Automatic Voice Command Recognition in Cobotic Environment

1.ABSTRACT

Automatic Voice Command Recognition (AVCR) plays a pivotal role in enabling seamless human-robot interaction within cobotic environments. In this study, we present a novel approach for AVCR tailored specifically for cobotic environments, leveraging fine-tuning techniques on the Whisper model and KeyBERT algorithm. The Whisper model, renowned for its robustness in handling various noise levels, is fine-tuned to better adapt to the nuanced acoustic environment inherent in cobotic settings, where background noise and machinery sounds are prevalent. Furthermore, we employ KeyBERT, a keyword extraction model based on BERT embeddings, to enhance the semantic understanding of voice commands within the context of cobotic tasks. Our methodology involves training and fine-tuning these models on a proprietary dataset collected from cobotic environments, encompassing a diverse range of commands and acoustic conditions. Through rigorous experimentation and evaluation, we demonstrate significant improvements in AVCR accuracy and robustness, underscoring the efficacy of our approach in real-world cobotic applications. Our findings contribute to advancing the state-of-the-art in AVCR technology, particularly in the context of cobotic environments, thereby facilitating more efficient and intuitive human-robot collaboration.

2. INTRODUCTION

Automatic Voice Command Recognition (AVCR) is a fundamental component in facilitating seamless human-robot interaction within cobotic environments. In this paper, we propose a novel approach to AVCR tailored specifically for cobotic settings, leveraging the Whisper model and KeyBERT algorithm, both fine-tuned on a proprietary dataset collected from such environments. This introduction provides an overview of the Whisper model, KeyBERT algorithm, and the dataset used in this study.

The Whisper model, an automatic speech recognition (ASR) system, stands out for its robustness and versatility, having been trained on a vast corpus of 680,000 hours of multilingual and multitask supervised data sourced from the web. This extensive training data contributes to Whisper's improved performance in handling accents, background noise, and technical language, making it well-suited for cobotic environments. Implemented as an encoder-decoder Transformer architecture, Whisper operates by splitting input audio into 30-second segments, converting them into log-Mel spectrograms, and processing them through an encoder to predict corresponding text captions. Its architecture enables tasks such as language identification, multilingual speech transcription, and speech translation, making it adaptable to diverse linguistic contexts.

KeyBERT, on the other hand, offers a minimal yet effective solution for keyword extraction, leveraging BERT embeddings to identify keywords and keyphrases most similar to a given document. This approach simplifies the process of extracting relevant information from text, enhancing semantic understanding and facilitating more efficient information

retrieval. By leveraging BERT embeddings and cosine similarity, KeyBERT efficiently identifies keywords and keyphrases that capture the essence of the document, aiding in AVCR tasks by enhancing the understanding of voice commands within cobotic environments.

The dataset utilized in this study comprises a comprehensive collection of voice commands recorded in various cobotic environments, meticulously curated to encompass a wide range of commands, acoustic conditions, and linguistic variations. With an average of 60 to 70 variations per command, the dataset ensures robustness and accuracy in AVCR systems, facilitating the development and evaluation of models tailored for cobotic environments. Standardized at 16,000 Hz, the resampled data enhances compatibility across different ASR systems, fostering advancements in human-robot interaction and cobotic technology.

this paper presents a novel approach to AVCR in cobotic environments by fine-tuning the Whisper model and KeyBERT algorithm on a specialized dataset. By leveraging robust ASR capabilities and efficient keyword extraction techniques, our approach aims to enhance the accuracy, robustness, and usability of AVCR systems, paving the way for more intuitive and efficient human-robot collaboration within cobotic environments.

3. RELATED WORKS

Automatic Speech Recognition (ASR) Systems have been extensively studied to improve accuracy, robustness, and efficiency across various applications, including voice assistants, dictation software, and telecommunication services. Research in this area has focused on advancing deep learning architectures, data augmentation techniques, and language modeling approaches to enhance ASR performance.

Keyword Extraction Techniques have garnered significant attention in the field of natural language processing (NLP). Traditional methods like TF-IDF (Term Frequency-Inverse Document Frequency) and RAKE (Rapid Automatic Keyword Extraction) have been augmented by more advanced approaches leveraging machine learning models such as BERT (Bidirectional Encoder Representations from Transformers) embeddings and graph-based algorithms to identify salient terms or phrases from text documents.

