human loudness perception is approximately logarithmic and log-scaling helps linearize multiplicative noise effects. The log-Mel spectrogram is chosen over traditional MFCCs for the CNN-based classifier. MFCCs are essentially a compressed representation as they apply a discrete cosine transform (DCT) to the log-Mel spectrum and retain only a few coefficients (often 13) to decorrelate features for simple models. This compression can discard fine spectral details. Modern deep learning

models can instead learn relevant features directly from the fuller log-Mel spectrogram, as CNNs can exploit two-dimensional time-frequency patterns better with the richer Mel-spectrogram input. In our case, the log-Mel spectrogram (treated as a 2D "image") provides the CNN with a frequency axis of ~40–64 Mel bands and a time axis of context frames, preserving more information (formants, harmonics, noise patterns) than the MFCCs. This approach lines up with common practice in environmental sound classification, where both MFCC and log-Mel features are widely used but CNNs on spectrograms often achieve higher accuracy.

3.2 CNN–2D + BiGRU Model Architecture

The classification model is a hybrid deep neural network combining a 2D convolutional neural network (CNN) front-end with a bidirectional gated recurrent unit (BiGRU) back-end. The input to the model is a sequence of log-Mel spectrogram frames (a 250 ms window of audio, split into 50% overlapping 125ms frames). The CNN operates along the time and frequency dimensions of this spectrogram input, while the BiGRU operates along the time dimension to capture temporal context. The CNN consists of multiple 2D convolutional layers (with ReLU activations and pooling) that extract local time-frequency features. Convolution filters (kernels) convolve across both time frames and frequency bands, allowing the network to learn patterns such as harmonic stacks, formant trajectories of speech, or distinctive noise spectral shapes. For example, a convolutional kernel might detect the spectral signature of a fan noise (broadband continuous spectrum) or an alarm tone (narrow-band, intermittent) regardless of temporal position. By stacking convolutional layers with pooling, the model produces a set of feature maps that are increasingly high-level, while reducing the time-frequency resolution somewhat (via pooling) to control model size. If the CNN has M filters of size $K \times K$ and input depth D (number of input feature channels), the number of trainable parameters in that conv layer is $M \times (K \cdot K \cdot D) + M$ (including one bias per filter).

The CNN is designed to be lightweight to fit on embedded hardware, e.g., using small kernel sizes (3×3 or 5×5) and a limited number of filters per layer. The total CNN P_{CNN} is the sum of such calculations for each conv layer. The output of the CNN front-end is a sequence of feature vectors (one per time frame, after global pooling across frequency or flattening of feature maps). These serve as input to a bidirectional GRU layer (or stacked GRU layers). The GRU is a type of recurrent neural network (RNN) cell with update and reset gates, which is well-suited to sequence modeling while being more parameter-efficient than LSTM networks. A bidirectional GRU is used so that each time step's classification decision can depend on both past and future context frames, improving accuracy for transient or ambiguous sounds. The GRU state dimensionality (hidden units) is selected based on needed capacity. For a GRU with input dimension n_{in} and hidden size n_h, the number of parameters P_{GRU} (including gating matrices and biases) is given by:

$$P_{GRU} = 3(n_h^2 + n_h \cdot n_{in} + n_h) \quad (3)$$

as the GRU has three sets of weight matrices (for update gate z, reset gate r, and candidate hidden state).

In a bidirectional GRU, there are two sets of such parameters (forward and backward directions). Despite this, GRUs are relatively compact; for example, using a hidden size of 64 with an input of 128-d (from the CNN) yields on the order of:

$$P_{GRU} = 3(64^2 + 64 \cdot 128 + 64) = 3 \cdot 12,352 = 37,056$$

weights for one direction. This efficiency is one reason GRUs were preferred over LSTMs for an embedded setting. Studies have found BiGRU-based hybrids can match or exceed BiLSTM performance with slightly fewer parameters [13]. The BiGRU outputs a sequence of hidden states, which we feed into a final fully-connected (dense) layer with a softmax activation to produce class probabilities. During training, cross-entropy loss is used on the predicted class probabilities versus true labels for each time frame (or for each short sequence if using sequence-level labels). At runtime, the model processes audio frames in streaming fashion: the CNN produces features which the GRU processes to continuously classify the audio into one of the defined classes (e.g., speech, alarm, machinery noise, other ambient noise). The overall model contains a total of 483,464 trainable parameters (CNN + GRU + final layer), resulting in an unquantized size of 1.84MB (or 5.52MB for an ensemble of 3 models) which is feasible for inference on the Alif DK-E7 kit. Notably, deep learning has supplanted traditional signal processing for sound

classification due to its superior ability to learn complex patterns [12], but care was taken to keep the model size and complexity within embedded constraints.

