Parameter-Efficient Finetuning of ASR Models for Maritime Radio Communications in Ireland

Abstract

The need for manual logging of critical Very High Frequency (VHF) radio communications by the Irish Coast Guard (IRCG) puts an additional cognitive load on watch officers, compromising their situational awareness. While Automatic Speech Recognition (ASR) offers a solution, generic foundational models underperform in the maritime domain due to acoustic mismatches and a scarcity of domain-specific training data. The goal of this work is to address this gap by developing a data preparation pipeline and evaluating different parameter-efficient finetuning (PEFT) methods in a low-resource scenario. A 2-hour labelled dataset was created and used to train the OpenAI whisper-large model. The effectiveness of two PEFT techniques, namely, Encoder Freezing and Low-Rank Adaptation (LoRA) was evaluated. These finetuned models outperformed the original model's baseline 62.8% Word Error Rate (WER). The constrained LoRA configuration proved most effective, achieving a 28.12% WER, a 52.2% relative reduction in error.

Keywords: Automatic Speech Recognition, Maritime VHF communication, Encoder Freezing, Low-Rank Adaptation, Dataset Preparation, Search and Rescue

1 Introduction

The maritime domain is vital to global commerce, underpinning the transport of goods and services worldwide. Maritime industries employ people across diverse sectors, including fishing, offshore renewable energy, communications infrastructure, and naval operations. In addition, many individuals go to sea for leisure, yet the marine environment remains inherently hazardous for all mariners [1]. According to the European Maritime Safety Agency's annual overview of marine casualties and incidents [2], there were 2,896 reported incidents involving ships within the territorial seas or internal waters of EU Member States in 2023. The report notes that 11 ships were lost, 809 people sustained injuries, and 29 people were lost. In 2024 alone, the Irish Coast Guard (IRCG) responded to 2,554 incidents, assisting thousands of individuals and saving over 500 lives[3].

The IRCG as a division of the Department of Transport in Ireland has the responsibility of conducting Search and Research (SAR) operations within the Irish SAR Region, which is 140,600 square kilometres in size, The IRCG has responsibility for shipping incidents that may impact the environment within Irelands Exclusive Economic Zone (EEZ) which is ten times larger than the land mass of Ireland at 880,000 square kilometres[4]. Coast Guard SAR operations, coordinated through Maritime Rescue Coordination Centres (MRCCs), provide critical assistance to vessels and crews when incidents occur. The rapid deployment of SAR assets significantly improves the chances of survival in maritime emergencies [5].

Effective VHF (Very High Frequency) radio communication is critical for coordinating SAR operations. During incidents, watch officers face a high cognitive load, as they simultaneously monitor multiple radio channels and manually log critical information. Incomplete or delayed logs can compromise situational awareness and operational decision-making [6].

Automatic Speech Recognition (ASR) technology is now commonly used in daily life through voice assistants like Apple's Siri and Amazon's Alexa, but its application in high-stakes environments like emergency response and maritime safety remains a significant challenge. An ASR system could automate the transcription of VHF communications, allowing officers to focus on coordination rather than documentation. However, the maritime domain presents formidable challenges for standard ASR models. VHF transmissions suffer from signal degradation, heavy background noise and weather, and a narrow bandwidth that limits audio quality[6, 7]. Furthermore, the domain is characterised by special vocabulary, including unique vessel and location names, as well as a wide diversity of regional and non-native accents from local and international crews [8].

Our initial evaluation of the state-of-the-art foundational models, including the top 4 models in the HuggingFace Open ASR leaderboard[1] and Whisper-large variants and the commercial API gpt-4o-transcribe revealed zero-shot WER over 40%, rendering them unsuitable for practical use (Section 5.2). This performance deficit, caused by a mismatch between generalist training data and the specific maritime domain, defines the core problem addressed in this work by first developing a robust pipeline to process raw, noisy VHF recordings into a high-quality, domain-specific speech dataset, and systematically investigating Parameter-Efficient Finetuning (PEFT) techniques to adapt pretrained ASR models in a low-resource context, aiming to significantly reduce WER. This research lays the groundwork necessary for creating a deployable ASR tool to enhance maritime safety and reduce cognitive load on IRCG personnel.

