Conversational agent in HCI a review

Yixuan Zhao

Electronics and Information, Guangdong Normal University of Technology,
Guangzhou, 510450, China

zhaoyixuan_137@qq.com

Abstract. Nowadays, AI technology is developing rapidly and slowly covering all areas of life. AI facilitates different parts of our lives and provides us with a lot of help. For example, in learning, students can acquire knowledge more conveniently through AI, and buyers can find suitable goods more conveniently through AI when shopping. At the same time, more and more technology is used in the field of voice agents, which allows humans to enjoy a lot of better services. In this article, we will study to better understand "how to understand human natural language and how the repository of knowledge is built." In the article we build with examples and deep learning models (CNN and RNN) through databases. Through repeated research and analysis, we can find that there are some limitations in this paper, such as the single learning model and the insufficient elaboration of data analysis and signal system technology. But at the same time, we also found a lot of future application prospects that voice agents can develop, it can be applied in many fields, such as finance, medical care, education and so on. For example, in children's education, parents can use voice agents to set time limits and monitor their children's progress. Existing digital interactive storytelling systems have limitations in terms of available storybooks and hand-crafted issues. Voice agents are becoming more popular in everyday scenarios, and more users are adopting devices like Siri and Google Assistant. In the future, conversational agents are expected to play an important role in oral communication with users.

Keywords: Conversational agent, CNN, RNN, Artificial intelligence

1. Introduction

In today's era of rapid technological development, voice agents have become a hot topic in the field of artificial intelligence. From Google Assistant to the iPhone's Siri, voice assistants are playing an increasingly important role in our lives [1]. Through speech recognition and natural language processing technology, they are able to understand human speech and hold conversations with people, greatly improving the human-computer interaction experience [2]. In the process of reference and practice, we have the following findings:

The emergence of voice agents is not only to provide convenient services, but also to improve the way people and machines communicate. The traditional way of human-computer interaction is usually operated by keyboard or mouse, and the appearance of voice agent allows people to interact with the machine through natural language, which greatly improves the user's operation efficiency and experience [3].

In addition to improving the human-computer interaction experience, voice agents can also improve operational efficiency. They can provide intelligent services by automating answers to frequently asked questions, providing real-time data and recommending relevant content [4]. For example, in the customer service industry, voice agents can greatly reduce the workload of manual customer service, improve customer satisfaction and service efficiency.

In the study of the development and application of voice agents, user experience and acceptability are crucial considerations. One of the goals of voice agents is to provide user-friendly interaction and ensure that users can easily obtain the required information and complete tasks. Therefore, it is particularly important to conduct in-depth research on users' perceptions, attitudes, and satisfaction with voice agents [5]. The user experience not only includes perception of the interaction process of voice agents, but also factors such as ease of use, reaction speed, and error handling in interface design. By carefully studying user feedback and needs, the design and interaction of voice agents can be improved, thereby improving the quality of user experience. At the same time, it is also crucial to evaluate the user's acceptance of voice agents, understand the user's acceptance of this interaction method, and determine the applicability of voice proxies in specific fields or scenarios.

Multimodal interaction is the combination of multiple interaction methods to enrich the user experience and provide greater flexibility. Voice agents can be integrated with other interaction methods, such as touch screens, gestures, and visuals, to achieve a richer and more intuitive interaction experience [6]. By integrating different perception channels, multimodal interaction can provide more choices and more precise interaction methods to meet the different preferences and usage environments of users. For example, combining image recognition in speech agents can achieve object recognition, image search, and other functions, enhancing the functionality and intelligence of the system.

Privatization and personalization are also important directions for the development of voice agents. Voice agents can provide customized services and suggestions to users by learning their preferences, habits, contextual information, and other personalized features. The key to studying personalized and personalized services is effective data collection and analysis mechanisms. As more and more user data become available, how to protect user privacy, ensure the legitimate use and security of data, is also a key consideration. At the same time, it is also necessary to consider how to design algorithms and models to effectively infer users' personalized needs and provide corresponding services.

