AI-Driven Conversational Voice Communication for Maritime Autonomous Surface Ships

Sanha Kim
Department of Computer Engineering
Chungnam National University
Daejeon, Republic of Korea
0009-0003-4615-3045

Jaebin Ku
Department of Computer Engineering
Chungnam National University
Daejeon, Republic of Korea
0009-0001-1694-9670

Eunkyu Lee
Autonomous Ship Research Center
Samsung Heavy Industries
Daejeon, Republic of Korea
0009-0000-4903-5288

Umar Zaman
Department of Computer Engineering
Chungnam National University
Daejeon, Republic of Korea
0000-0002-4257-4868

Kyungsup Kim
Department of Computer Engineering
Chungnam National University
Daejeon, Republic of Korea
0000-0002-0166-1439

***Abstract*—Although voice communication plays a crucial role in maritime operations due to its safety and efficiency, Maritime Autonomous Surface Ships (MASS) operating unmanned have not been fully integrated with existing voice communication systems. Therefore, this study proposes an integrated voice communication system for MASS. The proposed system combines state-of-the-art Speech-to-Text (STT) technology, Text-to-Speech (TTS) technology, and Large Language Models (LLM) to enable MASS to autonomously interact via VHF voice communication. In particular, the LLM generates appropriate responses that comply with standardized maritime communication protocols by learning Standard Marine Communication Phrases (SMCP) using Retrieval-Augmented Generation (RAG). The developed system enhances the autonomy of MASS and enables seamless interaction with existing manned ships and control centers. Experimental results demonstrated high speech recognition accuracy and real-time processing performance, confirming its practicality in maritime communication environments.**

***Index Terms*—Large Language Model, LLM, Retrieval-Augmented Generation, RAG, Communication System, MASS**

I. Introduction

As Maritime Autonomous Surface Ships (MASS) advance, many traditional aspects of shipping are becoming autonomous [1]. MASS have the potential to enhance the efficiency of the maritime industry and reduce operational costs through their ability to make decisions and take actions without human intervention. Particularly in areas such as route planning, collision avoidance, and situational awareness, the autonomy of MASS is revolutionizing traditional ship operations [2]. However, despite the development of these autonomous systems, complete integration with existing maritime communication systems, such as voice communication using Very High Frequency (VHF), remains a challenge [3].

Currently, VHF voice communication in the maritime industry serves as an essential tool that enables immediate and situational awareness-based information exchange between ships [4]. Communication between ships, based on standardized communication protocols like the Standard Marine Communication Phrases (SMCP), helps prevent collisions, maintain navigational safety, and effectively handle various situations that arise in the marine environment [5]. However, MASS lack the capability to understand VHF voice communication and generate appropriate responses suitable for the situation, unlike humans. This could hinder MASS from playing a significant role in collaboration with existing manned ships and communication with control centers. Particularly, voice-based communication is crucial for enhancing the autonomy and transparency of MASS, necessitating an effective technical solution for this.

In this study, we propose an integrated voice communication system for MASS to overcome these limitations. The proposed system is designed to enable MASS to autonomously interact through VHF voice communication by combining state-of-the-art speech recognition technology (STT, Speech-to-Text), speech synthesis technology (TTS, Text-to-Speech), and Large Language Models (LLMs). The system converts VHF voice data into text, generates context-appropriate responses through an LLM-based response generation module, and then converts these back into speech to be transmitted over VHF channels. Specifically, the responses generated by the LLM are based on SMCP data learned using Retrieval-Augmented Generation (RAG) technology, ensuring compliance with standardized maritime communication protocols.

The proposed system is designed to operate exclusively on an internal network to ensure stable functionality even in maritime environments where internet connectivity is unreliable. This design prevents external entities from accessing the system outside of related systems within the same internal network, thereby enhancing security and preventing disruptions in system usage due to unstable networks. Additionally, the system is configured so as not to interfere with the VHF communication processes employed by existing vessels when communicating with other ships. As a result, MASS can seam-

lessly interact with existing manned ships and control centers within the maritime communication network. This is expected to secure reliability and efficiency without causing confusion to counterparts interacting with the system in complex marine environments.

