Applying Automatic Speech Recognition on Intelligent Human-Robot Interfaces for Operational Usage

IOANNIS GIACHOS, VASILEIOS-STYLIANOS LEFKELIS,
EVANGELOS C. PAPAKITSOS*, PETROS SAVVIDIS, NIKOLAOS LASKARIS
Department of Industrial Design & Production Engineering,
University of West Attica,
12241 Egaleo, Athens
GREECE

*Corresponding Author

Abstract: - This paper deals with the implementation of a readily available automatic speech recognition (ASR) system in a human-robot interface (HRI), intended for operational uses. Automatic speech recognition is a very important process that has occupied artificial intelligence for over 70 years. The aim is to build the prerequisites with a basic code for the full integration of a modern advanced automatic speech recognition system into an intelligent human-robot interface, designed by the authors, and which is part of a developing robotic system. At the beginning of this paper, a brief discussion of the history of ASRs and the techniques used is presented.

Key-Words: ASR, NLP, NLG, NLU, Speech Recognition, Dialog System, OMAS-III, HRI, Operational Robots, TTS.

Received: March 19, 2024. Revised: November 9, 2024. Accepted: December 6, 2024. Published: February 10, 2025.

1 Introduction

In a previous study, we referenced intelligent virtual assistants, [1]. This study defined this continuously evolving artificial function as the "brain" of an advanced humanoid robot. Reference was made to the area of communication between an intelligent virtual assistant and its environment, through an advanced human-robot interface (HRI). Such an interface should include a unit for collecting natural language input, a unit for processing received sentences, a unit for the semantic representation of these sentences [2], and, additionally, a unit for generating and reproducing natural language.

The first of the units above is the focus of this paper and, by extension, one of the core components of a robotic system being developed by the authors that is based on the communicative capabilities of the OMAS-III systemic conceptual model [3], and experimentally employs a small natural language dictionary in Greek. This unit is known as Automatic Speech Recognition (ASR), and its main function is to convert incoming speech into text.

Nowadays there are many applications that do not require ASR systems, such as chat boxes, which are at the forefront of research to expand their application to more sectors in our daily life, such as examples from the agriculture sector, [4]. However, there is a wide range of advanced ASR systems available on the market. Countless daily applications around the world use those ASR systems. Such applications are parking entrance/exit, telephonists, home assistants, etc., but also social, industrial, and operational robots. The authors' interest is directed towards operational robots, where the environment is often quite difficult for human presence. One such case is the taking of soil, water, or air samples in an open area on land or sea, with the only control procedures that of a speech-dialogue model.

The structure of the paper consists of six chapters. The first chapter is an introduction. The second chapter briefly discusses ASR technology, its historical development, how it works, the methods it uses, and, finally, the recent research interest in its application on operational robots. In the third chapter, there is a discussion on this system's development. The fourth chapter details the programming and testing code, for integrating the ASR unit into our developing system. The fifth chapter discusses this developing system and how it provides possibilities for expanding its capabilities. The final, sixth, chapter presents conclusions and suggestions for future research.

2 ASR (Auto Speech Recognition)

Automatic Speech Recognition (ASR) is a technology used by the Speech Recognition Unit in human-machine dialogue systems. Nowadays, it is considered one of the three primary areas of artificial intelligence (AI) research that contribute to Conversational AI, [5]. Speech recognition technology enables the conversion of spoken language (i.e., an audio signal) into written text [6], which is often used as a command, as in digital assistants [7]. ASR started with simple systems that responded to a limited range of sounds and have evolved into advanced systems that respond smoothly to natural language.

2.1 History

The history of Automatic Speech Recognition (ASR) begins approximately 70 years ago, in the 1950s [8], [9], [10]. Following the important moments of ASR history, the first research in voice recognition and synthesis emerged in the late 1950s. One of the initial major efforts was made by Bell Labs, which developed the “Audrey” system, [11], in 1952. This system could identify the numbers between 1 and 9 when spoken by a single voice.

Ten years later, IBM introduced the “Shoebox” system [12], which could understand and respond to 16 English words. Through collaborative research, by the late 1960s, technology could support words containing four vowels and nine consonants.