The development of voice agents has also promoted the advancement of artificial intelligence technology. To be able to better understand and answer human questions, voice agents need to have strong speech recognition, natural language processing, and machine learning capabilities. Therefore, the research and development of voice agents involves the research of speech recognition technology, natural language processing technology, machine learning algorithms and other fields.

However, the development of voice agents also faces some challenges. First, there are still some problems with the accuracy and stability of speech recognition technology, especially when it comes to noisy environments and accents. Secondly, natural language processing technology is faced with complex problems such as semantic understanding and contextual reasoning. In addition, user privacy and data security are also a big problem in the development of voice agents.

Related to this, Conversational Agents Replying with a Manzai-style Joke and Building a storytelling conversational agent through The parent-AI collaboration articles provide some interesting research directions. The former introduces a joke reply method in the style of Japanese crosstalk, which increases the interactive experience of users through the way of dialogue. The latter explores the collaboration between parents and AI by combining the parent's story with the AI's analytical power to achieve an interesting storytelling voice agent.

In short, as an important research direction in the field of artificial intelligence, voice agents can not only improve human-computer interaction experience, improve operational efficiency, but also promote the development of artificial intelligence technology. However, to achieve better voice agents, further research and technological innovations are still needed in areas such as speech recognition, natural language processing and data security. It is believed that in the near future, voice agents will play a more important role in various fields.

2. Literature review

In total, there are about four steps about the component. The first step is to introduce the conversational machine and the mainly used techniques, the second part is to list the common features of chatbot and analysis. Meanwhile, the next step is to give example and analysis the advantage and disadvantage. In conclusion, summarizing the result of chatbot's experiment is crucial. The components of the model are as follows. Firstly, speech input. The voice receives the user's voice input, which can be done through a microphone or other voice capture device [1]. Secondly, Automatic Speech Recognition. Voice agents use speech recognition technology to convert users' speech input into understandable text form. This involves capturing sound signals from the surroundings using microphones or other audio devices. Thirdly, Natural Language Understanding. At this stage, the program processes the text to understand the user's intentions. Technologies include language modeling, entire recognition, and entity analysis to help understand user input. This may involve using machine learning models to convert the sound signals into textual form for further processing and analysis. Fourth, Dialogue Management. It is respondable for handling interactions with users. It involves determining the system's response and generating appropriate responses. Also, it is based on rules, machine learning or deep learning models to support dialogue control for different tasks. Fifth, it's Dialogue Generation [2]. In this step, the voice agent generates a natural language conversation response based on the instructions of the conversation management module. Dialogue generation involves selecting appropriate dialogue templates, natural and fluent sentences, and responding to user needs. Sixth, Speech Synthesis. The process of synthesizing speech involves converting the text generated by the conversation into auto speech output. Speech synthesis technology uses proper text models and voice libraries to generate realistic speech output. Seventh, it's Multimodal Interaction. This step discusses multimodal interaction with voice agents, including the use of forms such as postures, facial expressions, and gestures to enhance user experience and agent performance. They analyze other signals and data from the environment, such as camera images or sensor data, to understand the current context and generate appropriate behavioral responses. Eighth, speech output. This step will be the dataset output which includes multimodal interaction, well, but mostly from the voice message.

After we introduce the component of voice agents, maybe we will discover the deeper path about its special points. Next, we are going to show you its common challenge.

3. The common challenges

In my point of view, the most complex points for me can be listed in two parts: Nutural language understanding and knowledge base creation. Well, the first challenge must be the natural language understanding. As is known to us that It is difficult to recognize and understand the natural language. The reason why natural language processing (NLP) is difficult to understand is because human language has a high level of complexity and diversity. People use rich grammar rules, vocabulary selection, contextual understanding, and inference of meaning when using language. These factors make accurate understanding and processing of language challenging [3]. In voice agent, the difficulties of NLP are reflected in the following aspects:

The first one is the Speech recognition accuracy, Speech agents first need to convert speech signals into text information. However, speech recognition technology faces challenges such as audio noise, unclear pronunciation, accents, and changes in speech speed, which can affect the accuracy of recognition. Therefore, correctly recognizing speech input is the first major challenge in NLP. Semantic understanding is also the process of extracting meaning and intention from text. It requires a deep understanding of the differences in context, structure, and vocabulary. For a voice agent, it needs to convert the user's voice input into accurate intentions and commands to correctly respond to user needs. And then it‘s the Context Understanding and Reasoning. Language is usually contextual, and understanding a sentence requires consideration of the content and context of the preceding and following text. The difficulty lies in accurately grasping important information and conducting logical reasoning to generate correct responses. In voice agents, this means recognizing previous conversation records, understanding the user's intentions, and responding accordingly. Well, but we can still find another way to figure out what the problem is. These are two crucial steps besides language models and entity recognition: the first one is the Intent Classification: In the paper, the author use semantic understanding, which refers to the process of transforming user input into actionable representations. It involves transforming natural language text to the level of meaning, so that agents can understand user intentions and perform corresponding operations. In another paper, the author also came up with an idea called “Context Parsing and Tracking” [4]. Voice agents are able to parse and track information in conversations to ensure accuracy. Context resolution and tracking enable agents to understand previous conversation history and user intentions, providing more accurate responses for other conversations.[5]

After that, the second complex point is the Knowledge Base Creation (it is difficult to achieve human-like) Voice agents cannot fully simulate human behavior because human behavior involves complex factors such as emotions, social interaction, and body language in addition to language expression. Voice agents currently mainly focus on language communication, but cannot fully simulate human emotions, experiences, and behavioral patterns. In terms of knowledge base creation, voice agents also have some limitations. The intent describes what the messages are, for example, for a transportation network bot, the question: “I want to know taxi rate in Islamabad.” has a “request_rate_taxi” intent. It would help to summarize the information and input the keyword [6]. Entities are created only if these are included in the message otherwise it will be left as an empty array. Entities are important for the bot to understand what precisely a user is asking about. Maybe the most clear part is the limitations of knowledge. The creation of a knowledge base requires a large amount of data and information sources. Voice agents may not be able to directly access and organize information from various sources, such as the internet, social media, and professional databases. This will limit the content and update speed of the knowledge base. Information integrity and accuracy may also be a big problem. The quality of the knowledge base depends on the completeness and accuracy of the information [5]. Voice agents may face some challenges when organizing and storing information, such as mixed information, incorrect classification, and ambiguous interpretation. Updating and filtering information also requires a certain labor cost. Furthermore, Knowledge structure and links is also a crucial part. Building a complete knowledge base requires establishing accurate knowledge structures and links. Voice agents may encounter difficulties in handling complex semantic relationships and knowledge levels. This may lead to limitations in providing accurate and in-depth knowledge answers for voice agents [8]. To address these limitations in knowledge base creation, here are two strategies. The first of all is the Multichannel data collection which was conveyed by a author in my read paper. Collaborate with various sources such as specialized databases, academic resources, and social media platforms to gather diverse data and information, augmenting the content of the knowledge base. Meanwhile, the another author of the paper highlight that Human-machine collaboration is excellent. Combine human expertise and machine intelligence to curate, verify, and enhance the knowledge base [7]. Employ machine learning and natural language processing techniques to assist in automating the extraction and updating of relevant knowledge.

By implementing these strategies, we can overcome the limitations and challenges.

4. Deep Learning Model Introduction

There is a detailed introduction to two "crazy killing" learning models in the field of voice proxy. It includes two learning models about their principles, each step of the application process, and analysis of data object themes. These two learning models are CNN and RNN, Here are some information about them. First of all, CNN is a neural network model used to process images. By continuously moving and observing local areas on the image, meaningful features are sought (feature extraction) and then integrated to find the overall features. CNN can apply the principle of sound signals and extracting features to find features such as audio frequency in time series.[1]