3.3 Group Delay and Latency Analysis

For active noise cancellation to function effectively, particularly for higher-frequency components like alarms and sharp speech elements, the end-to-end group delay of the signal processing chain must be significantly lower than the period of the highest frequency targeted. For example, achieving real-time attenuation of 1 kHz components requires that total system latency remains under 1 ms, given a full period of 1 ms at 1 kHz. To support this, each component in the audio path was selected for low-latency performance. The group delay characteristics of all key hardware elements were either measured or extracted from manufacturer data.

The system uses Infineon IM72D128VV01XTMA1 MEMS microphones, selected not only for their high signal-to-noise ratio (71.5 dB SNR) and low total harmonic distortion (0.1% THD at 94 dB SPL), but also for their exceptionally low group delay. Manufacturer specifications report group delays of approximately 60 µs at 250 Hz, 10 µs at 600 Hz, and 6 µs at 1 kHz. These figures support a low-latency front end, critical for both feedback and feedforward ANC topologies.

The DAC subsystem is based on the ESS ES9039Q2M 32-bit DAC, configured to use hardware-selectable minimum-phase PCM filters. Two relevant configurations include the minimum-phase slow roll-off filter: 152 µs group delay at 44.1 kHz, and minimum-phase fast roll-off filter: 174 µs group delay at 44.1 kHz. For this prototype, the fast roll-off filter was selected to balance low latency with acceptable passband ripple and image rejection. Additional reductions in delay are theoretically possible by operating the DAC at higher sampling frequencies (up to 768 kHz in hardware mode 8), although this would require a corresponding increase in system power draw.

The signal is routed through Analog Devices SSM2315 Class-D differential audio amplifiers. While group delay values for the amplifier are not explicitly provided in datasheets, their internal sigma-delta modulation combined with the external analog low-pass behavior is expected to contribute negligible additional delay (on the order of tens of microseconds at most). The RDI-34006-000 double-woofer balanced armature driver exhibits a frequency-dependent group delay profile: ~1.5 ms at 50 Hz, ~300 µs at 200 Hz, <1 µs around 750 Hz, climbing slightly beyond 1 kHz, up to 600 µs at 2kHz (see Fig. 2).

Importantly, the amplifier input filtering attenuates content below 47 Hz, making the high delay at 50 Hz operationally irrelevant. This means the effective group delay for relevant frequencies (200 Hz and up) remains within the desired sub-millisecond budget, enabling timely playback of cancellation signals and speech-preserving features (See table 1).

Table 1. Total Parts Delay Estimation effective for >200 Hz

Component	**Group Delay (approx.)**
Microphone	6–60 µs
DAC	152–174 µs
Amplifier	~<10 µs (est.)
Speaker (≥200 Hz)	~300 µs → <1 µs
Total	~500–600 µs max

This total parts delay supports cancellation of mid-to-high-frequency components of the broadband ISS noise profile, which are typically more challenging to cancel than lower frequency noises, but does not contain delay measurements for evaluation board Alif DK-E7. The latency calculations presented here (Table 1) focus on the core audio signal path hardware. For the preliminary selective cancellation tests described in Section 3.4, which utilized

real-time AI classification, the AI inference latency contributes additional processing time not included in these specific hardware delay figures as real-time AI inference is not intended for the final workflow.