[1] HuggingFace Open ASR Leaderboard: https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

2 Related Work

The Global Maritime Distress and Safety System (GMDSS) standard was implemented in 1999 by the International Maritime Organisation (IMO); and it provides procedures for maritime distress and safety communications. The GMDSS relies primarily on voice communication over VHF radio, Medium Frequency (MF), and High Frequency (HF) radio [9, 10] Research shows mariners prefer traditional voice communication, such as VHF radio, when compared to alerting systems such as Digital Selective Calling (DSC) [11, 12].

Application of ASR in maritime communication is an emerging research area. Mokaram and Moore [13] addressed a foundational gap by developing the Sheffield SAR Corpus, one of the first datasets focused on land-based SAR scenarios. Although not maritime-specific, it underscored the value of domain-relevant datasets and the complex dynamics of real-time, human-to-human distress communication. Simulated SAR exercises illustrated that MRCC watch officers, when overwhelmed, often failed to log critical communications in real time, highlighting the need for AI solutions to support situational awareness [6]. The evolution of ASR has shifted from early statistical methods, such as Hidden Markov Models (HMMs), to modern end-to-end neural architectures. Early research on the latter approach largely utilised encoder-decoder architectures such as Listen, Attend and Spell [14]. The advent of the Transformer architecture [15] and self-supervised learning frameworks, such as Wav2Vec 2.0 [16], marked a significant leap in performance. However, the 2022 release of OpenAI's Whisper [17] model trained on a massive and diverse dataset demonstrated robust zero-shot capabilities across various domains, and marked a shift towards transformer-based architectures, which have since gained traction for maritime ASR. The number of studies on VHF speech transcription increased after the release of the Whisper model. Still, these generic models significantly underperform in specialised environments, such as maritime VHF communications due to severe acoustic and lexical mismatches.

Prior research in adapting ASR for the maritime domain reflects this technological progression. Gözalan et al. [18] used hybrid TDNN-LSTM (Time Delayed Neural Network – Long Short-Term Memory) models with separately adapted acoustic and language models. They achieved a 41% WER on 27 hours of real VHF data, highlighting the difficulty of this domain. With the rise of large, pretrained ASR models, focus shifted to finetuning. Nakilcioğlu et al. [19] finetuned a Wav2Vec 2.0 [16] model on 62 hours of English and German maritime data, achieving 31.59% WER. Dat et al [8] achieved a 14.34% WER by finetuning Whisper-large-v3 on over 200 hours of real VHF data, demonstrating the effectiveness of transfer learning when large domain-specific datasets are available. Martius et al. [7] explored using synthetic data to train models when real data is unavailable, reducing WER of Whisper-large-v3 from 44.61% to 35.50%. While finetuning shows significant performance improvements, it is computationally expensive and requires large datasets, which are scarce in specialised domains. This has led to the exploration of PEFT methods. Martius et al. [7] found that with their synthetic data, Encoder Freezing improved WER, while Lor-rank Adaptation performed poorly in their setup, suggesting sensitivity to their hyperparameter choices. In a different approach, Lall & Liu [20] used contextual biasing to inject domain vocabulary at inference time, dramatically reducing the WER of Whisper-medium model from 27.82% to 11.12% on simulated data without altering model weights.

Our review of the literature revealed a clear research gap in systematic evaluation of prominent PEFT methods, such as Encoder Freezing and LoRA on real, noisy VHF data in a low-resource scenario. Furthermore, there is a lack of documented, repeatable pipelines for processing raw maritime audio, hindering reproducibility. This study aims to address these gaps.