Second, RNN is a neural network model that excels in processing sequence data. It can remember previous information and apply it to the current situation. This memory ability makes RNN very useful in processing language and audio sequences. Also, RNN can understand continuous speech or text inputs, consider previous contextual information, establish memory of past information, and consider contextual relationships when processing language and audio sequences. Otherwise, these two deep learning models are mostly used in these steps. First of all, it's sound input. CNN will help us analyze various details of sound signals, just as we do when observing images. It can extract different features from audio, such as high and low pitches, and even tell us which instrument's sound is. Second, it's sound recognition. RNN helps us convert sound into text. It will read sound features and generate corresponding text. In this way, we can convert the conversation content into readable text. Furthermore, it's natural language understanding. It helps us understand the intention and information in the conversation [3]. By memorizing previous conversations, RNN can analyze the structure of sentences, the meaning of vocabulary, and extract semantic information. In this way, the agent can better understand our meaning and provide correct answers. Last but not least, Dialogue management. It can help agents remember previous conversation history and generate coherent responses based on context [2].

In summary, the speech translation system uses CNN to extract audio features. For 1D audio sequence data, CNN can extract different frequency and time features by using different sizes of kernel functions and filters, and convert the audio signal into a 2D heat map.[6] This transformation is helpful for subsequent feature extraction and analysis. RNN performs voice text translation, text understanding models perform context processing, dialogue management techniques organize dialogue processes, and fine-tuning techniques optimize [5]. By using BERT or LSTM based methods, RNN can model long-term dependencies and contextual information, improving the accuracy and performance of voice text translation. Thus, the translation and comprehension tasks from audio to text were achieved. The combination of these technologies enables the system to better adapt to different speech and dialogue scenarios, and provide more accurate and coherent speech translation and response. The combination of these technologies enables the system to better adapt to different speech and dialogue scenarios, and provide more accurate and coherent speech translation and response [8].

5. Dataset

Datasets can hold information such as medical records or insurance records, to be used by a program running on the system. People mostly would apply the dataset to do the calculation.

Table 1. The comparison between three common models:

Name of the dataset	public/private	size	Applicable model
dataset1	public	range between 64 × 64 and 256 × 256	CNN
dataset2	public	the number of features in the input data that are fed into the network at each time step.	RNN
dataset3	public	the number of samples processed in one forward and backward pass during training.	Transformer

There are some advantage and disadvantage of applying CNN model. First of all, it's Spatial Hierarchy. CNNs are very good at capturing the spatial hierarchy of features. In images, this means they can learn to detect simple features like edges and textures in lower layers, and combine them to detect complex patterns like shapes and objects in higher layers [4]. This hierarchical feature extraction is particularly valuable for image analysis. It's very effective on image and video data as well. Because of its high accuracy, CNNs are state-of-the-art in a variety of computer vision tasks, including image classification, object detection, image segmentation, and even video analysis. However, the most challenging limitation of CNN is Big Data Requirements [5]. Because CNNs typically require large amounts of labeled training data to perform well. That would increase the pressure of CNN. What's else, memory usage would be also a problem. Deep CNN architectures would spend a lot of memory during training and inference. This may limit their deployment on resource-constrained devices.

Except CNN, there are also some advantage and disadvantage of applying RNN model. One of the most benefit one is Sequential information processing. RNN is suitable for processing sequence data and can capture the evolution process of sequence information, so that makes them suitable for tasks where the order of data points is important, such as speech recognition, language modeling, and video analysis [5]. Furthermore, Parameter sharing is also the meaningful feature of RNN. In RNNs, the same set of weights are used at each time step, which means fewer parameters need to be trained compared to some other sequence models. That would increase its efficiency. But, there are also some disadvantage. Coming first is Short-term memory. Traditional RNNs have limited short-term memory, so they may struggle to capture long-term dependencies in sequences. This limitation has led to the development of more advanced RNN variants such as LSTM (Long Short Term Memory) and GRU (Gated Recurrent Unit) [4]. What's more, Limited parallelization. Compared with CNN, RNNs inherently process sequences sequentially, which limits their parallelization during training and inference. So, that would obviously cost more time for RNN to parallelize the producing process [4].