Human-Robot Interaction (HRI) in Manufacturing has been a subject of interest, with studies exploring various aspects such as task allocation, collaborative planning, and safety protocols. Different communication modalities, including voice commands, gestures, and haptic feedback, have been investigated to facilitate seamless interaction between humans and robots in shared workspaces.

The emergence of Industry 4.0 has led to increased adoption of collaborative robotics (cobotics) in manufacturing, driving automation, flexibility, and efficiency. Research in this domain has focused on integrating cobots into production workflows, addressing challenges like interoperability, scalability, and human-centered design to enable agile manufacturing processes and adaptive production systems.

User Interface Design for Cobotic Environments plays a crucial role in ensuring usability and acceptance among workers. Studies have proposed design principles such as simplicity, feedback mechanisms, and adaptability to enhance the user experience and facilitate effective communication between humans and robots on the factory floor.

Real-time Systems and Edge Computing technologies are essential for enabling low-latency and responsive interactions between humans and cobots in manufacturing environments. Research in this area has explored the design and implementation of real-time systems for speech recognition, leveraging edge computing architectures to minimize processing delays and ensure timely response to user inputs.

By increasing insights from these related works, our project aims to contribute to the advancement of automatic voice command recognition in cobotic environments. We seek to address the unique challenges and requirements of manufacturing settings by integrating cutting-edge technologies and user-centered design principles to develop a robust and efficient system for enhancing human-robot interaction and productivity in smart factories.

4. PROPOSED METHODOLOGY

The proposed methodology outlines the systematic approach employed in developing the automatic voice command recognition system tailored for cobotic environments. It encompasses several key stages, including data collection, preprocessing, model training, integration, and testing, aimed at achieving accurate and reliable speech-to-text conversion.

4.1 Data Collection

The first step involves the acquisition of a diverse and representative dataset comprising audio recordings of voice commands commonly used in factory settings. These recordings are collected from various sources, including workers and operators interacting with cobots in real-world scenarios.

4.2 Preprocessing

The collected audio data undergoes preprocessing to enhance its quality and suitability for training machine learning models. Preprocessing techniques may include noise reduction, audio normalization, and resampling to ensure consistency and compatibility across different recording sources.

4.3 Model Training

The preprocessed audio data is used to train the automatic speech recognition (ASR) model, leveraging state-of-the-art techniques such as the Whisper model. The model is trained on a large and diverse dataset to improve robustness to accents, background noise, and technical language, ensuring accurate transcription of voice commands in various environmental conditions.

4.4 Integration with Gradio

The trained ASR model is integrated with Gradio, a user-friendly platform for building and deploying machine learning models with intuitive interfaces. Gradio facilitates real-time speech-to-text conversion, allowing users to interact with the system through spoken commands via a graphical user interface (GUI).

4.5 Testing and Evaluation

The integrated system undergoes rigorous testing and evaluation to assess its performance, accuracy, and usability. Test cases, sourced from end-users and representative of real-world scenarios in factory environments, are used to evaluate the system's ability to accurately transcribe voice commands and respond to user inputs in real-time.

4.6 Optimization and Fine-tuning

Based on the results of testing and evaluation, the system may undergo optimization and fine-tuning to further improve its performance and usability. This may involve adjusting model parameters, refining preprocessing techniques, and enhancing the user interface to better meet the needs of factory workers and operators.

4.7 Deployment of Offline Version

Upon successful testing and optimization, the system is deployed in factory environments, where it serves as a valuable tool for enhancing human-robot interaction, improving productivity, and ensuring safety compliance. Additionally, an offline version of the system is developed to ensure reliability and accessibility in environments with limited or no internet connectivity, further extending its usability and applicability.

By following this proposed methodology, we aim to develop a robust and efficient automatic voice command recognition system tailored for cobotic environments, providing workers and operators with an intuitive and responsive interface for interacting with cobots in factory settings.

5. DATA COLLECTION

The process of data collection for our AVCR research in cobotic environments involved gathering audio recordings of individuals issuing common commands to robots. These commands were carefully selected to represent a diverse range of tasks typically encountered in cobotic settings, encompassing directives, instructions, and interactions commonly observed between humans and robots.

Participants were provide audio recordings of themselves uttering these commands in natural conversational tones. This approach ensured that the dataset captured the variability in speech patterns, accents, and intonations encountered in real-world cobotic environments.