3 Methodology

The first phase of this research is focused on creating a high-quality domain-specific dataset from recorded raw IRCG audio. The second phase covers optimising pretrained ASR models with PEFT techniques under low resource conditions.

3.1 Dataset Preparation

The dataset used in this study was derived from an initial corpus of 189 hours of VHF audio recordings sourced from the IRCG, containing only marine radio traffic captured for lawful safety and security purposes[2]. The raw recordings were characterised by extensive periods of silence and ambient noise, necessitating a robust preprocessing pipeline to create a high-quality, labelled dataset suitable for model training and evaluation.

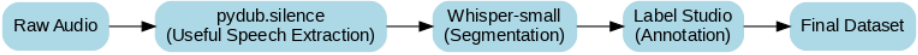

Fig. 1. Dataset Preparation Pipeline

The dataset was prepared through a three-stage process (Fig. 1). For the initial speech extraction, we compared two approaches: an energy-based tool (*pydub.silence* [21]) and a pre-trained neural network (*Silero VAD* [22]). The *pydub.silence* library was selected as it achieved 100% recall rate, which was critical for ensuring no speech segments were missed in this step. The extracted speech audio was then fed into the Whisper-small model to generate initial sentence-level transcripts and corresponding timestamps. To build a high-quality labelled dataset, these machine-generated clips and transcripts were imported into the Label Studio [23]. This annotation platform was chosen because it provides an efficient interface for domain experts to manually review

[2] All VHF audio recordings were collected in full compliance with the Wireless Telegraphy Act and the General Data Protection Regulation (GDPR). No personal identifiers or sensitive information relating to individuals were included, and the data has not been shared publicly.

each transcription alongside the corresponding audio clip, thereby creating the gold-standard dataset.

This pipeline successfully reduced the 189 hours of raw audio to approximately 27 hours of active communication, yielding 25,683 segments. Out of this, 1,530 segments have been manually transcribed during the course of this work, amounting to 2 hours of labelled audio data for further experiments. The dataset was partitioned into training (1 hour), validation (0.5 hours), and test (0.5 hours) sets in a 2:1:1 ratio.

3.2 ASR Model Optimisation

We used OpenAI's Whisper-large as the base model in our fine-tuning experiments. The Whisper-large model was trained on a substantial 680,000 hours of multilingual, weakly-supervised data [[17]. However, we selected it because of its suboptimal zero-shot performance on our target domain relative to other state-of-the-art models (5.2). This characteristic made it the ideal candidate for clearly evaluating the impact of fine-tuning. The two distinct PEFT strategies, namely Encoder Freezing and LoRA were systematically evaluated. The premise of the Encoder Freezing approach is that the Whisper encoder has already learned robust acoustic representations, and the decoder is adapted to map existing features to the target domain's vocabulary. For this method, 4 models were trained using four learning rates: 5×10^{-5}, 1×10^{-5}, 5×10^{-6}, and 1×10^{-6}. This range is informed by similar study by Matrius et al. [7].

The second method, LoRA, works by injecting small, trainable low-rank matrices into the frozen layers of the baseline model. The weight update ΔW is approximated with the product of two matrices B·A, where $A \in R^{d\times r}$ and $B \in R^{r\times d}$ and the rank $r \ll d$. Two sets of LoRA experiments were conducted. The first applied LoRA broadly to the query (q_proj), key (k_proj), value (v_proj), and output (o_proj) projection layers, while the second applied it more sparsely to only the query and value layers. For each configuration, we performed a grid search over a set of hyperparameters to determine the optimal combination. We evaluated learning rates of 1×10^{-3} and 1×10^{-4}, based on the work of Liu et al. [24], in combination with LoRA rank (r) values of 16 and 32, as recommended by established tuning guides[3]. This resulted in four distinct hyperparameter combinations being evaluated for each model.

