The Development Trend of Intelligent Speech Interaction

Yishuang Ning[1,2,3,4(✉)], Sheng He[1,2,3,4], Chunxiao Xing[1,2], and Liang-Jie Zhang[3,4]

[1] Research Institute of Information Technology, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
ningyishuang@126.com
[2] Department of Computer Science and Technology, Institute of Internet Industry, Tsinghua University, Beijing 100084, China
[3] National Engineering Research Center for Supporting Software of Enterprise Internet Services, Shenzhen, China
[4] Kingdee Research, Kingdee International Software Group Company Limited, Shenzhen, China

Abstract. To make the computers have capabilities of listening, speaking, understanding and even thinking is the latest development direction of human-computer interaction. As one of the most convenient and natural ways for communication, speech has become the most promising way of human-computer interaction in the future, which has more advantages than other interaction ways. As one of the most popular artificial intelligence (AI) technologies, intelligent speech interaction technology has been widely applied in many industries such as electronic commerce, smart home and intelligent industry as well as manufacturing. It will change the user behavior habits and become the new mode of human input and output. In this paper, we state the current situation of intelligent speech interaction at home and abroad, take many examples to illustrate the application scenarios of speech interaction technology and finally introduce its development trend in the future.

Keywords: Speech interaction · Intelligent technology · Multi-modality integration · Development trend

1 Introduction

In recent years, with the increasing popularity of portable intelligent terminal devices, it has become a new trend to use mobile applications to perform human-machine instant messaging anytime and anywhere. As one of the most natural, effective and convenient ways to obtain information, speech has become an important tool for human-machine communication. For example, the "2012–2013 China Instant Messaging Annual Monitoring and User Behavior Research Report" [1] shows that speech accounts for more than 50% of the instant messaging functions that users frequently use. A large number of intelligent speech

R. Xu et al. (Eds.): ICCC 2019, LNCS 11518, pp. 169–179, 2019.
https://doi.org/10.1007/978-3-030-23407-2_14

interaction applications (e.g. Apple Siri, Google Now, Microsoft Cortana, Baidu and Sogou speech assistants, etc.) have sprung up all over the world. As one of the key technologies in these years, intelligent speech interaction which uses speech as the main information carrier has made machines to not only have the ability to listen and answer, but also be able to understand and learn. It also makes the human-machine interaction increasingly more harmonious. Currently the ultimate goal of speech interaction has become understanding the whole scenario more and more accurately.

With the continuous expansion of the demand for intelligent speech interactive applications, the intelligent speech industry supported by key technologies (e.g. big data, cloud computing and mobile Internet, etc.) has developed rapidly, attracting the continuous attention of domestic and foreign research institutions and enterprises.

In many business areas or industries, speech has become a new way of input and output in the future. With the significant progress of speech technologies, the speech interaction technology has gradually gone to the market from libraries. It is currently applied in many areas, including telecommunication, medical, automotive electronics, home services and consumer electronics products. For example, you can use your speech to control a machine to operate as what you want or make a machine start to speak. This is the input and output of speech, which will become an input and output mode for future calculations.

However, what are the most important things about speech interaction? What is its ultimate goal? What are the application scenarios for enterprise? In this paper, we will first briefly introduce the speech interaction supporting technologies, then give an introduction to the current situation and many real-world application scenarios, and finally investigate the development trend of intelligent speech interaction.

The rest of the paper is structured as follows. Section 2 makes a brief description to the speech interaction support technologies, including basic speech technology, intelligent technology and big data technology. Section 3 introduces the current situation of speech interaction industry. Sections 4 and 5 present the application fields and the development trend of intelligent speech interaction, respectively. Section 6 summarizes the paper.

2 Speech Interaction Supporting Technologies

Intelligent speech interaction mainly studies the processing and feedback of speech information between human and machine. Generally speaking, there are three supporting technologies for speech interaction, as can be shown in Fig. 1.

2.1 Basic Speech Technology

The basic speech technology includes speech recognition, speech synthesis, speaker recognition and emotion recognition.