The third common model of dataset is the Transformer. There are also some advantage and disadvantage of applying the Transformer model. First of all, it's Parallelization. Transformers can process input data in parallel rather than sequentially, which makes them very efficient, especially when processing long sequences [8]. That is similar to CNN. This parallelization reduce the training and inference time period, resulting in faster model development and deployment. It also has strong modeling capabilities and is suitable for processing various data forms such as sequences and images. However, Large memory footprint would be one of the disadvantages of it [7]. The reason is that the parameters of pretrained Transformer models are very large, making them challenging to deploy on resource-constrained devices or scenarios with limited memory. Because of the large memory footprint, it would also lower the efficiency of Transformer, especially when pretrained on large corpora, can require large amounts of data. In addition, Interpretable but complex. While Transformers can achieve impressive results, their inner workings can be complex and difficult to explain, that would increase the understanding difficulty.[9]

Table 2. The features of dataset:

	Feature 1	Feature 2	Feature 3
Data point 1	Data Storage: it refers to the use of recording media to retain data using computers or other devices. Maintly for saving the data.	Data Retrieval: it is a process that locates, extracts and presents information from data repositories.	Data Replication and Clustering: it is an ongoing process that creates replicas of all business-critical data and applications, synchronizes the data, and distributes it across a network
Data point 2	Data Backup and Recovery: it is the practice of duplicating your organization's data to ensure its protection in any type of data loss event. That would help to make sure the model could use the data again and again.	Data Security: it is the process of safeguarding digital information throughout its entire life cycle to protect it from corruption, theft, or unauthorized access. That would be important to protect the model research.	Concurrency Control: it is a procedure in DBMS which helps us for the management of two simultaneous processes to execute without conflicts between each other. That would help to achieve what do CNN and transformer can do

6. Application Background

In the field of voice agency, we can also gain a detailed understanding of its shining side in different fields. This time, we will elaborate on the process and results of investment in the fields of finance, healthcare, and education, respectively. The first one is Financial field. Voice agents have a wide application background in the financial field [1]. On the one hand, voice agents can be used for voice trading, allowing users to perform trading operations such as buying and selling through voice commands, improving trading efficiency [2]. On the other hand, voice agents can also be used for customer service and support, providing personalized suggestions and solutions, answering user questions, processing account information, and more. This helps to provide a better user experience, improve customer satisfaction, and reduce operational costs. The second one is Medical field. In the medical field, voice agents have an important application background. Voice agents are used for medical diagnosis and monitoring, converting conversations between doctors and patients into text through speech recognition technology, helping doctors diagnose and develop treatment plans more [3]. At the same time, voice agents can also be used to provide medical consultation and advice, answer patients' questions, and provide health management guidance. In this way, patients can more conveniently obtain professional opinions and promote communication and cooperation between doctors and patients. Furthermore, it's Education**.** The application of voice agents in the field of education is also receiving increasing attention. Voice agents can provide online learning support and personalized education services for students. Students can gain answers to questions, learning advice, and guidance through conversations with voice agents, helping them better understand and master knowledge. At the same time, voice agents can also provide educational games and interactive experiences, stimulating students' interest and motivation in learning [3]. In addition, voice agents can also be used for communication and collaboration between teachers and students, providing immediate feedback and evaluation, and promoting the improvement of teaching effectiveness. To sum up, voice agents have a wide range of application backgrounds in the fields of finance, healthcare, and education. They play an important role and generate more innovation and value in various fields [1].

7. Analysis

In the analysis process, it would focus on delving deeper into the problems. There is some information that identified in these papers and analyze them in detail, exploring the limitations of the paper for various research directions and relevant solutions. Here is the following information. In the first aspect, it is about Machine Learning Model Gap. We only briefly introduced the application of CNN and RNN neural network models in each step,[6] but did not deeply analyze the operational mechanisms of these two models themselves. And the article also did not attempt to explore other deep learning models, such as the Transformer, and lacked a comparison that is more suitable for speech adoption compared to other models (theoretical support)**.** In the second aspect, it's about Domain problem related research gap [8]. In fact, the application scenarios they mentioned are too concise. As they only briefly discuss the theoretical issues of voice agents and do not excessively explain them in practice, the content in various fields is relatively rough and lacks innovation. Third aspect is the lack of certain content mention about data handled information. It shows on the focus of the article is to introduce the design methods and technological advancements, it didn't mention enough knowledge about the signal system and data processing. Also, it lacks sufficient reference to the theory of the dataset and does not provide specific information on various aspects of the dataset. It just simply mentions the origin and simple application of its scale.