In the 1970s, significant steps were taken to improve the accuracy and speed of speech recognition systems. A notable example is Carnegie Mellon University’s “Harpy” [13], which could recognize around 1,000 words, roughly equivalent to the vocabulary of a three-year-old child. This system was developed under “The Speech Understanding Research” (SUR) program [14], which was run by the US Department of Defense and DARPA [15], and was one of the largest of its kind in the history of speech recognition. Another important advancement in the 1970s was Bell Laboratories’ introduction of a system capable of interpreting multiple voices.

In the 1980s, more sophisticated techniques were introduced, such as Hidden Markov Models (HMMs) [16], which allowed for more precise and flexible speech analysis. These techniques remain the foundation of many modern speech recognition systems.

In the 1990s, commercial applications of speech recognition systems emerged, with companies like Dragon Systems, launching products for personal computers. These systems began to be used in applications such as text transcription and telephone services. BellSouth introduced the Voice Portal (VAL), an interactive voice recognition (IVR) system [17], accessible by dialing in. This system laid the groundwork for many other phone tree systems still in use today.

Due to the surge in online data usage, the 2000s saw the integration of machine learning and deep learning techniques into ASR. For instance, Google used massive data sets and powerful computing resources to improve ASR accuracy. By 2001, ASR technology had achieved around 80% accuracy. For much of the decade, there was little advancement until Google launched Google Voice Search, [18]. As an app, speech recognition became accessible to millions of users. It was also significant because processing could be offloaded to Google’s data centers, reducing the processing load on users’ devices. Additionally, Google gathered data from billions of searches, aiding in predicting what a user was saying. At that time, Google’s English Voice Search System included 230 billion words from user searches.

Today, speech recognition and text synthesis systems are integrated into many robotic applications. Robots like Pepper [19] and NAO [20] use these systems to communicate with people in natural language. Technologies such as Apple’s Siri [21], Amazon’s Alexa [22], and Google Assistant [23] have made speech recognition and text-to-speech generation a crucial part of our daily lives.

2.2 ASR’s Tasks

An Automatic Speech Recognition (ASR) unit typically performs the following tasks [9], [10], [24], [25], [26]:

- It analyzes the incoming audio signal, extracting features such as pitch, frequency, and amplitude. This provides a representation of the audio signal that can be analyzed for further processing.
- It maps the extracted features to phonemes or word subunits. Traditionally, Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) were used for this task. However, more recent advancements have utilized Deep Neural Networks (DNNs) and Recurrent Neural Networks (RNNs), which have improved performance significantly.
- It runs language models, aiming to predict the likelihood of word sequences, which helps in understanding the structure and patterns of language. N-gram models have historically been used, but advanced neural network architectures are increasingly common now.

- It runs under Low Latency for Real-Time Response to provide timely, immediate transcriptions, which is essential for real-time applications.
- It manages in the best way the Vocabulary and Grammar, as the size and diversity of the vocabulary and grammar that the ASR system uses greatly affect its accuracy in recognizing and transcribing words.
- ASR systems use techniques to adjust to the unique vocal characteristics of different speakers.
- Capability in Multilingual Support that is provided to it, with training on a variety of datasets, representing different linguistic characteristics, adapting to the nuances of each language.

Among these tasks, there are various challenges in ASR's speech processing that must be addressed to avoid impacts on the system's accuracy and overall performance. The main challenges are:

- Background noise can hinder the ASR's ability to separate speech from ambient sounds, thus affecting accuracy.
- Variations in speech styles and pronunciations across speakers (Accents, Dialects, and Speech Variability) present challenges in achieving uniform accuracy.
- Not recognizing words or phrases that are not included in the training data, which is challenging in dynamic settings where new or specialized terms may be introduced.
- The Prosodic Features that means capturing the rhythm, intonation, and stress patterns of speech, which is complex yet critical for understanding the conveyed meaning.
- In the case of Limited Training Data, the training dataset is small and not representative of the target user population. Then the system may not generalize well to real-world scenarios.
- The management of real-time data processing requirements, while ensuring accurate transcription, remains a significant technical hurdle.
- It needs the capability for a smooth transition between languages, especially in multilingual contexts, being a challenge if the system is not specifically designed for multilingual support.