All experiments were conducted under a unified training framework. The training dataset consisted of 763 audio segments totalling 60 minutes, with a validation set containing 385 segments of 30 minutes, structured in the required JSON format. A linear learning rate scheduler with a 10% warmup phase was employed for all runs, alongside a batch size of 16 and FP16 mixed-precision for computational efficiency. Model performance was evaluated on the validation set every 25 training steps. All experiments were executed on a single NVIDIA A100 GPU with 80GB of VRAM.

[3] LoRA Hyperparameter Guide by Unsloth AI. Available at: https://docs.unsloth.ai/get-started/fine-tuning-llms-guide/lora-hyperparameters-guide

4 Evaluation Metrics

To quantitatively assess performance, two distinct metrics were aligned with the two primary tasks of the methodology.

4.1 F1 Score for Speech Extraction

To evaluate the speech extraction tools, *Silero VAD* and *pydub.silence*, our primary metric was Recall, also known as the True Positive Rate (TPR). For our data preparation pipeline, high Recall is the most critical factor because it measures a tool's ability to find all actual speech segments, ensuring no valuable audio is accidentally discarded. We also considered the False Positive Rate (FPR), which measures the proportion of non-speech segments incorrectly identified as speech; a lower FPR is better. The performance trade-off between these two metrics is visualized on a Receiver Operating Characteristic (ROC) curve, which plots the TPR against FPR. An ideal model would be placed in the top-left corner of this plot, indicating a high Recall with a low FPR.

4.2 Word Error Rate (WER)

The primary metric for evaluating ASR model performance is the WER. WER measures the difference between model-generated transcription (hypothesis) and the ground-truth transcription (reference) by calculating the minimum number of edits required to match them.

$$WER = \frac{S+D+I}{N} \tag{1}$$

Where *S* is the number of substitutions, D is the number of deletions, I is the number of insertions, and *N* is the total number of words in the reference text. A lower WER indicates higher accuracy, with 0 being a perfect transcription. All texts were normalised (i.e., converted to lowercase and punctuation removed) before calculating WER to ensure an accurate comparison.

5 Results and Discussion

Our experiments yielded clear results regarding data preparation, baseline model performance, and the effectiveness of PEFT techniques. All ASR models were evaluated on the unseen 0.5-hour test set.

5.1 Data Preprocessing Performance

The performance of *Silero VAD* and *pydub.silence* was quantified using a 5-hour audio corpus, segmented into 621 manually labelled 30-second chunks. The objective was to select a model with the highest recall to ensure no speech segments were lost during data preprocessing. Hyperparameter tuning was performed for both tools: Silero VAD's speech probability threshold was varied (0.2 to 0.8), while *pydub.silence's* RMS energy *silence_thresh* (dBFS) and *min_silence_len* (ms) were adjusted.

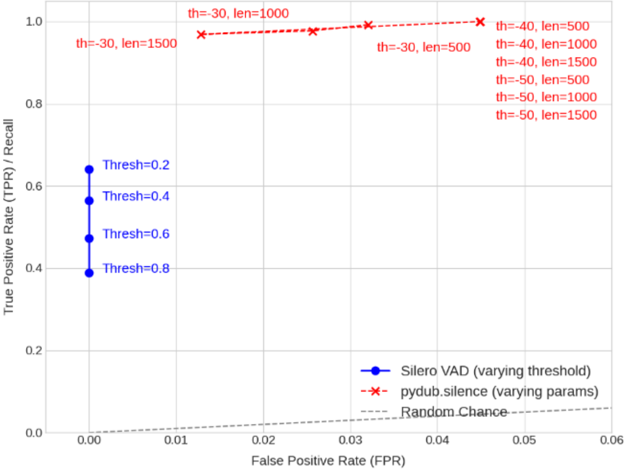

Fig. 2. ROC Curve Comparison of Silero VAD and pydub.silence

The resulting ROC curve (Fig. 2) shows that *Silero VAD* maintains a perfect precision (0% FPR) but suffers from poor recall, which peaked at 64% with a threshold of 0.2. This means 36% of speech-containing segments were incorrectly discarded. In contrast, *pydub.silence* achieved 100% recall with *silence_thresh* settings of -40 dBFS and below, irrespective of the *min_silence_len*. For our data preparation task, the perfect recall of *pydub.silence* is ideal, as it ensures the complete preservation of valuable speech data.