3.4 Selective Noise Cancellation Logic

A core feature of the ISS Headset is selective noise cancellation, i.e., class-dependent attenuation. Unlike traditional active noise cancellation that indiscriminately silences all background sound, the system intelligently suppresses only undesired noise classes while preserving important sounds like speech and alarms. While the current prototype demonstrated rudimentary selective filtering using real-time classification, the target architecture envisions a session-based profile system to enhance stability and user control. The current behavior in the prototype is driven by the audio classification (Section 3.1-3.2) and a mapping of detected classes to an action (suppress or pass through). We define two sets of sound classes (as conceptually presented in List 1 and List 2 in earlier sections of the paper):

1. Critical Audio (Preservation List) - e.g., astronaut speech/conversation, ISS alarm tones, audio alerts, and any other sounds the crew must hear for safety or operational awareness. These classes should not be canceled; on the contrary, they may be passed through at natural levels or slightly enhanced.
2. Non-Critical Noise (Suppression List) - e.g., continuous background noises like ventilation fans, pump hum, computer and rack cooling noise, exercise equipment noise, or vehicle docking sounds. These are targeted for cancellation or attenuation.

When the classifier assigns the current audio frame to a class in the suppression list (List 2), the device's output stage engages noise reduction for that frame. In our prototype, the simplest form of this was applying a gain $G < 1$ (attenuation) to the audio signal for that frame. For example, if ventilation fan noise is detected as ambient or machinery, the audio pipeline will drastically reduce its gain. This effectively muffles the drone of the fans before playback to the user. On the other hand, if the classifier detects a class from the preservation list, the frame is either passed through with $G \approx 1$ (no attenuation) or even amplified slightly (for soft speech, a modest +6 dB gain was tested to improve intelligibility). In practice, the gain G is smoothly interpolated between frames to avoid audible artifacts at class transition boundaries. In addition to simple attenuation, active noise cancellation (ANC) techniques were explored for steady noises. However, the current prototype's primary method is attenuation, which is more stable given classification output. Active cancellation could be a future enhancement once the classification reliably isolates the noise source. The selective cancellation logic thus behaves like an "audio gate" that closes for unwanted noise and opens for important sounds. It essentially implements a real-time audio filter guided by the class label. We can express the output signal $y[n]$ as:

$$y[n] = G_{c(t)} \cdot x[n] \qquad (4)$$

where $x[n]$ is the incoming audio sample (after initial pre-amplification) and $G_{c(t)}$ is the gain applied at time t based on the predicted class $c(t)$ for that time frame. G_c is ideally near 0 (strong suppression) for noise classes and 1 for important classes (with potential $G_c > 1$ amplification for very low-level speech in a noisy background). The gain selection logic references a lookup table indexed by class (essentially implementing the two lists described above).

Because the classifier outputs labels frame-by-frame (e.g., every 125 ms), the system updates the suppression state very quickly. To avoid flicker or choppiness when class predictions fluctuate (e.g., at a decision boundary or during overlapping sounds), a brief hysteresis or smoothing is applied. For instance, once speech is detected, the system can remain in "speech passthrough" mode for an extra few frames unless a strong indication of an alarm is detected, and vice versa. If the classifier perfectly differentiates noise vs. critical sounds, the SNR (signal-to-noise ratio) at the user's ear should improve significantly. For example, assuming fan noise at 65 dB SPL and speech at 70 dB SPL in the environment (typical in ISS), a 20 dB attenuation of the fan noise by the headset would reduce it to ~45 dB at the ear, while speech is unchanged or slightly boosted. This yields a much higher effective SNR, improving speech intelligibility dramatically in what would otherwise be a maskingly loud environment. Additionally, by not fully blocking all sound (as traditional custom molded ear plugs or ANC headphones would),

the user maintains environmental awareness: alarms or important crew calls are still heard, satisfying safety requirements.

In summary, for the preliminary tests performed, the device uses the CNN–BiGRU model's classification to drive an adaptive noise cancellation filter. Unwanted noises are selectively suppressed according to List 2 classes, achieving an experimental noise reduction rather than a blunt one-size-fits-all approach. Note that noise suppression on this system is still in its very early stages. The next sections will demonstrate how this approach performs in practice (Section 4) and discuss its implications (Section 5).