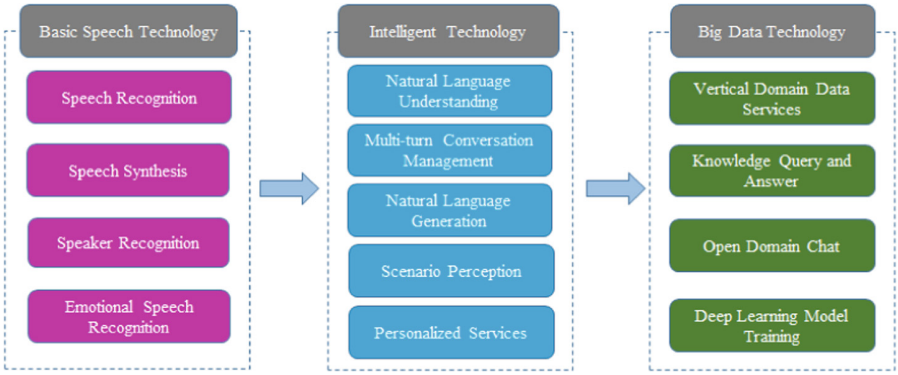

Fig. 1. Three supporting technologies for speech interaction.

Speech Recognition. Speech recognition is the ability of a machine or program to identify words or phrases in spoken language and convert them to a machine-readable format [2]. It is considered as the artificial ear. It uses acoustic model and language model to transcribe speech signals to texts. Acoustic model represents the relationship between linguistic units of speech and audio signals, while language model matches speeches with word sequences to help distinguish between words that sound similar.

Speech Synthesis. Speech synthesis, also called text-to-speech (TTS) is the technology that converts normal language text into natural and fluid speech [3]. It is the artificial production of human speech. Therefore, it is usually regarded as the artificial mouth. Currently the ultimate goal of speech synthesis has become synthesizing human-like speeches with high expressiveness.

Speaker Recognition. Speaker recognition or voiceprint recognition refers to the automated method of identifying or confirming the identity of an individual based on his or her voice [4]. It is the identification of a person from characteristics of voices. With the appearance of Amazon's smart speakers, this technology is rapidly used in many scenarios, such as public security, smart home and financial certification.

Emotional Speech Recognition. Emotion speech recognition is the process of identifying human emotion based on his or her speech [5,6]. Its main task is to analyze human expressions in multiple modalities such as text, speech or video and recognize the underlying emotions [7]. It is usually used in customer service scenarios to evaluate the quality of service (QoS) of agents.

2.2 Intelligent Technology

The second one is intelligent technology which includes natural language understanding, multi-turn conversation management, natural language generation, scenario perception and many personalized services.

Natural Language Understanding. Natural language understanding is the process that converts human languages in to machine-readable, structured and complete semantic expression [8,9]. It is the subtopic of natural language processing (NLP) in artificial intelligence. It can be used in many commercial fields such as automated reasoning, machine translation, question answering and large-scale content analysis.

Multi-turn Conversation Management. Multi-turn conversation management plays an important role in the human-machine interaction systems [10]. Its main functionality includes two aspects. The first one is to maintain or update the dialog state while the second one is to generate dialog strategy based on the dialog state [10].

Natural Language Generation. Natural language generation is a technology that simply generates natural language from a machine-representation system [11]. With the development of deep learning technologies, this technology is adopted to help transform their business and innovate on how they better engage with their customers by enterprise organizations.

Scenario Perception. Scenario perception [12] is a technology the purpose of which is to perceive the current situation through sensors and related techniques. As an interdisciplinary concept, this technology has many in-depth researches in computer science, cognitive science, psychology and linguistics. With this technology, intelligent devices can adaptively change with the current environment and push related services for users.

Personalized Services. The personalized services are implemented to provide or recommend related information to meet users' demand according to various channels and users' settings [13]. This technology breaks the traditional passive service model, can fully utilize various resource advantages to optimize the industrial chain and conduct all-round services for the purpose of meeting individual user demand actively.

2.3 Big Data Technology

The last one is big data technology [14,15], such as vertical domain data services, knowledge query and answer, open domain chat and large scale deep learning model training. In many real-world scenarios, users will use various of softwares

or applications to understand the online speeches. All the speech data will be uploaded to the cloud and use the application programming interfaces (APIs) to transform speeches into texts. When the speech files are saved, they will actually be even larger than the text and image files. Therefore, speech is also a very important object based on the big data processing technology in the future.

3 The Current Situation of Speech Interaction Industry

In this section, we will introduce the current situation of speech interaction industry. As an important part of intelligent speech interaction, speech recognition or speech search has been widely used in many application scenarios. It has become the needs of mobile Internet times. Nowadays, in many applications (e.g. Baidu search or Google search, etc.), there is a speech search functionality. When users would like to buy a product or search a place, they just need to speak to the applications and then the applications will help them to find the corresponding information.