5.2 Zero-Shot Baseline Performance

To quantify the difficulty of the maritime domain and establish a performance baseline, a zero-shot evaluation was conducted on a suite of pre-trained ASR models on the 30-minute test set. The results shown below reveal that all evaluated models exhibit high WER, underscoring the necessity of domain-specific adaptation.

The best-performing model was whisper-large-v3, yet it still achieved a high WER of 39.87%. Other prominent models, including the top-ranked model on the Hugging Face ASR leaderboard (*nvidia-canary-qwen-2.5b*) and OpenAI's API (*gpt-4o-transcribe*), also struggled significantly, posting WERs of 47.76% and 49.71%, respectively.

Notably, the whisper-large model, which was selected as the baseline for our fine-tuning experiments, was among the weaker performers with a WER of 62.80%. The performance of other models, such as *granite-speech-3.3-2b* (73.76%), further highlights the domain gap. These high error rates across a diverse set of powerful, general-purpose

ASR systems confirm that off-the-shelf models are inadequate for reliable transcription in this specialized environment, providing a clear rationale for the fine-tuning methods investigated in this study.

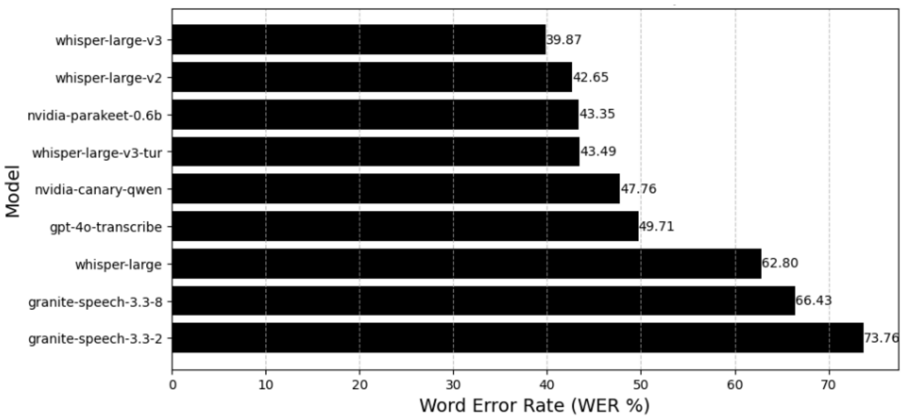

Fig. 3. Zero-shot performance evaluation of general-purpose ASR models

5.3 Model Training Behaviour

The training logs provided crucial insights into the stability and hyperparameter sensitivity of the different finetuning strategies. In the Encoder Freezing experiments, demonstrated a generally stable training process (Fig. 4). However, the model's convergence was highly sensitive to the choice of learning rate. The run using a learning rate of 5×10^{-5} exhibited ideal behaviour. Both the train and eval losses dropped rapidly and stabilised at a value near zero. Correspondingly, the evaluation WER for this run showed a consistent downward trend, indicating effective learning. In contrast, lower learning rates resulted in much slower convergence and higher final error rates, confirming that while this method is robust against divergence, its performance is contingent on proper learning rate selection.