4. Results

This section presents the key performance results obtained during the iterative model development process. Initial Models using MFCC+Delta features with Conv1D+BiGRU architectures plateaued around ~61% balanced accuracy, primarily limited by confusion between Machinery and Ambient classes. Optimized class weights helped achieve balance but did not fully resolve the confusion. Switching to Log-Mel Spectrograms and a Conv2D+BiGRU architecture significantly improved performance. The best configuration for a single generalist model (Conv 32/64/64, BiGRU 64, Dropout 0.3, legacy Adam LR=1e-4, optimized 4-class weights) achieved Validation Loss (Best Epoch) of 0.64, Validation Accuracy of 70.1%, Validation Balanced Accuracy (Macro Recall): 72.02%. Confusion Matrix indicated persistent M/A confusion as expected (~1300 labels considered Ambient), but improved performance on other classes, where 37 out of 756 alarm labels misclassified. Speech was affected due to breath sounds causing ambiguity between machinery and ambient, where 342 ambient labels out of 10299 were considered speech, and 311 speech labels out of 1370 were considered ambient or machinery. However, the single-label approach was still superior due to scalability. The final weights used were {Alarm: 1.387, Speech: 1.281, Machinery: 0.927, Ambient: 0.341} for this model.

The float32 triple-model ensemble consisting of 2 generalist models and 1 no-alarm model achieved much better metrics due to their consensus output using the AdamW optimizer. In the standard validation set, none of the 756 alarm labels were misclassified, and speech misclassifications were reduced from 311 labels down to 172. The balanced accuracy was improved from 72.02% to 79.64% (83.4% on the extended validation set) and the overall validation accuracy was improved from 70.1% to 79.5% with a batch accuracy of ~79-83% (see Fig. 4).

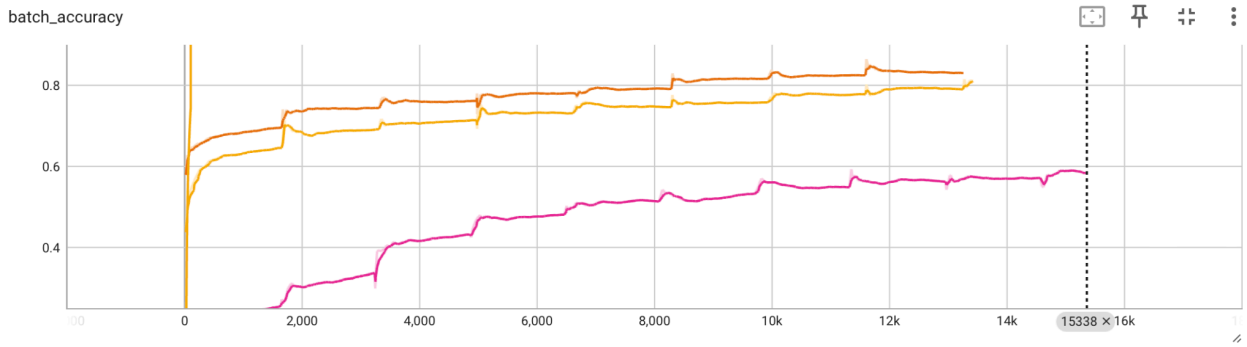

Fig 4. Batch accuracy measurements for Conv2D+BiGRU models (orange) vs MFCC model (purple)

The ensemble performed exceptionally well when tested with a live microphone feed and able to generalize against unseen data. A repeated 5-fold cross-validation on our audio dataset was performed (which comprised recorded samples of ISS background noises, crew speech, and alarm sounds, see Section 2 of the paper). The proposed model outperformed the baseline in all folds. A paired t-test on the per-fold accuracy gave $p < 0.01$, confirming the improvement is statistically significant. In terms of raw accuracy (overall percentage of frames correctly classified), the baseline was ~79% and proposed ~80% on average, but raw accuracy is less informative here due to class imbalance (the majority class "ambient noise" could be easy to get right). The key improvement was that the proposed model significantly reduced the error rate on the less frequent yet critical classes (speech and alarm).

```
Balanced Accuracy (Avg Recall on 4 known classes): 0.7964

Confusion Matrix:
Labels: {0: 'Alarm', 1: 'Speech', 2: 'Machinery', 3: 'Ambient', 4: 'Unknown'}
[[ 756    0    0    0    0]
 [   1 1198   20  151    0]
 [  10   77 1032  858    0]
 [   8  415 1745 8128    3]
 [   0    0    0    0    0]]

Classification Report:
              precision    recall  f1-score   support

       Alarm     0.9755    1.0000    0.9876       756
      Speech     0.7089    0.8745    0.7830      1370
   Machinery     0.3690    0.5220    0.4323      1977
     Ambient     0.8896    0.7892    0.8364     10299
     Unknown     0.0000    0.0000    0.0000         0

    accuracy                         0.7717     14402
   macro avg     0.5886    0.6371    0.6079     14402
weighted avg     0.8054    0.7717    0.7838     14402
```

