Advances in Data Security and Privacy in Cloud Computing – Research Article

International Journal of Distributed Sensor Networks

International Journal of Distributed Sensor Networks
2022, Vol. 18(11)

DOI: 10.1177/15501329221134479
journals.sagepub.com/home/dsn
SAGE

An online intelligent electronic medical record system via speech recognition

Xin Xia[1,2*], Yunlong Ma[3*], Ye Luo[1] and Jianwei Lu[1,4,5]

Abstract

Traditional electronic medical record systems in hospitals rely on healthcare workers to manually enter patient information, resulting in healthcare workers having to spend a significant amount of time each day filling out electronic medical records. This inefficient interaction seriously affects the communication between doctors and patients and reduces the speed at which doctors can diagnose patients' conditions. The rapid development of deep learning–based speech recognition technology promises to improve this situation. In this work, we build an online electronic medical record system based on speech interaction. The system integrates a medical linguistic knowledge base, a specialized language model, a personalized acoustic model, and a fault-tolerance mechanism. Hence, we propose and develop an advanced electronic medical record system approach with multi-accent adaptive technology for avoiding the mistakes caused by accents, and it improves the accuracy of speech recognition obviously. For testing the proposed speech recognition electronic medical record system, we construct medical speech recognition data sets using audio and electronic medical records from real medical environments. On the data sets from real clinical scenarios, our proposed algorithm significantly outperforms other machine learning algorithms. Furthermore, compared to traditional electronic medical record systems that rely on keyboard inputs, our system is much more efficient, and its accuracy rate increases with the increasing online time of the proposed system. Our results show that the proposed electronic medical record system is expected to revolutionize the traditional working approach of clinical departments, and it serves more efficient in clinics with low time consumption compared with traditional electronic medical record systems depending on keyboard inputs, which has less recording mistakes and lows down the time consumption in modification of medical recordings; due to the proposed speech recognition electronic medical record system is built on knowledge database of medical terms, so it has a good generalized application and adaption in the clinical scenarios for hospitals.

Keywords

Electronic medical record system, speech recognition, linguistic knowledge base, semantic analysis, electronic medical record

Date received: 17 January 2022; accepted: 28 June 2022

Handling Editor: Yanjiao Chen.

Introduction

In recent years, electronic medical records (EMRs) have rapidly spread in medical institutions both at home and abroad, bringing great conveniences to healthcare professionals in recording patients' conditions. The recording template of EMR has indeed improved the efficiency of physicians' writing to a great extent, but in the meanwhile, it has also resulted in

[1]School of Software, Tongji University, Shanghai, China
[2]East Hospital, School of Medicine, Tongji University, Shanghai, China
[3]College of Electronic and Information Engineering, Tongji University, Shanghai, China
[4]College of Rehabilitation Science, Shanghai University of Traditional Chinese Medicine, Shanghai, China
[5]Engineering Research Center of Traditional Chinese Medicine Intelligent Rehabilitation, Ministry of Education, Shanghai, China

*Xin Xia and Yunlong Ma contribute equally to the article work as the first authors.

Corresponding author:
Jianwei Lu, College of Rehabilitation Science, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China.
Email: jwlu33@tongji.edu.cn

similar or even identical medical records of patients with different conditions. This homogenization of medical records has led to a large number of incorrect records, which seriously affects patient care.[1–4] In addition, the information inputs of EMR recordings can also take large time consumption, which affects the communications between doctors and patients for the detail information of patients and slows down the treatment process to some extent.[4–6]

The two problems of EMR mentioned above are explained in detail below. First, because the only way of information inputs of EMR systems used in hospitals are manually through computer keyboards at present, this mode of human–computer interaction is really less efficient.[7] Doctors have to spend a large time consumption in inputting information about patients' conditions in their daily clinical work, which seriously affects the efficiency of communication between doctors and patients, blurs the patient's treatment histories and blocks the efficiency of the doctor's clinical work. Moreover, the information input manually can be distorted to a certain extent, which means that the doctor may not be able to record the patient's entire condition in details. Second, EMR suffers from homogeneity problems. The current EMR systems will provide many templates, and the doctor only needs to choose the appropriate template to fill in some specific details about the patients' conditions. However, this approach also creates another problem: EMR records of different patients may be similar probably, which obviously affects the quality of EMR records and even leads to many inappropriate treatment plans prescribed to patients by medics. And as we know, frequent information input operates of EMR system through computer keyboards will obviously causes many inevitable recording mistakes by carelessness following continuous inputs, and which directly affects the veracity and quality of medical records.