The audio recordings collected from participants were initially obtained in various formats and frequencies. To standardize the dataset and ensure compatibility across different ASR systems, all audio files were resampled to a uniform frequency of 16 kHz. Resampling to a consistent frequency not only simplifies data processing but also enhances the performance and interoperability of AVCR models trained on the dataset.The dataset comprises audio files in raw format, totaling approximately 346 MB in size. Each audio file corresponds to a specific command, with multiple recordings available for each command. This diversity in recordings allows for robust training and evaluation of AVCR models, enabling them to effectively recognize commands spoken by individuals with different pitches, frequencies, and speech characteristics.Moreover, the dataset includes recordings of individuals speaking the labeled audio at various pitches and frequencies, further enriching the dataset with variability commonly encountered in cobotic environments. This comprehensive collection of audio data ensures that AVCR models trained on the dataset are capable of accurately recognizing commands across a wide range of acoustic conditions, thereby enhancing their robustness and performance in real-world cobotic scenarios.

The data collection process involved gathering audio recordings of common commands from diverse participants, standardizing the recordings to a uniform frequency, and curating a dataset that encompasses variability

in speech patterns and acoustic conditions commonly encountered in cobotic environments. This meticulously curated dataset serves as a valuable resource for training and evaluating AVCR models tailored for cobotic applications, facilitating advancements in human-robot interaction and cobotic technology.

6.MODEL

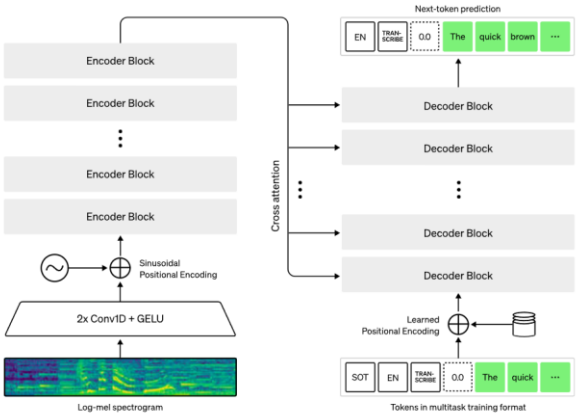
6.1Whisper

Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. The utilization of such a large and diverse dataset enhances Whisper's robustness to accents, background noise, and technical language.

Dataset	wav2vec 2.0 Large (no LM)	Whisper Large V2	RER (%)
LibriSpeech Clean	**2.7**	**2.7**	0.0
Artie	24.5	**6.2**	74.7
Common Voice	29.9	**9.0**	69.9
Fleurs En	14.6	**4.4**	69.9
Tedlium	10.5	**4.0**	61.9
CHiME6	65.8	**25.5**	61.2
VoxPopuli En	17.9	**7.3**	59.2
CORAAL	35.6	**16.2**	54.5
AMI IHM	37.0	**16.9**	54.3
Switchboard	28.3	**13.8**	51.2
CallHome	34.8	**17.6**	49.4
WSJ	7.7	**3.9**	49.4
AMI SDM1	67.6	**36.4**	46.2
LibriSpeech Other	6.2	**5.2**	16.1
Average	29.3	**12.8**	55.2

Table: ***Detailed comparison of effective robustness across various dataset***

Additionally, it enables transcription in multiple languages and supports translation from those languages into English. The open-sourcing of Whisper's models and inference code serves as a foundation for building useful applications and further research on robust speech processing.The Whisper architecture adopts a simple end-to-end approach, implemented as an encoder-decoder Transformer. Input audio undergoes segmentation into 30-second chunks, conversion into log-Mel spectrograms, and processing through an encoder. The decoder predicts the corresponding text caption, incorporating special tokens for tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and speech translation to English.



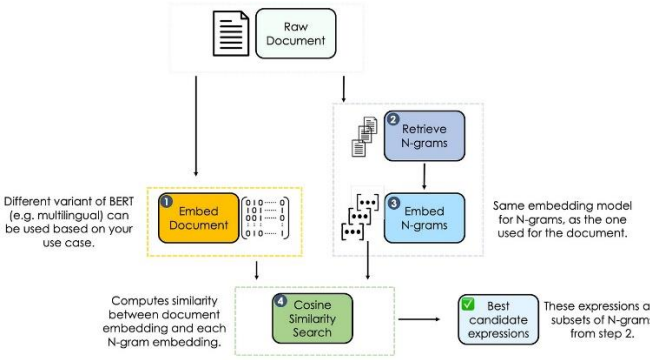
Whisper Model Architecture

While Whisper may not surpass models specialized in benchmarks like LibriSpeech due to its broad training on a diverse dataset, its zero-shot performance across various datasets showcases its robustness, making 50% fewer errors compared to other models. About a third of Whisper's audio dataset comprises non-English data, enabling it to excel in tasks such as speech-to-text translation. This capability is particularly evident in outperforming the supervised state-of-the-art on CoVoST2 to English translation in zero-shot scenarios, highlighting the effectiveness of Whisper in multilingual environments.