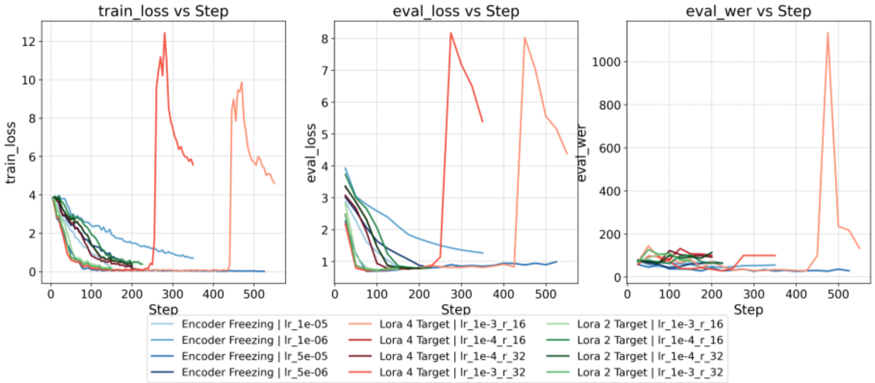

Fig. 4. Whisper-large training logs

The LoRA experiments highlighted a critical sensitivity to the learning rate, leading to training instability in certain configurations. As shown in Fig. 4, the runs configured with a high learning rate of 1×10^{-3} initially began to converge before suddenly diverging. This divergence is visualized as dramatic spikes where the eval WER, train and eval loss explode to extreme values. This indicates that the weight updates became too large, causing the instability in training. This occurred regardless of whether LoRA was applied to two target modules or four target modules. This confirms that the learning rate was the dominant factor driving the model's failure, not the number of adapted modules. In contrast, LoRA runs with a lower learning rate of 1×10^{-4} did not exhibit this divergence and converged successfully.

5.4 PEFT Performance and Comparative Analysis

The finetuning experiments demonstrate the significant impact of PEFT on adapting the *Whisper-large* model to the maritime domain. Whisper-large was selected for its suboptimal zero-shot performance, to evaluate the transformative impact of the PEFT methods. As illustrated in Fig. 5, nearly all PEFT configurations substantially outperformed the zero-shot baseline WER of 62.8%.

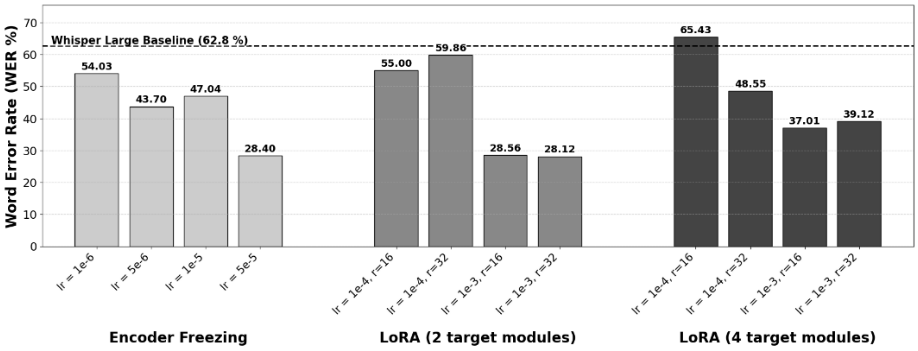

Fig. 5. Performance of finetuned models on 30-minute test set (*lr* = learning rate, *r* = rank)

The Encoder Freezing method proved to be highly effective, achieving its best result of 28.40% WER with a learning rate of 5×10^{-5}, which represents a 54.8% relative reduction in error compared to the baseline. However, this approach was highly sensitive to the learning rate, with other values yielding worse performance. The LoRA experiments showed strong results that were dependent on hyperparameter configuration as well. Applying LoRA to four target modules (*q_proj*, *k_proj, v_proj*, *o_proj*) resulted in a minimum WER of 37.01%. While it is a significant improvement over the baseline, this approach was outperformed by Encoder Freezing and proved more susceptible to instability. Counter-intuitively, the most effective strategy overall was a more constrained application of LoRA to only two target modules (*q_proj* and *v_proj*). This configuration achieved the study's lowest WER of 28.12% with a learning rate of 1×10^{-3} and a rank of 32, slightly surpassing the best performance of the Encoder Freezing method. By limiting the adaptation to only the query (*q_proj*) and value (*v_proj*) projection layers, the model is directed to update the most critical components for ASR adaptation: (a) what acoustic features to focus on (query); and (b) what content to extract (value). The key (*k_proj*) and output (*o_proj*) projection layers adapted in the four-module setup provide greater model capacity. This increased capacity can be beneficial with large datasets. However, in a low-resource scenario, it can allow the model to learn spurious correlations specific to the training data. This harms the model's ability to generalise to unseen data.