II. Related Work

A. Traditional Maritime Communication

Communication in traditional ships heavily relies on Very High Frequency (VHF) and the Automatic Identification System (AIS), which serve as pivotal tools for ensuring maritime safety and facilitating short-distance ship-to-ship and ship-to-shore communication [3]. VHF typically covers distances from 5 to 40 nautical miles, depending on atmospheric conditions and antenna height, enabling effective exchanges of navigational intentions among vessels or with coastal stations to minimize confusion. To address language barriers in VHF communications, the International Maritime Organization (IMO) introduced the Standard Marine Communication Phrases (SMCP), standardizing key expressions for ship-to-ship and ship-to-shore exchanges [6].

AIS devices, mandated under the 2002 Safety of Life at Sea (SOLAS) convention for ships exceeding 300 gross tonnage (GT), automatically transmit vessel position and motion data in the VHF band, and can extend coverage globally via Satellite AIS (S-AIS) [7]. While AIS is effective for digitally processed data and promotes situational awareness, fully interpreting analog VHF transmissions remains technically challenging for Maritime Autonomous Surface Ships (MASS). Consequently, an integrated strategy is necessary to seamlessly merge MASS operations with existing communication systems.

B. Speech-Text Conversion

To enable the system to interpret communications occurring on VHF, Speech-to-Text (STT) and Text-to-Speech (TTS) technologies are utilized. STT is a model that converts speech into text, helping the system interpret speech. Recent STT research focuses on improving speech recognition performance by utilizing large-scale unlabeled data to develop models that can generalize to more diverse environments [8]. Through this, models are being developed that can handle multiple languages and various tasks with a single model while maintaining high accuracy. By leveraging this, the system can interpret and process requests even in diverse language environments.

TTS is a model that converts given text into speech through speech synthesis, allowing it to output speech. It consists of a speech synthesis model that converts given text into acoustic representations and a vocoder that outputs the converted acoustic representations into actual speech [9]. The configured model can convert text to speech naturally and quickly, enabling communication through VHF by delivering speech. The integration of such STT and TTS is particularly helpful in the MASS environment by enhancing automation transparency, allowing human operators to more effectively monitor and interact with autonomous systems.

C. Large Language Model and Retrieval-Augmented Generation

To generate appropriate responses aligned with VHF requests and the surrounding situation of Maritime Autonomous Surface Ships (MASS), we utilize Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs). LLMs, pre-trained on vast amounts of textual data, excel at language understanding and generation, enabling them to handle complex linguistic patterns [10]. However, their knowledge can be limited when it comes to domain-specific or recent information because their training data may not cover every new development. To address this, we supplement LLMs with RAG, ensuring they can utilize the latest maritime information.

RAG overcomes LLM limitations by retrieving relevant external information before generating a response [11]. Instead of relying solely on the LLM's stored parameters, this method consults external documents or databases to incorporate real-time knowledge. As a result, the LLM can flexibly generate contextually accurate responses, simply by updating the external information without extensive retraining. By combining LLMs with RAG, our system consistently produces protocol-compliant, situation-specific responses, enhancing communication efficacy in maritime environments.

III. Method

The system consists of a voice module, an AIS data reception and management module, and a response generation module. All modules configured in the system are designed to be constructed without going through the internet due to the specificity of the maritime environment.

The response generation module utilizes an LLM using RAG to generate responses appropriate to the current situation. In the process of generating responses in this module, it uses the AIS data managed through the AIS reception and management module. The system is configured based on an environment utilizing the Korean language due to the limitations of the experimental setup, and the structure of the system can be confirmed in Figure 1.

A. System Architecture

The system is built on a server featuring an AMD Ryzen 7 9700X processor, a GeForce RTX 4080 SUPER (16GB VRAM), and four DDR5 PC5-44800 32GB memory modules. This single server runs the LLM, STT, and TTS models simultaneously. The LLM used in our system—Llama 3.1 8b—is quantized via Ollama, reducing its original size from approximately 34GB to about 4.9GB, and is executed on the GPU. Both the STT and TTS models also utilize the GPU for their operations.