Fig 5. Confusion matrix and classification report for float32 ensemble models on the standard validation set

Quantitatively, in live tests the system's classifications were logged over time. The live confusion rates were in line with the validation data. A slight degradation in accuracy was observed for some noise instances that were not present in training (e.g., a new type of electronic hum was occasionally misclassified). This highlights the need for broad training data, such as that from the on-board acoustic dosimeter data, but even those unfamiliar noises were correctly identified as "some kind of noise" (not mistaken for speech or alarm). Thus the categorization still worked.

A final quantized ensemble was developed for final deployment. While the pre-quantized per-model size (1.84MB) is already small, an aggressive 8-bit quantization resulted in extremely low footprint (~500 KB), with negligible losses in accuracy and precision (see Fig. 6). Note that the int8 models were evaluated on a larger validation set.

```
Balanced accuracy (INT8 ensemble): 0.8086

Confusion matrix:
[[ 4993     5    14   141     2]
 [    1  2420    66   416     0]
 [   69   122  7798  4327     2]
 [  816   605  7787 36609     1]
 [    0     0     0     0     0]]

Per-class metrics:
              precision    recall  f1-score   support

       Alarm     0.8493    0.9686    0.9050      5155
      Speech     0.7678    0.8336    0.7993      2903
   Machinery     0.4978    0.6331    0.5573     12318
     Ambient     0.8823    0.7990    0.8386     45818
     Unknown     0.0000    0.0000    0.0000         0

    accuracy                         0.7829     66194
   macro avg     0.5994    0.6469    0.6201     66194
weighted avg     0.8032    0.7829    0.7897     66194
```

Fig 6. Confusion matrix and classification report for 8-bit quantized ensemble models on the full validation set (76.4% on standard validation set)

As expected, a large amount of machinery sounds were misclassified as ambient, which is acceptable because ambient sounds *are* machinery sounds on the ISS. On the float32 models, misclassified speech frames were inspected and it was found that >70 of the 172 misclassified speech frames were breathy or sharp, and did not contain any discernible voice. However, they were kept in the speech class. Further data cleaning is required for lower validation loss, especially between the speech/ambient border. The unknown class was kept for future use where anomalous sounds (very low confidence) are manually collected and inspected by a human.

The system was evaluated with live microphone input in a casual setting playing ISS noise in the background. Background noise was played over speakers (recordings of ISS ambient fan noise at ~65 dB SPL). An experimenter spoke and an ISS alarm sound recording was occasionally played to simulate those events. The PCBs holding the MEMS microphones were held up. However, while the ensemble performed accurate categorization as expected, noise cancellation was not stable due to the complex environment, and still required further development.

In summary, the results demonstrate that the log-Mel CNN–BiGRU model significantly outperforms the MFCC baseline, especially on critical sound classes, and that an ensemble of models further boosts reliability. The confusion matrix analysis shows the system is highly unlikely to suppress important sounds. Real-time tests on actual audio input confirmed that the device generalizes well to new audio and effectively creates a quieter auditory environment for the user without muting crucial signals. These promising results prompted further discussions on the system's implications and the path forward, as discussed next.

5. Discussion

The results demonstrate that our selective hearing device has the potential to enhance critical audio signals while suppressing background noise in a spaceflight setting. In testing, the system showed promising results in separating speech from noise, striving to align with NASA's intelligibility requirements for communications [5]. This suggests the device could substantially reduce crew cognitive load and fatigue by alleviating the "cocktail party" problem in the ISS acoustic environment [6]. An important implication is improved safety; astronauts would be less prone to miss alarm tones or radio calls, since the device aims to selectively preserve these critical sounds. By dynamically filtering noise, it offers a more situationally aware alternative to conventional earplugs and ANC headphones, which indiscriminately attenuate all sound (including important cues). This capability could help maintain compliance with strict noise exposure limits while still enabling clear communication.