Then there are suggestions for fill out the gap and solve the limitation. First of all, about the learning model problem. When it comes to the issue of learning models, the two most prominent gaps are that the two neural network models involved in CNN and RNN only briefly explain their applications in the process and do not specifically analyze the work completed by the internal mechanisms of CNN and RNN [7]. In fact, discussions can be carried out in a specific structural form: for example, how the pooling part of CNN's structure complete their work in speech proxies, Furthermore, a detailed explanation of how the internal operating mechanism of RNN recurrent neural networks in the field of conversation agents can be achieved by introducing recurrent connections, introducing memory units in the network, transmitting state information, capturing long-term dependencies in sequences, and processing temporal data is more meaningful. When it comes to the lack of sufficient theoretical support for comparing other deep learning models, some scholars inevitably question whether other models such as the Transformer model can complete this task. Why can only CNN and RNN be mentioned, or whether these articles mention that other deep learning models can work together? These can all be explained briefly, Helps to better understand the content of the model. Second, about the domain expansion issues. When it comes to the issue of domain deployment, it's actually quite easy to say that there is a lack of specific explanations for application scenarios [9]. Perhaps the pursuing is not the result of investment in the future and don't like to only see its functions. We may be more curious about how the functions displayed in each domain are implemented through internal entity structures. This allows for a deeper understanding of how the field unfolds, such as in the financial field, how voice agents analyze investors' voices and provide certain data processing to predict future financial development trends; How to assist children in learning in the education industry by conducting data analysis through conversations with children and accurately delivering knowledge about connections; How to obtain relevant data through patient voice messages and provide assistance to patients in the medical industry, and so on! These can all be explained in detail, the more detailed the better. Third, focus on Dataset issues. The two issues reflected on the dataset mainly lie in the lack of specific analysis of signal systems and digital signal processing, including the use of a large amount of voice data for training and testing, which involves data collection, cleaning, labeling, and processing. [10]In addition, the two key steps of speech recognition and speech synthesis are also related to signal systems, including the processing and conversion of speech signals. These can actually be explained; Then, specific analysis can be conducted on the records of the dataset, including its advantages and disadvantages, future trends, and comparison with other datasets, rather than simply understanding the scale and origin of the dataset. Last but not least, about the Language problem**.** The issue is related to the language would be the essential and basic problem for the conversational model to improve. The main improvement is to try to provide the official or common language for user to choose. [8] In addition, the model should make sure that it can recognize the multi language input and give corresponding language for output.

8. Future direction

In the future, conversational agents have broad prospects and potential, which will have a significant impact on people's lives in various fields. With the continuous progress of artificial intelligence technology and the deepening development of research on natural language processing and machine learning, it will develop into more intelligent, natural, and human beings. Here are some fields for the future prospects of dialogue agents. First of all, Providing personalized experience**.** Future Conversation agents will be able to better understand and adapt to users' personalities and needs. [2]They will lead advanced machine learning algorithms and big data analysis technology to provide personalized conversation experiences for each user by learning their historical data and behavior patterns. This will make the agent closer to the user and provide more accurate and targeted support. Next, is Multimodal interaction. Future agents will support multiple interaction methods, including voice, image, gesture, etc. By combining visual and speech recognition technologies, it can better perceive users' emotional expressions and intentions, achieving a more natural and rich interactive experience. Users can communicate with agents through voice and interact with them through images or gestures, enhancing the diversity and flexibility of interaction [3]. Then is about Social robots**.** In the future, future agents will not only be tools or services, but also more social and humanized robot partners. They will be able to understand and simulate emotions, establish emotional connections with users, and provide emotional support and companionship. This will play a positive role in lonely or emotionally supportive populations, providing users with a friendly, understanding, and supportive communication environment. Furthermore, Cross language interaction**.** Future agents will be able to achieve cross language interaction. Through natural language processing and machine translation technology, agents can understand and respond to input from multiple languages. This will help break down language barriers, promote cross-cultural communication and global cooperation. [10]In summary, future agents will become important partners and tools in people's lives. They can not only provide personalized services and support, but also adapt to diverse interaction methods and domain requirements. Through intelligent, personalized, and emotional design, agents will bring people a more convenient, rich, and meaningful communication experience.