B. Voice Communication

The voice module handles speech occurring in communication between MASS and VTS centers acting as control stations or other manned ships. Communication between MASS and VTS centers and other manned ships is composed in a way that when requests for information about MASS or action requests

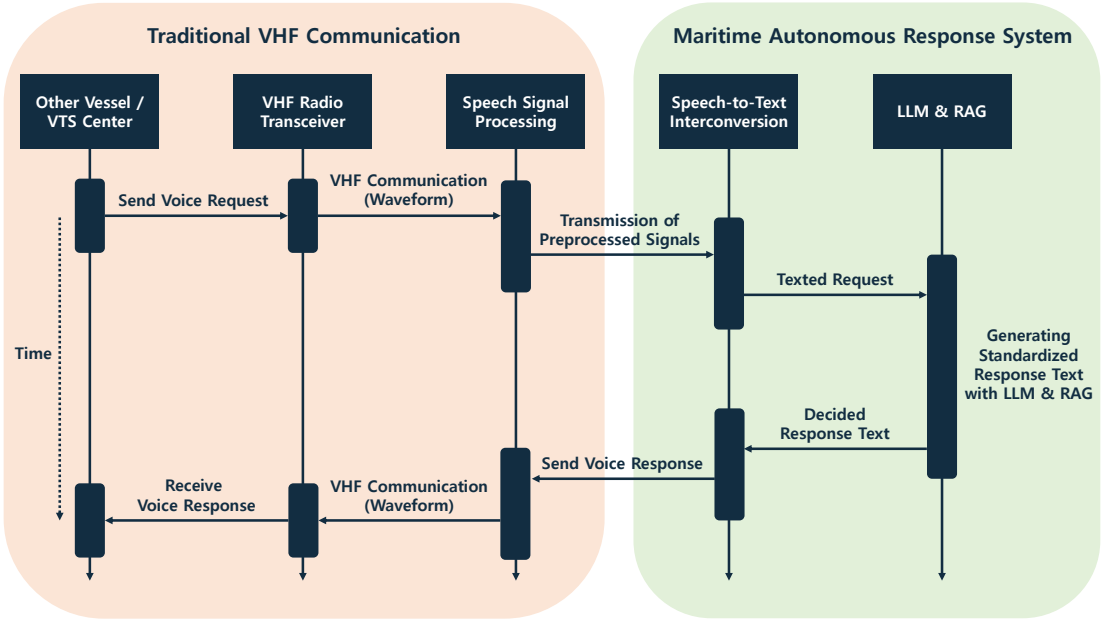

Fig. 1. Diagram illustrating the operational flow of the Maritime Autonomous Response System. It demonstrates the integration of VHF communication, speech signal processing, and LLM-based decision-making to generate standardized responses for maritime communication.

to MASS are given, MASS responds to them or reports after action. Voice communication is conducted through VHF used in maritime voice communication so as not to disrupt existing maritime voice communication processes.

The waveform of speech occurring on VHF is delivered to the system, and the waveform undergoes preprocessing through filtering to improve quality. Preprocessing enhances the areas corresponding to the voice by emphasizing frequencies through bandpass filtering and regional frequency emphasis. Through this process, even speech that could be interpreted by humans but not by the system due to noise can be interpreted by the system, improving performance.

C. Speech-Text Conversion Process

The preprocessed waveform is translated into text using STT, allowing the system to interpret the meaning. In this study, the STT model used is Whisper, developed by OpenAI. Whisper was developed in a weakly supervised manner and shows strong performance in multiple languages and various tasks. It also shows high robustness in noisy environments, with little performance degradation in environments with noise and background noise. Since the environment where the current system is applied uses VHF speech, even with preprocessing, there is a lot of noise included, so speech interpretation is performed through this model to secure high accuracy.