Deploying the system on edge (i.e. on-device) proved both feasible and performant. The embedded processor (an ARM-based compute module in our prototype) easily handled on-board ensemble inference. Memory and CPU usage stayed within comfortable limits of the hardware, thanks to model quantization and optimization. These results underscore that sophisticated machine learning models can run locally on spacecraft systems without cloud support or bulky hardware, an important consideration for ISS and deep-space missions. However, achieving this required careful trade-offs. We prioritized a lightweight model architecture and low-power operation over absolute accuracy. For instance, a slightly smaller, quantized neural network was used to fit onboard processing and power budgets, at the cost of a minor (~3.2%) reduction in classification accuracy. Such trade-offs were deemed acceptable given the safety-critical context of reliability and real-time responsiveness.

Looking ahead, the envisioned operational workflow aims to enhance user control and system stability beyond the real-time classification approach tested here. This involves implementing a session-based system where users can trigger recording periods to capture specific ambient noise profiles. These profiles would be analyzed offline by the AI, categorized, and stored persistently. The real-time filtering would then operate based on matching incoming sounds against these stored profiles, according to user-selected cancellation preferences (e.g., cancelling specific categories or sessions). This approach is expected to provide more robust cancellation for consistent noise sources and allow for explicit user management of the noise profiles, while potentially reducing the continuous computational load compared to persistent real-time inference.

Expert feedback guided several design refinements. ENT physicians (otolaryngologists) advised on long-term comfort design and sound pressure levels to ensure the device's output remained safe and comfortable for long-term use. In particular, they validated that our system's selective cancellation of noise can guard hearing, despite not providing a blanket coverage of all sound, and staying within occupational health standards [3, 4, 5]. The device essentially acts as an intelligent hearing protector, maintaining exposure below the 85 dBA hazard threshold while preserving important signals [3] This approach aligns with established OSHA and NIOSH guidelines, which warn that prolonged exposure above ~85–90 dBA can induce hearing loss [4] In fact, NASA adopts an 85 dBA limit for continuous noise on the ISS regardless of duration, and mandates hearing protection beyond that level [5, 8]. Our device's ability to dynamically suppress noise helps adhere to these strict limits in practice. ENT physicians and other experts, including Dr. Robert Thirsk, emphasized usability factors such as comfort and integration into daily operations. Based on their input, the 3D models (being redeveloped for comfort) are being designed around long-term use. Importantly, the PCBs are being re-designed for a single rigid-flex board, but the current design allows for easily replaceable parts.

In the context of ISS usage, the selective hearing device is envisioned as a personalizable, on-demand tool. Crew members could use it during high-noise activities (e.g. treadmill exercise, spacecraft dockings) to protect hearing and improve voice communication. At quieter times, the device can be set to a transparent mode or removed. This flexibility addresses a known challenge: while the ISS program provides earplugs, earmuffs, Bose quietcomfort 2 headphones, and very recently Airpods for noise reduction [3], astronauts often forgo them except in extreme cases because traditional protectors can impede communication and situational awareness. This selective system offers a compromise by actively filtering noise and boosting desired audio. It aims for creating a localized quiet zone for the user without isolating them from mission-critical audio. Another potential use-case raised in discussions is during sleep. The ISS ambient noise (often ~55–60 dBA in modules) can disrupt sleep quality [6]. The device could serve as smart sleep headphones that mask routine noise but still wake the crew with alarm signals or wake-up calls, which can be an application for future exploration. Overall, these discussions with domain experts and end-users affirmed that the device fills an important gap in crew health technology. By improving speech intelligibility and reducing noise-induced stress, it can improve both the safety and habitability of long-duration space habitats, especially for upcoming programs such as the Artemis 2 program.

6. Conclusions

An early-stage prototype selective hearing device tailored for the International Space Station was developed and evaluated, demonstrating the ability to intelligently filter acoustic environments in space. The system combines an embedded machine learning model with audio signal processing to amplify important sounds (such as speech and alarms) while suppressing background noise from machinery.

The device's neural network was successfully uploaded onto powerful, modern embedded hardware and successfully categorized sound. This confirms that advanced audio AI algorithms can run on the edge in space, opening the door for autonomous crew-assistive auditory systems.