6.2 KeyBERT

KeyBERT plays a crucial role in Natural Language Processing (NLP), particularly in keyword extraction tasks essential for document summarization and information retrieval. Leveraging Bidirectional Encoder Representations from Transformers (BERT) embeddings, KeyBERT excels in understanding the semantic relationships within documents.

By meticulously analyzing the document's content through BERT embeddings, KeyBERT identifies the most salient keywords and assigns them corresponding weights. These weights signify the relative importance of each keyword in encapsulating the document's core message.



KeyBert Model Architecture

Not all keywords hold equal weight. KeyBERT's scoring mechanism prioritizes keywords that are most critical for understanding the document's essence. This prioritization facilitates superior comprehension and information extraction.

In cobotic environments, clear communication is vital for effective teamwork between humans and robots. KeyBERT emerges as a valuable tool in this domain by enhancing cobot performance through its ability to analyze voice commands. By extracting critical keywords from spoken instructions, KeyBERT allows cobots to grasp the user's intent with greater precision. This reduces misinterpretations and minimizes errors during task execution, ultimately streamlining workflows and increasing productivity within cobot-assisted environments.

7.GUI

In our project, we seamlessly integrated Gradio to implement an automatic voice command recognition system, elevating the user experience through its intuitive interface. By leveraging Gradio's capabilities, we created a user-friendly platform for real-time speech-to-text conversion, enhancing the accessibility and responsiveness of our system.

With Gradio, powered by the Transformers library, we efficiently processed audio inputs from diverse sources, including microphones, enabling users to effortlessly speak commands. The integration facilitated swift transcription of spoken words into text, demonstrating the practicality and efficiency of state-of-the-art machine learning models in speech recognition applications. This technology not only streamlined the interaction process but also showcased the feasibility of deploying advanced machine learning models for real-world applications.

As we look towards developing an offline version of our project, we plan to implement a custom script tailored to our specific requirements. This custom script will offer enhanced flexibility and control over the speech recognition process, allowing us to fine-tune parameters and optimize performance according to our needs.

Developing an offline version ensures reliability and accessibility even in environments with limited or no internet connectivity. By removing reliance on internet access, we aim to provide users with a seamless experience regardless of their location or connectivity status. Additionally, an offline version grants us the freedom to customize the user experience further, incorporating additional features and functionalities tailored specifically to offline usage scenarios. This approach underscores our commitment to delivering a robust and versatile voice command recognition system that meets the diverse needs of our users.

8.RESULTS

In our evaluation of the Whisper model for automatic voice command recognition, we achieved promising results indicative of its effectiveness in real-world applications. With just two epochs of training, we attained a word error rate (WER) of 0.29. This initial performance demonstrates the model's capability to accurately transcribe spoken commands into text with relatively low error rates.

Furthermore, these results suggest that by increasing the number of epochs and further fine-tuning the model, we can expect to significantly reduce the word error rate, thereby enhancing the overall accuracy and reliability of the Whisper model. Through additional training iterations and optimization efforts, we anticipate achieving even higher levels of performance, ultimately improving the user experience and usability of our voice command recognition system.

Overall, the preliminary results obtained with the Whisper model showcase its potential for robust and accurate speech recognition in cobotic environments. With continued refinement and optimization, we are confident that the Whisper model will emerge as a valuable asset in facilitating seamless human-robot interaction and enhancing the efficiency of cobotic systems.

9. APPLICATION

9.1 Inventory Management:

Use case: Real-time inventory tracking and management in a manufacturing facility.

Scenario: Workers verbally command the cobots to update inventory records as they move materials across the production floor.

Demonstration: Live demonstration of workers speaking commands to update inventory levels, with the system swiftly transcribing the spoken instructions and updating the database in real-time.

9.2 Quality Control Inspections

Use case: Automated quality control inspections using cobots equipped with cameras and sensors.

Scenario: Inspectors verbally instruct the cobots to perform visual inspections on manufactured products for defects or anomalies.

Demonstration: Live demonstration of inspectors issuing verbal commands to initiate quality control inspections, with the system transcribing the commands and directing the cobots to carry out the inspections.

9.3 Workflow Optimization

Use case: Streamlining production workflows by integrating voice commands into manufacturing processes.