2.3 ASR Methods

Traditional speech recognition uses a genetic approach, simulating the entire process of producing sounds to evaluate a speech sample. It starts from a language model that includes the most likely sequences of words produced, such as an n-gram model [27], to a pronunciation model [28] for each word in the sequence (e.g., a pronunciation table), to an acoustic model that translates these utterances into sound waves (e.g., a Gaussian mixture model). Then, when it receives some audio input, it finds the most likely text sequence that would result from the given audio signal, according to the genetic process of models. Overall, traditional speech recognition attempts to model Pr(audio|transcript)*Pr(transcript), and take its argmax over possible transcripts [29].

With the evolution of neural networks, every element of the traditional speech recognition model could be replaced by a neural model, with better performance and greater generalizability. For example, we could replace an n-gram model with a neural language model and a pronunciation matrix with a neural pronunciation model. However, each of these neural models needs to be trained separately on different tasks, while errors in any model of the process can affect the entire prediction.

This led to the development of end-to-end ASR architectures. These are discrete models that simply take audio input and output text, where all architecture elements are trained together toward the same goal, [30]. The model encoder acts as an acoustic model to extract speech features, which can be directly fed to a decoder that extracts the text.

2.4 ASR on Operational Robots

To gain an insight into the current research interest, a search was carried out in the SCOPUS database, with the proposal "speech recognition operational robot", for the five-year period 2019-2024. The search engine returned just eight articles.

- In the 1st article [31], reference is made to the need for understanding the habits of people by collaborative robots in the industrial environment, for the better functioning of work groups. Speech is also included within this context.
- Similarly, the 2nd one [32] deals with collaborative robotics. More specifically, the article investigates the effectiveness and accuracy of voice interfaces, for operating robotic features through extremely low latency, associated with 5G network links intended for Arduino robots.
- The 3rd one [33] is a survey of technologies, enabling the design of a multimodal interactive robot for military communication. The author aims to identify existing automation capabilities in multimodal communication that can enhance

the proposed Interactive Robotic System (IRS), an AI-integrated robotic platform designed to improve the speed and accuracy of military operational and tactical decision-making.

- In the 4th one [34], the author proposes the use of efficient lightweight models for speech command classifiers, applied to human-computer interactions in robotic applications. This will avoid the current deep learning methodology with complex networks that require memory and energy.
- The 5th article [35] refers to the speech recognition and machine learning technologies that have been applied to the banking industry thanks to artificial intelligence, and how in the future artificial intelligence-related technologies will be applied to even more scenarios.
- In the 6th paper [36], the author aims to increase the intelligence of the humanoid agent Nao, using big data by activating multisensory perceptions, including auditory and speech-related stimuli. For this reason, he proposes a definition of artificial intelligence that focuses on enhancing Nao's learning and interaction capabilities.
- The research in the 7th paper [37] explores the application of embedded agent design methodologies, to develop a multimodal human-computer interface for managing the visualization of cyber events. This interface is primarily intended to assist security analysts and operators in visualizing cyber incidents or attacks, particularly when dealing with graphical information. Control methods for the visualization include visual gestures and voice commands.
- The 8th paper [38] informs us that while voice recognition systems provide a method of controlling robotic systems useful in law enforcement or military settings, users still need to learn the available commands and practice pronunciation to ensure accurate recognition. The results of a small pilot study indicate an overall preference for VR systems by users, despite their perception of the additional complexity and difficulty in learning to use them. It also suggests that the desktop mode was considered easier to use and users reported being more confident in using it.

From the above papers, it is largely clear that human-machine interfaces used in robotic applications for both operational and industrial purposes must be simple to use by humans, with high perception by machines, and lightweight in the use and consumption of resources.

3 System Development

As mentioned in the Introduction, the authors develop a robotic system with an advanced intelligent human-robot interface in Greek that is based on the communicative capabilities of the OMAS-III systemic conceptual model, [3]. Key parts of this interface are the functions of an ASR, for excellent speech-to-text creation, as well as a text-to-speech module, [10]. Both of these functions require applications that use advanced speech recognition algorithms that can detect and analyze voice data with high fidelity, machine learning technologies, and deep neural networks to improve accuracy, even in noisy environments, or when speakers are located in noisy environments and have different accents and intonations. Also, speech production must be natural and comprehensible, to ensure effective communication with users. Synthesized speech should sound human, with appropriate variations in pitch, rate, and emphasis, in Text to Speech (TTS) technology. In addition, the speed of response is critical for the success of the system, especially when it is used in dynamic environments or in applications where immediate response is essential.