However, there are still many challenges for the speech interaction technology. (1) Although the near-field speech recognition has achieved significant improvements (can even reach 97%), the accuracy of the far-field speech recognition is still very low; (2) When there are many speakers, that is, when the speeches come from different directions, it will be difficult to give accurate response; (3) When the surrounding environment may have noise or reverberation, the technology is relatively less mature; (4) When the speech files are uploaded to the cloud through the Internet from different places, due to the differences of the speech quality, how to combine them for speech recognition is also a big problem.

4 Application Fields of Artificial Intelligence

To get good results in speech recognition, in addition to exploring application scenarios, many artificial intelligence (AI) [16] technologies should be introduced. In the AI industry, there are three application fields that can best bring value to people: speech, image, and natural language processing. Therefore, it will be natural to use the AI technologies in speech interaction, image processing and semantics understanding. In this section, we only discuss the application scenarios for speech interaction.

4.1 Smart Home and Intelligent Enterprise Fields

As is known to us that smart home [17] is still controlled by mobile phones currently. In this case, speech interaction can undoubtedly greatly enhance the operation experience of smart home. Amazon Echo is the first smart speaker that is made to play the role of controlling devices at home or offices. As one of the most successful products for speech interaction in foreign countries, Amazon Echo has become the best-selling electronic product in 2016. Relying on this

product, Amazon gradually seized the information entrance of people's lives. They can use it to broadcast the daily news and weather forecast, schedule trips, or control the electrical appliances at home, etc. As a speech-based interaction controller, one of its core capabilities is to integrate various services on the Internet into the connector of the smart speaker. For example, the "IFTTT" is a function that can trigger events when conditions are met. For example, a photo will be posted to the Sina Weibo automatically after it has been taken. It turns such an "IF-THEN" condition and event-driven process into a rule which will fulfill the intention when the condition is met. As a matter of fact, such a process can also be controlled by smart speakers, which means that the future speech recognition will be seamlessly connected to the various services on the Internet with the mouse as we can see today. In addition, Google also launched the same device, Google Home, which has been maken as another input mode at home or in the office. Likewise, some Internet companies in China (e.g. Baidu, Alibaba, Jingdong and Xiaomi, etc.) have published their smart speaker products.

However, these products mainly focus on the personal consumption level entertainments, lacking of applications in enterprise scenarios. At present, the top international enterprise resource processing (ERP) vendors (e.g. Oracle, SAP, Salesforce and Acumatica, etc.) are building enterprise digital AI assistants and applying them in many enterprise fields, such as procurement, finance, human resource (HR), customer relation management (CRM) and distribution management. These AI assistants can be connected to a variety of back-end application systems, helping employees simplify user interaction, and improving efficiency and user satisfaction. In China, as the leading enterprise in the enterprise Internet, Kingdee also proposes the business digital assistant solution for enterprise users which takes smart speakers and chatbots as the product design, and takes financial office scenarios as the core businesses. It is actively deploying the AI scenarios to seize the industrial development opportunities. In the future, there will be increasingly more hardware devices that use speech as the new input and output.

4.2 Vehicle Field

With the development of speech technologies, the value of speech interaction is gradually transformed to the "big connection" stage. As one of the core connected hardware devices, the speech interaction platforms for vehicle are gradually applied in vehicle navigation systems. The entering of speech interaction technology cannot only free the drivers' hands to make convenient driving, but also make them concentrated on the surroundings to ensure the safety.

As the leading enterprise in the speech interaction field, Amazon has made great efforts to commercialize on the production vehicles of many global automobile companies (e.g. Ford, Volkswagen, Hyundai, BMW and Nissan, etc.) [18]. In the CES of 2018, Amazon has won the strong support of Toyota, Batten and Jeep. As a strong competitor, Google is relying on the high carrying rate of its Android Auto to upgrade the Google Assistant to make connection to a large scale of vehicles such as Ford, Nissan, Volkswagen and more than 40 brands

of more than 400 models. In China, as the world's largest car ownership country, the speech interaction battlefield competition is also very fierce. Early In 2017, Baidu first released the conversational AI operating system, DuerOS, and has applied in vehicle scenarios. The speech synthesis technology of iFlyTek also successfully cooperated with many large automobile factories such as Volkswagen, Beiqi and Volvo [18].

4.3 Wearable Device Field

Compared with communication through text information, the universal applicability of speech recognition is much stronger. Depending on its inherent characteristics, the speech interaction technology has been applied in the wearable device field. More and more wearable devices (e.g. smart watch, smart glasses) have been implanted the speech interaction systems to make the interaction become much more convenient. Suppose when you are tired from overusing your eyes, smart glasses can scan the texts and help you "read" them out.