Also, Whisper can perform all operations without internet connection by downloading the model in a local environment, making it usable even in autonomous ship environments where security-related elements are important. In this study, the turbo model of Whisper is used with default settings. The turbo model has similar performance to the large model, which is the largest and highest-performing among Whisper models, and has a processing speed between the smallest and second smallest models in the Whisper model types. Therefore, it is a suitable model for the system that needs to consider fast processing speed and high accuracy, which are important in a real-time communication background. Through such a model, the requests interpreted into text are delivered to the response generation module to generate appropriate responses.

The responses generated and delivered through the response generation module are converted into speech through a TTS model so that they can be delivered through VHF. The TTS model used in this study is Melo-TTS, developed by MyShell-ai, and can generate speech in multiple languages, including English and Korean [12]. This model provides settings and model files for each language so that it can be used without additional training. This research is based on Korean, so the setting files and model files for Korean speech are used. The library that runs the model attempts to connect to download the setting files and model files every time it runs, so the code was modified to ensure there is no problem using the model offline. Also, the model's playback speed was adjusted so that it reads the text at a speed similar to that of a real person speaking, so that the listener does not feel awkward. The responses converted into speech through this model are delivered as speech through VHF.

D. Response Generation Process

The response generation process generates responses using an LLM fine-tuned with RAG. The requests interpreted into text through the voice module are processed using the data table organized through the AIS data reception and manage-

ment module to generate appropriate response texts based on scenarios learned through RAG.

We utilize a Large Language Model (LLM) that has been pre-trained on a vast dataset and demonstrates excellent performance in various natural language processing tasks to generate responses to requests. The LLM employed is the Llama 3.1 8b model developed by Meta. The Llama 3 model is a next-generation large-scale language model created by Meta AI, based on the latest Transformer architecture [13]. It has been trained on data in multiple languages, including Korean, enabling it to perform tasks such as high-level text generation, document summarization, and question answering.

However, despite being trained on extensive data, the Llama model tends to generate responses based solely on the information included in its training dataset. Its response accuracy may diminish when dealing with the latest information that emerged after its training or with knowledge in specific domains not covered by the training data. To overcome these limitations, we incorporate additional training data by integrating Retrieval-Augmented Generation (RAG).

We deliver pre-constructed maritime communication scenarios through the Retrieval module that constitutes our Retrieval-Augmented Generation (RAG) system. The engine that searches these received scenarios adopts the KkmaBM25 algorithm. The KkmaBM25 algorithm combines the Kkma algorithm—a Korean morphological analyzer—with the BM25 algorithm, which is used in existing RAG systems for efficient information retrieval. This combination is advantageous for enhancing retrieval performance for maritime communication scenarios composed in Korean.

Kkma is a Korean morphological analyzer developed by the Natural Language Processing Laboratory at Seoul National University [14]. It provides various functions necessary for Korean natural language processing, such as morphological analysis, part-of-speech tagging, and syntactic parsing. Kkma is designed to accurately analyze the complex lexical structures and grammar of the Korean language. As an open-source tool, it is widely used in various research and industrial fields. By utilizing Kkma, the system can effectively grasp the core content of queries and documents by reflecting the intricate structures of Korean. Additionally, specialized terms used in maritime communication are added to the dictionary where Kkma stores pre-classified words, enabling the classification of terms that occur specifically in maritime contexts.

The BM25 algorithm is a ranking function widely used in the field of information retrieval to evaluate the relevance between documents and queries [15]. Based on probabilistic language models, it calculates relevance by considering the frequency of terms within documents and the distribution of terms across documents. BM25 computes relevance by utilizing term frequency—which reflects the importance of a term by indicating how often it appears in a document—and inverse document frequency—which indicates how rare a term is across the entire document set. It also applies a correction factor to minimize the impact of document length. BM25 demonstrates simple yet effective performance, providing accurate retrieval results by reflecting the actual linguistic similarity between documents and queries.