All of the above requirements can be provided by standardized communication and application programming protocols (APIs) that can be easily integrated into different robotic systems. More specifically, the following will be used:

- Google Speech Recognition API [39], a service that enables voice-to-text conversion, providing accurate and fast voice recognition.
- Google Text-to-Speech API [39], a service for converting text to speech.

Python, a powerful programming language that provides a wealth of libraries and tools for the development and management of complex systems, will be used as a programming environment for the integration of the subsystems. Python is known for its simple and understandable syntax, which makes it ideal for rapid application development, while there are many libraries that can be used for various functions, such as audio processing, voice recognition, text-to-speech, and many more. It has a large community of users and developers, offering extensive documentation and support.

More specifically:

- Google’s cloud speech and cloud text-to-speech libraries, required to communicate with Google APIs, will be installed.
- API keys to access Google services will be generated and configured.
- The Google Speech Recognition API will be used, to record and recognize the user’s voice.
- The Google Text-to-Speech API will be used, to convert the recalled texts into speech.
- A unified system is being developed that will integrate both functions, allowing the user to communicate with the system through voice.

4 Integration Method / Programming and Testing Code

Aiming at APIs integration into the core code of the interface, extra code was developed in a very popular programming environment, which is the IDE Pycharm (community edition), [40]. The Python interpreter used is Pycharm’s default, which is the Python 3.10 interpreter.

4.1 Programming Libraries

The additional packages used are the following:

- Pydub 0.25.1
- Speech Recognition 3.10.4
- gTTS 2.5.1
- sounddevice 0.4.6
- numpy 1.26.4
- difflib 3.10.3.

In addition to the above, the ‘ffmpeg’ package [41] should be installed, as well as declared in the system’s PATH of the computer.

An uninterrupted internet connection is also a necessary condition for the operation of ASR.

Additionally, we check for the correct display of the following interpreters in Pycharm:

- PyAudio
- SpeechRecognition
- certifi
- cffi
- charset-normalizer
- click
- colorama
- gTTS (Google Text To Speech)
- idna
- numpy
- pip
- pycparser
- pudub
- requests
- setuptools
- sounddevice
- typing_extentions
- urlib3.

4.2 Basic Code

Firstly, a base loop is required to run the command: “check_mic_device_index()”. This returns the “device_index” number for Microphone array. This number will be included in the main command of the base loop as follows: “run_speech_to_text_greek(device_index=1,language=Language.GREEK)”.

The command calls the procedure that communicates through a microphone.

The following code is completed with the necessary commands and includes:

- Nine (9) Libraries
- Two (2) Classes in Language Selection (i.e., ENGLISH and GREEK)
- The Class with Standard Commands
- The Function Play Sound in Start up
- The Static methods
- The Microphone Determine Function
- The Speech-to-Text Function
- The Basic Loop

4.3 Testing Code

It should be noted at this point that the smart HRI developed by the authors, among others, has the ability to recognize and learn unknown words. As it has been documented in a previous work [42], when an unknown word enters the system, it then undertakes to start a meaningful dialogue with its user, until it forms the necessary knowledge and integrates the new word into its dictionary.

The ASR used in this work has an immediate response to the Greek language and with absolute accuracy in the performance of known Greek words, where it was tested. However, a test that is worth presenting in this work is the study of the behaviour of the ASR on completely unknown words. This test is of particular importance for the HRI, because in order to be able to learn new words, they must come as they sound and not through a “percentage” approach. To ensure that this advanced ASR does not know the input testing words and to check whether it will render them correctly and not approximately, an artificial language was used that the ASR certainly does not know. The language chosen is SostiMatiko, [43]. The choice was not made by chance, since this language is based on Greek roots, which is the language of HRI, but its writing is in Latin characters. The ASR is set to the