These results are comparable within the existing maritime ASR literature. Our final WER of 28.12% represents a significant improvement over the 41% WER reported by Gözalan et al. [18] and the 31.59% WER achieved by Nakilcioğlu et al. [19]. It also surpasses the 35.50% WER obtained by Martius et al. [7] with synthetic data. While Dat et al. [8] achieved a superior 14.34% WER, it is critical to note that their result relied on a massive dataset of over 200 hours. In contrast, out study demonstrates the efficacy of PEFT in a low-resource environment, achieving competitive result with 2 hours of labelled audio. This highlights a computationally efficient path for adapting large pretrained models to specialized domains where data is scarce.

6 Conclusion

This research addressed the challenge of adapting general-purpose ASR models for the challenging domain of maritime VHF communications. A scalable data preparation pipeline was developed that prioritised data integrity, ensuring no loss of valuable speech from the raw recordings. The experiments conducted demonstrate that PEFT is essential and highly effective in low-resource conditions. The finetuned models significantly outperformed their high-error baseline, with a LoRA configuration applied on two modules of the self-attention mechanism of the Whisper-large model, achieving a final WER of 28.12%. This represents a remarkable 55.2% relative reduction in error from the zero-shot baseline.

The findings confirm that domain adaptation is non-negotiable for achieving usable ASR performance in specialized fields. For low-resource conditions, this study shows that multiple PEFT techniques provide a computationally efficient path to state-of-the-art results. Both Encoder Freezing (28.40% WER) and a well-tuned LoRA (28.12% WER) proved to be highly competitive, nearly matching each other in performance. A key finding is that a more constrained LoRA adaptation, targeting only two modules, outperformed a broader four-module approach. This suggests that limiting the model's capacity during fine-tuning can act as a powerful regularizer, preventing overfitting on limited data and leading to better generalisation.

This study provides a strong foundation and a clear, evidence-based methodology for developing a deployable ASR tool that can significantly reduce the cognitive load on IRCG officers and enhance overall maritime safety. Future work will build upon this foundation by focusing on the path to operational deployment. The immediate priority is to expand the annotated dataset, which will enable more robust training and facilitate a comparative study between PEFT and full fine-tuning. Concurrently, contextual biasing will be explored to improve the model's handling of domain-specific terminology such as vessel names and coastal locations. The ultimate objective is to integrate these advancements into a streaming-capable model for real-world evaluation of its latency, robustness, and usability within the IRCG's live operational environment.

References

[1] M. Malyszko, "Fuzzy Logic in Selection of Maritime Search and Rescue Units," *Applied Sciences (Switzerland)*, vol. 12, no. 1, Jan. 2022, doi: 10.3390/app12010021.

[2] "ANNUAL OVERVIEW OF MARINE CASUALTIES AND INCIDENTS 2023 ANNUAL OVERVIEW OF MARINE CASUALTIES AND INCIDENTS 2024 European Maritime Safety Agency," 2024.

[3] Department of Transport, "The Irish Coast Guard responds to 2,554 incidents in 2024," Government of Ireland. Available at: https://www.gov.ie/en/department-of-transport/press-releases/irish-coast-guard-responded-to-2554-incidents-in-2024/.

[4] Marine Institute. Available at: https://www.marine.ie/site-area/news-events/news/map-ireland-bigger-you-think, "The Map of Ireland is Bigger Than You Think."