By combining the two algorithms into the KkmaBM25 algorithm, we can efficiently calculate the similarity between a given query and input scenarios using the BM25 algorithm on Korean requests analyzed by Kkma, thereby determining the scenario most similar to the query [16]. The RAG configured in the system leverages this KkmaBM25 to extract scenarios suitable for maritime communication-specific queries composed in Korean. The extracted scenarios are then used by the LLM to generate appropriate response texts, which are passed to the Text-to-Speech Conversion stage to be converted into speech, allowing them to be transmitted via VHF communication.

IV. Result

A. System Performance Overview

In this study, the proposed system was experimentally verified in a controlled maritime simulation environment to evaluate its communication and information processing capabilities for Maritime Autonomous Surface Ships (MASS). The experiments were conducted based on key performance indicators that include VHF-based voice communication processing accuracy, AIS data collection and real-time updating performance, and the quality of contextually appropriate response generation. These indicators were defined as critical factors in determining whether MASS can autonomously perform communication while harmonizing with existing maritime communication protocols. The GUI configured within the system can be verified in Figure 2.

B. Evaluation of Speech Recognition (STT)

The voice communication module was evaluated by interactions using a microphone. Real VHF-based voice communication is limited to a frequency range of 300Hz to 3000Hz and is subject to noise caused by factors such as communication distance or weather conditions. To simulate these characteristics within the system, the audio input from the microphone was filtered using a bandpass filter to restrict the frequency range and increased with white noise 5%, making it similar to actual VHF audio. The main performance indicators and results of the module under these conditions are as follows.

- **Accuracy:** The Whisper model demonstrated high robustness even in noisy environments, achieving an average Word Error Rate (WER) of 8.2%. The errors primarily involved words with similar pronunciations, such as recognizing ”ETB” as ”ETV,” and most of the misidentified words were proper nouns like pier names or destinations. Other requests, such as information queries and action commands, were accurately interpreted. This indicates that the model effectively extracts text even under simulated VHF communication conditions.
- **Processing Speed:** For speech recordings with an average length of 10 seconds, the average transcription delay time was measured at 0.4 seconds, which is sufficient

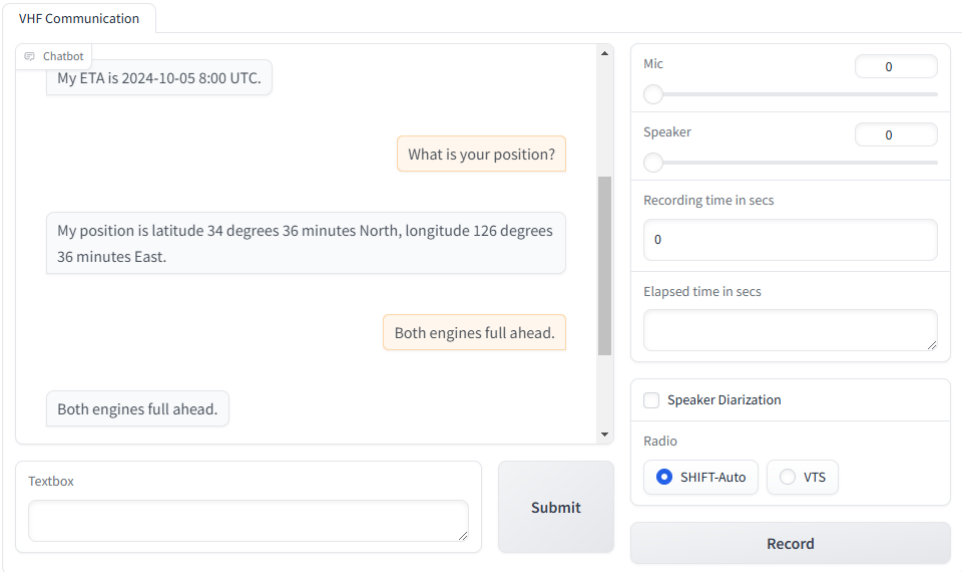

Fig. 2. User interface of the maritime communication system, showcasing the VHF communication functionality, real-time message exchange for seamless maritime operations. The requests and responses visible through the chat GUI are displayed in Korean but have been translated into English for better understanding.

processing speed to support real-time interaction between MASS and manned ships.