4.4 Customer Service Field

Whether it is a website, software, APP or other entity product, customer service is a always its important part. The huge number of repetitive human services have brought huge costs to enterprises. Faced with the trend of "user-centric" personalized service, how to effectively reduce customer service costs is a great concern for enterprises. Compared with traditional human services, there are many advantages for intelligent customer service: (1) the cost is much lower and the efficiency is higher; (2) it can be applied to various fields; (3) for user consultation, it can search information with the big data technology and respond to user quickly and accurately.

4.5 Electronic Commerce Field

This kind of speech technology can not only be used in hardware device systems, but also to redefine the software systems in electronic commerce field. To take an instance for example, Google's smart messaging system, Smart Messaging or Google Allo, which let you search what you are looking for through interacting with it via speech, such as display the restaurants nearby after hearing what we say.

So what's happening now? In fact, whether it is Amazon Echo or Google Allo, such messaging devices or softwares can be used in the enterprise softwares or each scenario of specific industries. Therefore, what's happening now is to solve the problems that exist in the real world, such as far-field speech recognition or speech synthesis with low quality speech samples. In this paper, we will give three examples to illustrate the application of speech input in electronic commerce scenarios (as can be seen in Fig. 2).

The first example is an Amazon application on the mobile phone. For example, when a user wanted to know the best-selling camera in the last month, the

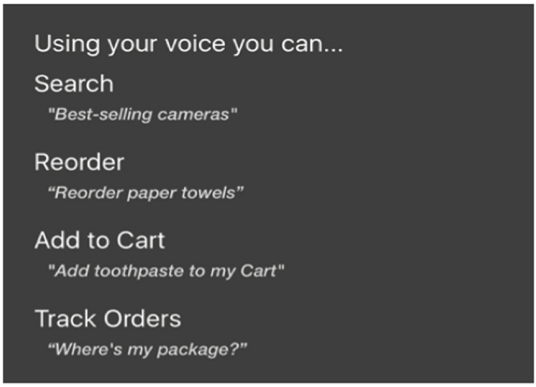

Fig. 2. Application of speech input in electronic commerce scenarios.

user would interact with the application via speech with the following three keywords: (1) the things that have been sold out; (2) camera; (3) best-selling. Then the application will try to understand the user intention and finally display the best-selling cameras on the Amazon website.

The second one is a repetitive order example. When a user told the application to reorder the hand sanitizer, it knows that the user has bought the hand sanitizer before, and will put it in your shopping cart automatically.

The last example is package tracking. When a user wanted to track where the package is, the application will check the orders the user has already placed, which of these orders have been shipped, and finally give the detailed information of the order.

From these examples above, we know that in a specific scenario, the keywords can be defined to form a thesaurus. After becoming a vocabulary of speech input and output in this scenario, the quality of speech recognition and speech synthesis will be improved significantly. It means that to use a new technology, it is best to find the application scenario first, then accumulate knowledge in the field, and then use the new technology to make better performance.

5 The Development Trend of Intelligent Speech Interaction

Finally, this paper will share four development trends of intelligent speech interaction.

5.1 Big Digitization

We are now in an digital economy era. There are almost 1.3 billion people who are using social media to communicate with each other on the Internet, and 9 billion sensors which have been adopted in the intelligent devices to obtain

data every day. In these cases, as an asset with a huge amount of data, it is necessary to redefine the technologies of speech recognition, speech synthesis and speech understanding in the environment of big data. Therefore, instead of putting it on a stand-alone machine to investigate the capability of speech recognition or natural language understanding, the future speech interaction will definitely explore new technologies in a digital ecosystem.

5.2 Internet of Things Industrialization

In the future, almost all the devices will be connected to the Internet. When everything is connected, every connected object will generate data. Once the data is generated, we need to interact with it to demonstrate its value. Therefore, it is possible for the speech input, output and understanding capability to become an important supporting technology in the connected objects.

5.3 Multi-modality Integration

All the intelligence comes from the understanding of different data. Speeches, images and texts will coexist. When we are conducting speech interaction, the text information and the capability of natural language understanding will assist in improving the performance of speech recognition and the quality of speech synthesis. Therefore, in the future, the multi-modality data (e.g. speech, text, image, etc.) will be integrated to enhance the ability of creating values as a data asset.