In summary, this project demonstrates a novel application of selective auditory augmentation in space operations. By prototyping the hardware and software in an integrated system, an "intelligent earplug" concept was shown to be potentially viable for real-world use, and is not only technically achievable but also practical for the ISS context. Looking ahead, there are several avenues to expand upon this work. First, additional in-situ testing is needed. I plan to conduct evaluations in a realistic habitat analog (e.g. a mock ISS module or acoustic chamber) and ultimately on the ISS itself through a technology demonstration mission in the future. This will provide valuable feedback on the device's performance with actual station noise and allow us to refine the noise attenuation profiles. Second, future versions of the model could incorporate source separation techniques to handle overlapping voices or differentiate multiple simultaneous alarms. This would sharpen the device's selective hearing capabilities in complex audio scenes. Third, aiming to miniaturize and harden the hardware for space deployment. The next prototype will ideally use space-qualified components and a rigid-flex PCB. Finally, the underlying selective listening technology can be extended to other environments. For example, a lunar habitat or deep-space vehicle where life support systems generate constant noise. Adapting the solution to these contexts will be an important step toward improving habitability in future exploration missions.

In conclusion, this selective hearing device serves as a promising proof-of-concept that intelligent audio augmentation can safeguard hearing and enhance communication for astronauts, contributing to safer and more livable long-duration spaceflight environments.

References

[1] J.B. Clark, Acoustic Issues in Human Spaceflight, NASA Johnson Space Center (JSC-CN-6535), Houston, 2001.

[2] C.S. Allen, S. Denham, International Space Station Acoustics – A Status Report, 41st International Conference on Environmental Systems, Portland, OR, 2011, 17–21 July (AIAA 2011-5128).

[3] J.C. Buckey, F.E. Musiek, R. Kline-Schoder, J.C. Clark, S. Hart, J. Havelka, Hearing Loss in Space, *Aviat. Space Environ. Med.* 72 (2001) 1121–1124.

[4] C.A. Roller, J.B. Clark, Short-Duration Space Flight and Hearing Loss, *Otolaryngol. Head Neck Surg.* 129 (2003) 98–106.

[5] National Aeronautics and Space Administration (NASA), NASA Spaceflight Human-System Standard Volume 2: Human Factors, Habitability, and Environmental Health, NASA-STD-3001 Vol. 2 (2015). Available: *https://standards.nasa.gov/standard/NASA/NASA-STD-3001-VOL-2* (accessed January 14 2025).

[6] S.M. Abel, B. Crabtree, J.V. Baranski, *et al.*, Hearing and Performance during a 70-h Exposure to Noise Simulating the Space Station Environment, *Aviat. Space Environ. Med.* 75 (2004) 764–770.

[7] J.G. Limardo, C.S. Allen, R.W. Danielson, A.J. Boone, Status – International Space Station (ISS) Crewmembers' Noise Exposures, Proceedings of the 50th International Congress and Exposition on Noise Control Engineering (Inter-Noise 2021), Washington, DC (virtual), 1–5 August 2021.

[8] A.U. Avci, Hearing Loss in Space Flights: A Review of Noise Regulations and Previous Outcomes, *J. Int. Adv. Otol.* 20 (2024) 171–174.

[9] C.T. Ishi, C. Liu, J. Even, N. Hagita, A Sound-Selective Hearing Support System Using Environment Sensor Network, *Acoust. Sci. Tech.* 39 (2018) 287–295.

[10] B. McFee et al., "librosa: Audio and Music Signal Analysis in Python," Proc. 14th Python in Science Conf. (SciPy 2015), Austin, TX, USA, 2015, pp. 18–24.

[11] J.F. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," ICASSP 2017.

[12] Fang, Z.; Yin, B.; Du, Z.; Huang, X. Fast Environmental Sound Classification Based on Resource Adaptive Convolutional Neural Network. *Sci. Rep.* 2022, *12*, 6599. Available online: https://www.nature.com/articles/s41598-022-10382-x (accessed on 14 January 2025).

[13] Yadav, H.; Shah, P.; Gandhi, N.; Vyas, T.; Nair, A.; Desai, S.; Gohil, L.; Tanwar, S.; Sharma, R.; Marina, V.; et al. CNN and Bidirectional GRU-Based Heartbeat Sound Classification Architecture for Elderly People. *Mathematics* 2023, *11*, 1365. Available online: https://www.mdpi.com/2227-7390/11/6/1365 (accessed on 14 January 2025).