[5] Z. Paladin, E. Fountoulakis, Z. Luksic, N. Kapidani, D. Ribar, and G. Boustras, "Advanced Mission Critical Communication in Maritime Search and Rescue Actions," in *2023 27th International Conference on Information Technology, IT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/IT57431.2023.10078557.

[6] J. Seger, "Coagency of humans and artificial intelligence in sea rescue environments A closer look at where artificial intelligence can help humans maintain and improve situational awareness in search and rescue operations." [Online]. Available: http://www.ep.liu.se/.

[7] C. Martius, E. Ç. Nakilcioğlu, M. Reimann, and O. John, "Refining maritime Automatic Speech Recognition by leveraging synthetic speech," *Maritime Transport Research*, vol. 7, Dec. 2024, doi: 10.1016/j.martra.2024.100114.

[8] P. Dat, J. M. Madhathil, and T. H. Dat, "Automatic Speech Recognition and Spoken Language Understanding of Maritime Radio Communications: A case study with Singapore data," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2025. doi: 10.1109/ICASSP49660.2025.10888216.

[9] E. Tzannatos, "GMDSS Operability: The Operator-Equipment Interface," *The Journal of Navigation*, vol. 55, no. 1, pp. 75–82, 2002, doi: 10.1017/S037346330100162X.

[10] S. Valčić, A. Škrobonja, L. Maglić, and B. Sviličić, "GMDSS Equipment Usage: Seafarers' Experience," *J Mar Sci Eng*, vol. 9, no. 5, p. 476, 2021, doi: 10.3390/jmse9050476.

[11] A. H. Patterson and P. S. McCarter, "Digital Selective Calling: The Weak Link of the GMDSS," *The Journal of Navigation*, vol. 52, no. 1, pp. 28–41, 1999, doi: 10.1017/S0373463398008133.

[12] L. Chen and J. Liu, "Identification of Shipborne VHF Radio Based on Deep Learning with Feature Extraction," *J Mar Sci Eng*, vol. 12, no. 5, p. 810, 2024, doi: 10.3390/jmse12050810.

[13] S. Mokaram and R. K. Moore, "The Sheffield Search and Rescue corpus," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5840–5844. doi: 10.1109/ICASSP.2017.7953276.

[14] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, Attend and Spell," Aug. 2015, [Online]. Available: http://arxiv.org/abs/1508.01211

[15] A. Vaswani *et al.*, "Attention Is All You Need," 2023.

[16] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2006.11477

[17] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.04356

[18] A. Gözalan *et al.*, "Assisting maritime search and rescue (Sar) personnel with ai-based speech recognition and smart direction finding," *J Mar Sci Eng*, vol. 8, no. 10, pp. 1–13, Oct. 2020, doi: 10.3390/jmse8100818.

[19] E. C. Nakilcioglu, M. Reimann, and O. John, "Adaptation and Optimization of Automatic Speech Recognition (ASR) for the Maritime Domain in the Field of VHF Communication," 2023, *arXiv*. [Online]. Available: http://arxiv.org/abs/2306.00614

[20] V. Lall and Y. Liu, "Contextual Biasing to Improve Domain-specific Custom Vocabulary Audio Transcription without Explicit Fine-Tuning of Whisper Model," Oct. 2024, doi: 10.1109/MLNLP63328.2024.10800265.

[21] J. Robert *et al.*, "pydub," 2011, *GitHub. Available at: https://github.com/jiaaro/pydub*.

[22] Silero Team, "Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier," 2024, *GitHub. Available at: https://github.com/snakers4/silero-vad*.

[23] M. Tkachenko, M. Malyuk, A. Holmanyuk, and N. Liubimov, "Label Studio: Data labeling software," 2020, *HumanSignal. Available at: https://labelstud.io/*.

[24] Y. Liu, X. Yang, and D. Qu, "Exploration of Whisper fine-tuning strategies for low-resource ASR," *EURASIP J Audio Speech Music Process*, vol. 2024, no. 1, Dec. 2024, doi: 10.1186/s13636-024-00349-3.