9. Conclusion

Overall, voice agent is a cutting-edge technology that provides users with intelligent and convenient communication through the integration of technologies such as speech recognition, natural language processing. It provides it with a natural and smooth interaction experience, using input and output as natural interaction methods to deeply communicate with computers [3]. It naturally and smoothly improves the efficiency and convenience of human-computer interaction, provides intelligent assistant functions, and acts as an intelligent assistant. It can answer questions, provide information, execute tasks, and control devices, supply personalized suggestions and services to meet user needs, and apply in multiple fields, Voice agents have potential applications in various fields such as finance, healthcare, and education, as well as improvements in barrier free interaction and data-driven technology, bringing significant contributions and breakthroughs to the fields of human-computer interaction and voice technology.[1] In explaining its unique operational processes, databases, deep learning models, application prospects, and temporary limitations for the future, we also continuously discuss deeper into exploring and discovering its novelty and charm. At the same time, the paper also pointed out its progress and development direction. We believe that the development of voice agent technology will bring revolutionary changes to fields such as intelligent assistants, smart homes, voice search, and voice navigation, greatly facilitating people's daily lives. In the future technological revolution society, it will play an important role in promoting the digital transformation and intelligent development of the industry; Of course, the viewpoints elaborated in the paper also list its limitations. Although voice agents still face challenges in accuracy, semantic understanding, and multilingual support, their continuous innovation and progress provide broad development space for their application fields. However, we have provided detailed and appropriate solutions and new valuable perspectives for this. Because we always believe that although voice agents still face challenges, we still make a difference in their continuous innovation and progress and provide broad development space for their application fields. The final section also lists potential application markets in the future. Of course, with the continuous progress of speech technology, language models, and artificial intelligence, we can boldly predict that in the future, speech agents will become more intelligent, personalized, and demonstrate enormous commercial and social value in multiple fields [2].

References

[1] Young, R. M., & Moore, J. D. (2018). Conversational AI: The science behind the Alexa Prize. AI Magazine, 39(3), 25-34.

[2] Schlangen, D. (2017). Situated interaction with embodied conversational agents. Synthesis Lectures on Human Language Technologies, 10(2), 1-184.

[3] Porcheron, M., Fischer, J. E., & Sharples, S. (2018). Voice interfaces in everyday life. Human–Computer Interaction, 1-36.

[4] Ali, A. (n.d.). Conversational AI chatbot based on encoder-decoder architectures with ... https://www.researchgate.net/publication/338100972_Conversational_AI_Chatbot_Based_on_Encoder-Decoder_Architectures_with_Attention_Mechanism

[5] Ali, N. (n.d.). Chatbot: A Conversational Agent employed with Named Entity Recognition Model using Artificial Neural Network

[6] Jackylyn L. Beredo, Ethel C. Ong. (n.d.). A Hybrid Response Generation Model for an Empathetic Conversational Agent.

[7] Kenro Go, Toshiki Onishi, Asahi Ogushi, Akihiro Miyata. (n.d.). Conversational Agents Replying with a Manzai-style Joke

[8] Chunjong Par, Chulhong Min, Sourav Bhattacharya, Fahim Kawsar. (n.d.). Augmenting Conversational Agents with Ambient Acoustic Contexts

[9] Kenro Go (2021) Conversational Agents Replying with a Manzai-style Joke. 221-230

[10] Zheng Zhang (2021) Building a storytelling conversational agent through parent-AI collaboration.1-6