Greek language so that there is a close relationship between the roots of the natural and the artificial language. The words chosen were nouns and are the following:

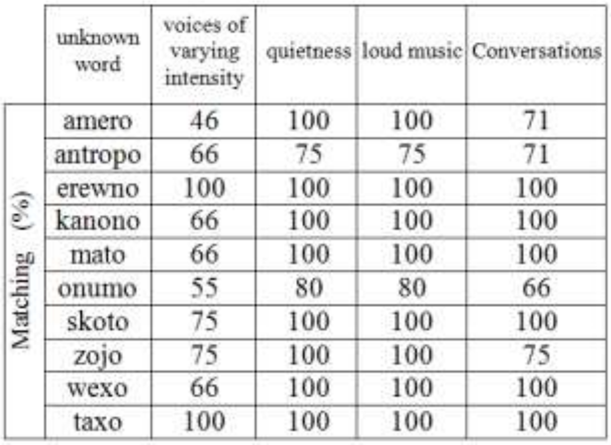
- *amero* which means ‘day’
- *antropo* which means ‘human’
- *erewno* which means ‘research’
- *kanono* which means ‘rule’
- *mato* which means ‘eye’
- *onumo* which means ‘name’
- *skoto* which means ‘darkness’
- *zojo* which means ‘animal’
- *wexo* which means ‘sound’
- *taxo* which means ‘speed’.

We also chose the test to be done in four different environments, which are:

- voices of varying intensity
- absolute silence
- loud music
- conversations.

The results are shown in Table 1.

Table 1. Results of ASR Testing in unknown words



	unknown word	voices of varying intensity	quietness	loud music	Conversations
Matching (%)	amero	46	100	100	71
	antropo	66	75	75	71
	erewno	100	100	100	100
	kanono	66	100	100	100
	mato	66	100	100	100
	onumo	55	80	80	66
	skoto	75	100	100	100
	zojo	75	100	100	75
	wexo	66	100	100	100
	taxo	100	100	100	100

As it can be seen from the above results, the ASR cannot respond well to the “voices of varying intensity” environment. The words from the environment are mixed with those of the test, resulting in a long delay in the response and many errors. On the contrary, in the “absolutely quiet” environment we have the greatest accuracy. The deviations that appear in two words exist due to an incorrect letter because the corresponding natural language words differ only in this letter, and there they sound about the same. It should be noted here that the response time is immediate. In the “loud music” environment we have corresponding results because there are no words to confuse the system. Here too, the response time was immediate. In the last environment “Conversations” the results are not too bad. In the case of conversations, a more stable intensity prevails than in the first case; it is enough for the speaker to give the word louder so that the system can ignore the lower noises. However, there is a delay in the response, as the system tries to capture more words.

5 Discussion

The code developed in Chapter 4 is the first experimental code for testing functionality and correctness. It is also under extensive testing to configure and incorporate functions for autonomy purposes. In the given code, there is only one operation cycle at a time and, after an audio signal, it waits for a short time until it receives the command given. Then it checks this command with a set of existing ones and results in the threshold. With this code, ASR was evaluated for its response to 10 unknown words in an unknown language. The results of this experimental code are quite good (Table 1).

The interface that has been developed provides the ability for Natural Language Understanding since each incoming sound becomes a word and is sent for analysis. In the further evolution of the code, there will be no commands but words, and every incoming sound will be checked every moment. This direction of development is consistent with the results of the previously mentioned small pilot study [38] that indicate an overall preference for speech recognition systems by the users, despite their perception of the additional complexity and difficulty in learning to use them. In this respect, the commands given in words of a natural language will make the robotic system herein more user-friendly.

In a previous work [44], algorithms for text-to-speech generation, speech formation, and transfer algorithms for people with disabilities [3] have already been developed. The present work completes these processes.

6 Conclusion

In this paper, we have presented an attempt to run an advanced ASR on an initially experimental code. The results obtained are quite satisfactory.

The code is being developed with the sole purpose of freeing a robot from the initially defined proposals and making it fully autonomous from the controlled hearing function.