Although this experiment has limitations due to differences from actual environments using VHF radios, it was confirmed that even in such tests, the Whisper model shows excellent performance in terms of noise resistance and transcription accuracy. Future studies need to include tests in actual environments using VHF radios to more precisely verify system performance.

C. Response Generation

The response generation module focused on generating contextually appropriate responses by combining Retrieval-Augmented Generation (RAG) and LLM. Test results showed that the module generated contextually appropriate responses in 96.3% of test cases, meeting the specific requirements of maritime communication.

RAG utilized the KkmaBM25 algorithm to search user requests from maritime communication scenarios composed in Korean. The KkmaBM25 algorithm effectively processed the complex grammatical structure of Korean and enhanced search performance by configuring a dictionary that includes specialized terms used in maritime communication. This search engine accurately retrieved scenarios appropriate to the request with an average delay time of 4 seconds.

The LLM (Llama 3.1 8b model) combined the retrieved scenarios with AIS data to generate responses. The generated responses complied with maritime communication protocols and showed a high level of contextual appropriateness by reflecting user requests and situational context. The responses were prepared to be delivered through VHF by converting them into speech through the TTS module.

In this process, the combination of RAG and LLM greatly enhanced the autonomy of MASS. However, there is a limitation in that the error rate is high for responses outside the scenarios learned due to limiting the scenarios for generating responses through RAG.

D. System Usability and Integration

Simulation tests based on the MASS environment focused on evaluating the overall usability and integration possibilities of the system.

- **Integration:** The modular architecture allowed for seamless integration with the VHF communication system, AIS data receivers, and internal decision-making systems. The user interface was designed to be straightforward, enabling vessel operators to easily understand and utilize the system.
- **Offline Operation:** All models, including Whisper and Melo-TTS, were designed to be executable in a local environment without internet connection. This significantly enhanced operational stability by eliminating internet dependency in the maritime environment. It was confirmed that the system operates normally even in network failure situations..

V. Conclusion

In this study, we proposed an integrated system that enhances the autonomy of Maritime Autonomous Surface Ships (MASS) and enables efficient communication and information processing while maintaining compatibility with existing maritime communication protocols. The system consists of a voice communication module, an AIS data management module, and a response generation module based on Retrieval-Augmented

Generation (RAG), and was verified in a controlled simulation environment.

The voice communication module showed high speech recognition accuracy and real-time processing performance in noisy maritime environments by utilizing the Whisper model. This provided a foundation for MASS to reliably interpret speech delivered through VHF and make autonomous decisions. However, since this study used ambient recording data rather than data obtained through actual VHF communication devices, future studies need to include additional tests using actual VHF devices.

The AIS data management module contributed significantly to enhancing the situational awareness capability of MASS. By effectively integrating static and dynamic AIS data to build comprehensive ship information, real-time decision support of the autonomous system became possible. Especially, the structure of data integration based on MMSI and real-time updating of dynamic data provided a foundation for MASS to quickly and accurately analyze the surrounding situation in complex maritime environments.

The response generation module successfully generated contextually appropriate responses by combining RAG and LLM (Llama 3.1). Through this, MASS could generate accurate responses reflecting the specificity of maritime communication, and they were perfectly compatible with existing VHF communication protocols by being converted into speech through TTS. However, a limitation was identified in that accuracy decreases for requests outside the learned scenarios due to the limited scenario-based training data. Future studies need to overcome this limitation by expanding the range of scenario data that can be utilized through RAG.

In simulation-based tests, the system sufficiently satisfied the real-time and durability required in maritime environments and was highly evaluated by users for its usability. Especially, the local execution of the Whisper and Melo-TTS models ensured stable operation even in maritime environments where internet connection is limited.