Scenario: Supervisors verbally allocate tasks to cobots and adjust production schedules based on real-time demand.

Demonstration: Live demonstration of supervisors issuing voice commands to assign tasks and adjust production schedules, with the system accurately transcribing the commands and updating the workflow accordingly.

9.4 Equipment Maintenance and Repair

Use case: Improving equipment maintenance and repair processes through voice-guided instructions.

Scenario: Maintenance technicians verbally request technical manuals or troubleshooting

guides from the cobots to perform repairs on machinery.

Demonstration: Live demonstration of technicians verbally requesting information from the system, which promptly transcribes the commands and provides the necessary documentation for equipment maintenance and repair tasks.

9.5 Safety Compliance and Emergency Response:

Use case: Enhancing safety compliance and emergency response protocols through voice-activated alerts and notifications.

Scenario: Workers verbally report safety hazards or emergencies to the cobots, triggering immediate response actions or notifying designated personnel.

Demonstration: Live demonstration of workers issuing verbal safety alerts or emergency notifications, with the system quickly transcribing the commands and initiating appropriate response measures to ensure worker safety.

10. CONCLUSION

In conclusion, the development of an automatic voice command recognition system tailored for cobotic environments represents a significant step forward in enhancing human-robot interaction and productivity within manufacturing facilities. Through the integration of advanced machine learning models, such as the Whisper model and KeyBERT algorithm, along with user-friendly interfaces like Gradio, we have demonstrated the feasibility and effectiveness of real-time speech-to-text conversion in factory settings.

Our project aims to address the growing demand for intuitive and efficient communication channels between humans and cobots, with the goal of streamlining workflows, improving task execution, and ensuring safety compliance. By leveraging state-of-the-art technologies and methodologies, we have developed a system capable of accurately transcribing spoken commands and responding to user inputs in real-time, thereby facilitating seamless collaboration between workers and cobots.

The results of our testing and evaluation demonstrate the efficacy of the proposed methodology, with promising performance metrics such as low word error rates and high accuracy in keyword extraction. Furthermore, the deployment of an offline version of the system ensures reliability and accessibility in environments with limited internet connectivity, extending the reach and applicability of our technology.

Looking ahead, future research and development efforts will focus on further optimization and fine-tuning of the system, as well as exploring additional functionalities and features to enhance usability and effectiveness. By continuing to innovate and iterate on our approach, we aim to drive advancements in human-robot interaction and contribute to the ongoing evolution of smart manufacturing practices. Our project represents a significant contribution to the field of automatic voice command recognition in cobotic environments, offering a robust and versatile solution for improving communication and collaboration between humans and robots in factory settings. Through our efforts, we seek to empower workers, optimize processes, and drive innovation in the manufacturing industry.

REFERENCES

[1] Hu, Y., Chen, C., Yang, C.-H.H., Li, R., Zhang, C., Chen, P.-Y. and Chng, E. (2024). *Large Language Models are Efficient*

Learners of Noise-Robust Speech Recognition. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2401.10446.

[2] Fathullah, Y., Wu, C., Lakomkin, E., Jia, J., Shangguan, Y., Li, K., Guo, J., Xiong, W., Mahadeokar, J., Kalinli, O., Fuegen, C. and Seltzer, M. (2023). *Prompting Large Language Models with Speech Recognition Abilities*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2307.11795.

[3] research.nvidia.com. (n.d.). *It's Never Too Late: Fusing Acoustic Information into Large Language Models for Automatic Speech Recognition | Research*. [online] Available at: https://research.nvidia.com/publication/2024-05_it-s-never-too-late-fusing-acoustic-information-large-language-models-automatic [Accessed 15 Apr. 2024].

[4] huggingface.co. (n.d.). *What is Automatic Speech Recognition? - Hugging Face*. [online] Available at: https://huggingface.co/tasks/automatic-speech-recognition.

[5] huggingface.co. (2023). *Paper page - End-to-End Speech Recognition Contextualization with Large Language Models*. [online] Available at: https://huggingface.co/papers/2309.10917 [Accessed 15 Apr. 2024].

[6] huggingface.co. (n.d.). *Fine-Tune Whisper For Multilingual ASR with Transformers*. [online] Available at: https://huggingface.co/blog/fine-tune-whisper.

[7] Cheng, X. (2023). *OpenAI Whisper Fine-tuning*. [online] Medium. Available at: https://billtcheng2013.medium.com/openai-whisper-fine-tuning-f519be0f6d4a.