The system should at all times seamlessly receive information from its environment. It will then be possible to integrate it into the developing robotic system, designed by the authors.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the authors used Google Translate services in order to improve the readability and language of their manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

[1] Giachos I., Papakitsos E.C., Savvidis P. and Laskaris N., Inquiring Natural Language Processing Capabilities on Robotic Systems through Virtual Assistants: A Systemic Approach, *Journal of Computer Science Research,* Vol. 5(2), 2023, pp. 28-36. https://doi.org/10.30564/jcsr.v5i2.5537.

[2] Tellex S., Gopalan N., Kress-Gazit H. and Matuszek C., Robots That Use Language, *Annual Review of Control, Robotics, and Autonomous Systems,* Vol. 3, 2020, pp. 32-55. https://doi.org/10.1146/annurev-control-101119-071628.

[3] Giachos I., Batzaki E., Papakitsos E. C., Papoutsidakis M. and Laskaris N., Developing a Natural Language Understanding System for Dealing with the Sequencing Problem in Simulating Brain Damage**,** *WSEAS Transactions on Biology and Biomedicine,* Vol. 21, 2024, pp.138-147. https://doi.org/10.37394/23208.2024.21.14.

[4] Symeonaki E., Arvanitis K., Piromalis D. and Papoutsidakis M., Conversational User Interface Integration in Controlling IoT Devices Applied to Smart Agriculture: Analysis of a Chatbot System Design, In: Bi, Y., Bhatia, R., Kapoor, S. (eds.), Intelligent Systems and Applications; IntelliSys 2019; *Advances in Intelligent Systems and Computing, Springer, Cham,* Vol. 1037, 2020, pp. 1071–1088. https://doi.org/10.1007/978-3-030-29516-5_80.

[5] Zijun X., Ruirui L., and Mingda L., Recent Progress in Conversational AI, arXiv: 2204.09719, 2022, p. 6. https://doi.org/10.48550/arXiv.2204.09719.

[6] Chen X., Rong Y., Qianqian D., Chengqi Z., Tom K., Mingxuan W., Tong X. and Jingbo Z., *Recent Advances in Direct Speech-to-text Translation,* arXiv:2306.11646, 2023, p. 10. https://doi.org/10.48550/arXiv.2306.11646.

[7] Luca A. H and Delphine R., A survey on privacy issues and solutions for Voice-controlled Digital Assistants, *Pervasive and Mobile Computing,* Vol. 80, 2022, p. 11. https://doi.org/10.1016/j.pmcj.2021.101523.

[8] Kincid J., A Brief History of ASR: Automatic Speech Recognition, Medium, 2018, [Online]. https://medium.com/descript/a-brief-history-of-asr-automatic-speech-recognition-b8f338d4c0e5 (Accessed Date: October 27, 2024).

[9] Shaip, A Comprehensive Overview of Automatic Speech Recognition (ASR), Medium, 2023, [Online]. https://weareshaip.medium.com/a-comprehensive-overview-of-automatic-speech-recognition-asr-2b208aae0305 (Accessed Date: November 2, 2024).

[10] Lefkelis V.S., *Development of a speech-to-text and text-to-speech system for robotic applications,* Diploma Thesis, Department of Industrial Design and Production Engineering, School of Engineering, University of West Attica, Athens, Greece, 2024. http://dx.doi.org/10.26265/polynoe-7162.

[11] Knight P., Smart Speaker, tell me about your acoustic sensor, *Physics World,* Vol. 33, No 12, 2021, p. 25. https://dx.doi.org/10.1088/2058-7058/33/12/27.

[12] Glantz R.S., SHOEBOX: a personal file handling system for textual data, fall joint computer conference (AFIPS '70 (Fall)), Houston, Texas, 1970, pp. 535–545. https://doi.org/10.1145/1478462.1478541.

[13] Lowerre B. P. and Reddy B. R., Harpy, a connected speech recognition system, *Acoustical Society of America,* Vol. 59(1), 1976, p. 97. https://doi.org/10.1121/1.2003013

[14] Furui S., *History and Development of Speech Recognition, in Speech Technology,* Springer, 2010, pp. 1-18. https://doi.org/10.1007/978-0-387-73819-2_1.

[15] Bonvillian W.B. and Van Atta R., ARPA-E and DARPA: Applying the DARPA model to energy innovation, *The Journal of Technology Transfer*, Vol. 36(5), 2011, pp. 469-513. https://doi.org/10.1007/s10961-011-9223-x.