The deployment of the system in an actual maritime environment has certain limitations. The scenarios used for training are centered around standardized expressions and are designed based on test environments relevant to developing autonomous ships. However, real-world maritime VHF communication is not standardized and does not strictly adhere to the expressions included in the training scenarios. Consequently, applying the system to real maritime communication could result in significant confusion.

In conclusion, the system in this study presented a foundation for MASS to secure autonomy and interaction transparency while harmonizing with existing maritime communication systems. This system shows the potential to enhance maritime safety and operational efficiency by supporting the communication and decision-making processes of MASS. Future studies need to further enhance the completeness of the system through experimental verification in actual maritime environments and expansion of training data reflecting various maritime communication scenarios.

ACKNOWLEDGMENT

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS2022-00155857, Artificial Intelligence Convergence Innovation Human Resources Development (Chungnam National University)), and the 'Regional Innovation Strategy (RIS)' through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2021RIS-004). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

[1] H.-J. Kim, H.-S. Roh, and J.-B. Yim, "Development and Performance Evaluation Results of Remote Control Systems for Maritime Autonomous Surface Ships," Journal of Navigation and Port Research, vol. 48, no. 4, pp. 335–341, Aug. 2024.https://doi.org/10.5394/KINPR.2024.48.4.335

[2] M. A. Hinostroza and A. M. Lekkas, "Temporal mission planning for autonomous ships: Design and integration with guidance, navigation and control," Ocean Engineering, vol. 297. Elsevier BV, p. 117104, Apr-2024.

[3] O. A. Alsos, P. Hodne, O. K. Skåden, and T. Porathe, "Maritime Autonomous Surface Ships: Automation Transparency for Nearby Vessels," Journal of Physics: Conference Series, vol. 2311, no. 1. IOP Publishing, p. 012027, 01-Jul-2022.

[4] J. Renner, "VHF maritime mobile communications: A systems approach to serving user requirements," IEEE Transactions on Vehicular Technology, vol. 26, no. 3. Institute of Electrical and Electronics Engineers (IEEE), pp. 213–222, Aug-1977.

[5] Žanić Mikuličić, Jelena Pavlinović, Mira Trgo, Branimir Role of maritime English in managing a vessel // Economic and Social Development (Book of Proceedings), 105 th International Scientific Conference on Economic and Social Development - "Building Resilient Society" / Kovac, Ivana; Misevic, Petar; Zahariev, Andrey (ur.).

[6] International Maritime Organization. "Standard Marine Communication Phrases (SMCP)." IMO.org. Accessed November 2024. https://www.imo.org/en/ourwork/safety/pages/standardmarinecommunicationphrases.aspx

[7] Felski, Andrzej, Krzysztof Jaskólski, and Paweł Banyś. "Comprehensive Assessment of Automatic Identification System (AIS) Data Application to Anti-Collision Manoeuvring." Journal of Navigation 68, no. 4 (2015): 697–717. https://doi.org/10.1017/S0373463314000897.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision." arXiv, 2022.

[9] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao et al., "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech." arXiv, 2020.

[10] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal et al., "Language Models are Few-Shot Learners." arXiv, 2020.

[11] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." arXiv, 2020.

[12] Zhao, Wenliang, Xumin Yu, and Zengyi Qin. 2023. MeloTTS: High-quality Multi-lingual Multi-accent Text-to-Speech. Computer software. GitHub. https://github.com/myshell-ai/MeloTTS

[13] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle et al., "The Llama 3 Herd of Models." arXiv, 2024

[14] D.-J. Lee, J.-H. Yeon, I.-B. Hwang and S.-G. Lee. "KKMA : A Tool for Utilizing Sejong Corpus based on Relational Database" Journal of KIISE: Computing Practices and Letters 16, no.11 (2010) : 1046-1050.

[15] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends® in Information Retrieval, vol. 3, no. 4. Now Publishers, pp. 333–389, 2009.

[16] K.-R. Lee (2024). langchain-teddynote [Computer software]. GitHub. https://github.com/teddylee777/langchain-teddynote