5.4 Trusted Interaction and Record

Because of the probability of error in the speech interaction, speeches and the results of recognition are generally hard to be used as a reliable data resource, such as for auditing or judicial field. As an incorruptible digital ledger of economic transactions, the blockchain [19] can be programmed to record not just financial transactions but virtually everything of value [20]. Therefore, through the blockchain technology, the whole process of speech interaction can be recorded and traced in the blocks. Based on the feature of irreversibility, it is possible to realize the trusted interaction and record of the speech. Beside the advantage of logging and feedback of errors in the interaction and recognition process, it is more likely to innovate some new valuable business interaction modes, for example, the enterprises can discuss and sign contracts more quickly just by speeches.

6 Conclusion

In the era of digital economy, intelligent speech interaction will become the new way of input and output of equipment operation in the future. The future speech interaction will inevitably explore new technologies in a digital ecosystem.

With the expansion of the influence of intelligent devices, intelligent speech interaction can be popularized to a wide range of people and integrated into users' lives and work, thus satisfying both of their actual demand and psychological needs. On the one hand, it will become an effective way of replacing a huge number of repetitive, customized human services and work categories. On the other hand, it will create new scenarios and promote new technologies to provide better user experience. With its extensive application in medical, financial, education and other important industries, enterprises should grab this opportunity to achieve multi-win relying on their data centers and platforms.

Acknowledgements. This work is partially supported by the technical projects No. c1533411500138 and No. 2017YFB0802700. This work is also supported by NSFC (91646202). This work is also supported by NSFC (91646202), the 1000-Talent program and the China Postdoctoral Science Foundation (2019M652949).

References

1. 2012–2013 China Instant Messaging Annual Monitoring and User Behavior Research Report [OL], 25 December 2018. https://wenku.baidu.com/view/3eb4d4d6f90f76c661371a83.html
2. Speech Recognition [OL], 25 December 2018. https://en.wikipedia.org/wiki/Speech_recognition
3. Speech Synthesis [OL], 25 December 2018. https://en.wikipedia.org/wiki/Speech_synthesis
4. Speaker Recognition [OL], 25 December 2018. https://en.wikipedia.org/wiki/Speaker_Recognition
5. Petrushin, V.: Emotion in speech: recognition and application to call centers. In: Proceedings of Artificial Neural Networks in Engineering (1999)
6. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn. **44**(3), 572–587 (2011)
7. Wu, B.Y., Jia, J., He, T., Du, J., Yi, X.Y., Ning, Y.S.: Inferring users emotions for human-mobile voice dialog applications. In: Proceedings of IEEE International Conference on Multimedia & Expo (ICME), Seattle, America (2016)
8. Natural-language Understanding [OL], 25 December 2018. https://en.wikipedia.org/wiki/Natural-language_understanding
9. Green, G.M.: Pragmatics and Natural Language Understanding. Routledge, New York (2012)
10. Sarikaya, R., et al.: An overview of end-to-end language understanding and dialog management for personal digital assistants. In: 2016 IEEE Spoken Language Technology Workshop (SLT), pp. 391–397, December 2016
11. Natural-language Generation [OL], 26 December 2018. https://en.wikipedia.org/wiki/Natural-language_generation
12. Roesener, C., Hareter, H., Burgstaller, W., Pratl, G.: Environment simulation for scenario perception models. In: Proceedings of IEEE International Workshop on Factory Communication Systems, pp. 349–352 (2004)
13. Otebolaku, A., Lee, G.M.: A framework for exploiting Internet of Things for context-aware trust-based personalized services. Mob. Inf. Syst. (2018)

14. Kashlev, A., Lu, S., Mohan, A.: Big data workflows: a reference architecture and the DATAVIEW system. Serv. Trans. Big Data (STBD) **4**(1), 1–19 (2017)
15. Zhang, L.J., Zeng, J.: 5C, a new model of defining big data. Serv. Trans. Big Data (STBD) **4**(1), 48–61 (2017)
16. Russell, S.J., Norvig, P.: Artificial Intelligence: A Modern Approach. Pearson Education Limited, New Delhi (2016)
17. Stojkoska, B.L., Trivodaliev, K.V.: A review of Internet of Things for smart home: challenges and solutions. J. Clean. Prod. **140**, 1454–1464 (2017)
18. Vehicle Voice Interaction: Transition from Vase to Just Need [OL]. http://baijiahao.baidu.com/s?id=1598227749471529133&wfr=spider&for=pc
19. Zheng, Z., Xie, S., Dai, H.N., Chen, X., Wang, H.: Blockchain challenges and opportunities: a survey. Int. J. Web Grid Serv. **14**(4), 352–375 (2018)
20. What is Blockchain Technology? A Step-by-Step Guide For Beginners [OL], 2 May 2019. https://blockgeeks.com/guides/what-is-blockchain-technology/