[16] Eddy S.R., Hidden markov models, *Current opinion in structural biology,* Vol. 6(3), 1996, pp. 361-365.

[17] Corkrey R. and Parkinson L., Interactive voice response: Review of studies 1989–2000, *Behavior Research Methods, Instruments, & Computers,* Vol. 34(3), 2002, pp. 342-353. https://doi.org/10.3758/BF03195462.

[18] Donepudi P.K., Voice search technology: an overview, *Engineering International,* Vol. 2(2), 2014, pp. 91-102. https://doi.org/10.18034/ei.v2i2.502.

[19] Pandey A. K. and Gelin R., A Mass-Produced Sociable Humanoid Robot: Pepper: The First Machine of Its Kind, *IEEE Robotics & Automation Magazine*, Vol. 25(3), 2018, pp. 40-48. https://doi.org/10.1109/MRA.2018.2833157.

[20] Gouaillier D., Hugel V., Blazevic P., Kilner C., Monceaux J., Lafourcade P., Marnier B., Serre J. and Maisonnier B., The nao humanoid: a combination of performance and affordability, arXiv:0807.3223v2, 2008, p. 10. https://doi.org/10.48550/arXiv.0807.3223.

[21] Bellegarda J.R., *Spoken Language Understanding for Natural Interaction: The Siri Experience, Natural Interaction with Robots, Knowbots and Smartphones,* Springer, New York, NY, 2013, pp. 3–14. https://doi.org/10.1007/978-1-4614-8280-2_1.

[22] Lopatovska I., Rink K., Knight I., Raines K., Cosenza K., Williams H., Sorsche P., Hirsch D., Li Q. and Martinez A., Talk to me: Exploring user interactions with the Amazon Alexa, *Journal of Librarianship and Information Science,* Vol. 51(4), 2018, pp. 984-997. https://doi.org/10.1177/0961000618759414.

[23] Michaely A. H., Zhang ., Simko G., Parada C. and Aleksic P., Keyword spotting for Google assistant using contextual speech recognition, *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Okinawa, Japan, 2017, pp. 272-278. https://doi.org/10.1109/ASRU.2017.8268946

[24] Yu D. and Deng L., Introduction. *In: Automatic Speech Recognition, Signals and Communication Technology,* Springer, London, 2015, pp. 1–9. https://doi.org/10.1007/978-1-4471-5779-3_1.

[25] Alharbi S., Alrazgan M., Alrashed A., Alnomasi T., Almojel R., Alharbi R., Alharbi S., Alturki S., Alshehri F. and Almojil M., Automatic Speech Recognition: Systematic Literature Review, *IEEE Access*, Vol. 9, 2021, pp. 131858-131876. https://doi.org/10.1109/ACCESS.2021.3112535.

[26] Foster K., What is Automatic Speech Recognition? A Comprehensive Overview of ASR Technology, AssemblyAI, (2023), Accessed: Nov. 1, 2024, [Online]. https://www.assemblyai.com/blog/what-is-asr/.

[27] Siu M., and Ostendorf M., Variable n-grams and extensions for conversational speech language modeling, *IEEE Transactions on Speech and Audio Processing*, Vol. 8(1), 2000, pp. 63-75. https://doi.org/10.1109/89.817454.

[28] Fosler-Lussier J.E., *Dynamic pronunciation models for automatic speech recognition,* University of California, Berkeley, 1999.

[29] Djeffal N., Kheddar H., Addou D., Mazari A. C. and Himeur Y., Automatic Speech Recognition with BERT and CTC Transformers: A Review, *2nd International Conference on Electronics, Energy and Measurement (IC2EM*), Medea, Algeria, 2023, pp. 1-8. https://doi.org/10.1109/IC2EM59347.2023.10419784.

[30] Li B., Chang S.Y., Sainath T.N., Pang R., He Y., Strohman T. and Wu Y., Towards fast and accurate streaming end-to-end ASR, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Barcelona, Spain, 2020, pp. 6069-6073. https://doi.org/10.1109/ICASSP40776.2020.9054715.

[31] Lee W.W.L., The future of collaborative robotics AI in Industry 5.0: An academic perspective with a practice approach, *10th International Conference on Socio-Technical Perspectives in Information Systems, STPIS* 2024, Hybrid, Jongkoping, Sweden. Vol. 3857, 2024, pp. 172 – 181.

[32] Abdulrazzaq A.Z., Ali Z.G., Mohammed Al-Ani A.R., Mohammed Khaleel B., Alsalame S., Snovyda V. and Kanbar A.B., Evaluation of Voice Interface Integration with Arduino Robots in 5G Network Frameworks, *Conference of Open Innovations Association FRUCT 2024* Helsinki, 2024, pp. 44-55.

https://doi.org/10.23919/FRUCT64283.2024.10749856.

[33] Sheuli P., A survey of technologies supporting design of a multimodal interactive robot for military communication, Journal of Defense Analytics and Logistics, Emerald Publishing Limited, Vol. 7(2), 2023, pp. 156-193. https://doi.org/10.1108/JDAL-11-2022-0010.

[34] Soltanian M., Malik J., Raitoharju J., Iosifidis A., Kiranyaz S. and Gabbouj M., Speech Command Recognition in Computationally Constrained Environments with a Quadratic Self-Organized Operational Layer, *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, 2021, pp. 1-6. https://doi.org/10.1109/IJCNN52387.2021.9534232.

[35] Li X., Application and influence of artificial intelligence technology in commercial banks, *2021 2nd International Conference on Computer Science and Management Technology (ICCSMT)*, Shanghai, China, 2021, pp. 455-458. https://doi.org/10.1109/ICCSMT54525.2021.00089**.**

[36] Baothman F.A., An Intelligent Big Data Management System Using Haar Algorithm-Based Nao Agent Multisensory Communication, *Wireless Communications and Mobile Computing,* Vol. 2021, 2021, pp. 1-15. https://doi.org/10.1155/2021/9977751.

[37] Szynkiewicz W., Kasprzak W., Zieliński C., Dudek W., Stefańczyk M., Wilkowski A. and Figat M., Utilisation of Embodied Agents in the Design of Smart Human–Computer Interfaces—A Case Study in Cyberspace Event Visualisation Control, *Electronics,* Vol. 9(6), pp. 1-36. https://doi.org/10.3390/electronics9060976.

[38] Carruth D.W., Hudson C.R., Bethel C.L., Pleva M., Ondas S. And Juhar J., Using HMD for Immersive Training of Voice-Based Operation of Small Unmanned Ground Vehicles, In: Chen, J., Fragomeni, G. (eds.), Virtual, Augmented and Mixed Reality, Applications and Case Studies, HCII 2019, Washington, DC, USA, *Lecture Notes in Computer Science,* Springer, Cham, Vol. 11575, 2019, pp. 34-46. https://doi.org/10.1007/978-3-030-21565-1_3.

[39] Google Cloud Documentation, Google, [Online]. https://cloud.google.com/docs (Accessed Date: November 3, 2024).

[40] Pycharm IDE, [Online]. https://www.jetbrains.com/pycharm/download/?section=windows (Accessed Date: November 3, 2024).

[41] FFmpeg, [Online]. https://www.jetbrains.com/pycharm/download/?section=windows (Accessed Date: November 3, 2024).

[42] Giachos I., Papakitsos E.C., Antonopoulos I. and Laskaris N., Systemic And Hole Semantics In Human-Machine Language Interfaces, *2023 17th International Conference on Engineering of Modern Electric Systems (EMES)*, Oradea, Romania, IEEE, 2023, pp.1-4. https://doi.org/10.1109/EMES58375.2023.10171635.

[43] SostiMatiko, [Online]. https://linguifex.com/wiki/User:SostiMatiko (Accessed Date: October 24, 2024).

[44] Giachos I., Batzaki E., Papakitsos E. C., Kaminaris S., and Laskaris N., A Natural Language Generation Algorithm for Greek by Using Hole Semantics and a Systemic Grammatical Formalism, *Journal of Computer Science Research,* Vol. 5(4), 2023, pp. 27–37. https://doi.org/10.30564/jcsr.v5i4.6067.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed to the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)