Therefore, in this article, for overcoming the shortcomings of the existing EMR, we propose to replace the input method of computer keyboards used in EMR with the method of speech recognition. Recently, speech recognition technology has developed rapidly and applied in many real-life scenarios.[8] However, a few studies have been conducted to apply speech recognition technology to the medical field, a highly specialized field. The terms and concepts used in medical systems are very different from the words that people usually use. In patient–doctor communication, it is a very challenging problem to accurately extract information about the patient's condition and the doctor's diagnosis and translate it into the terms and concepts used by the medical system. As we know, because of the large amount of patients' visits, doctors have to spend a large time consumption in inputting information about patients' conditions in clinic. Obviously, it seriously affects the efficiency of communication between doctors and patients, blocks the understandings of doctors for patients' conditions in details, and also consumes plenty time due to prescribe proper treatment plans for patients. Moreover, the information input manually can be distorted to a certain extent, which means that the doctor may not be able to record the patient's entire condition in details. The current EMR systems applied in clinics will provide many recording templates, and the doctor only needs to select an appropriate template for filling in some specific details about the patients' conditions, so it also causes another problem: EMR records of different patients may be similar probably, which affects the quality of EMR records and even leads to many inappropriate treatment plans prescribed to patients by medics. Meanwhile, frequent information input operates of EMR system through computer keyboards will produce many inevitable recording mistakes by carelessness, which directly affects the veracity and quality of medical records.

In this work, we develop and build an online EMR system based on speech interaction. The system integrates a medical linguistic knowledge database as the dictionary database for the input voice words, a specialized language model for sorting out the medical knowledge, a personalized acoustic model aiming to construct personalized acoustic models to distinguish different doctors, for improving the system's recognition accuracy, and build a fault-tolerance mechanism for the recognition of the words in speeches. At first, for testing the proposed system, we construct the medical speech recognition data sets using audio and EMRs from the real medical scenarios of hospital daily works. To further upgrade the accuracy of speech recognition in the medical scenarios, we propose an improved forward great matching algorithm, based on the medical scenario data sets, our proposed recognition algorithm significantly outperforms other machine learning algorithms. Furthermore, compared with traditional EMR systems that rely on computer keyboards as the information input access, our proposed system is much more efficient, and its accuracy rate upgrades with the increases of online utilization time. And the rest of this article is organized as follows: Section "Research background" reviews some of the work related including EMR speech recognition. Section "Methodology" gives the method of our proposed EMR system based on speech recognition. Section "Results and discussion" presents the experimental results. In section "Conclusion," we conclude the work.

Research background

This article aims to propose an online EMR system based on intelligent speech recognition (ISR)

technology. Here, we first review the related work of EMR and speech recognition.

EMR

EMR is used to manage the personal health status data of patients. It involves the collection, storage, transmission, processing, and utilization of patient information. It standardizes healthcare staff's medical behaviors, reduces medical errors, and improves the quality of healthcare services. EMR has outstanding performances in the aspects of experiences in the clinic for patients. Meanwhile, the informatization of medical records is critical in the construction of hospital information systems. Therefore, EMR is the foundation of hospital digital and the personalized development of medical health information. It is also an inevitable demand for the global medical and health industry information application in clinics.[4–7]

On one hand, the template design approach of EMRs has indeed improved the efficiency of physicians' writing to a great extent. On the other hand, it has also resulted in similar or even identical medical records of patients with different conditions. This homogenization of medical records has led to a large number of incorrect records, which seriously affects patient care.[1–4] In addition, entering EMRs can also take a lot of time, which affects the communication between doctors and patients and slows down the treatment process to some extent.[4–6]

Speech recognition

As a key issue of human–computer interaction, speech recognition has made rapid development in the past decade. Speech recognition technology is also named automatic speech recognition (ASR). ASR aims to convert the vocabulary contents of human speeches into computer-readable input information, such as press keys, binary codes, or character sequences.[8] ASR is a technology for resolving the problems of "comprehending" the meanings of human languages by machines. People are committed to making the machines understand human voice instructions and controlling the machine through voice.

Since the development of speech recognition technology, it can be totally parted into three stages from the technical direction:

1. From 1990s to first decade in the 21st century, speech recognition mainly focuses on small vocabularies and isolated word recognition based on GMM-HMM,[9] with a slowly improved recognition rate (RR) And GMM-HMM is a method using Gaussian mixture model (GMM) for describing the probability distribution function in sound production.
2. From 2009 to 2015, the accuracy of speech recognition has been significantly improved with a large vocabulary continuous speech recognition, and iFLYTEK also released the voice cloud platform, which has been endowed with a strong environmental learning capability, improves the robustness to noises and accents. There are modeling technologies such as recurrent neural network (RNN),[10] long-term and short-term memory network (LSTM), bidirectional long-term and short-term memory network (BLSTM),[11] convolutional neural network (CNN),[12] and so on. In general, based on features mapping, reducing speech signals diversity, powerful speech recognition systems can be combined with these networks.
3. In recent years, voice technology companies are training deeper and more complex networks by end-to-end technologies.[13–15] Meanwhile, they have further greatly improved the performances of voice recognition using end-to-end technology. In 2017, the accuracy of voice recognition surpassed that of human beings for the first time. End-to-end technology mainly solves the problem that the losses of the lengths of output sequences.

With the breakthroughs of speech recognition, ASR is prominent in computer technology and application.[16] The first large-scale research on speech systematization in the world originated from Bell Laboratory in the 1950s. The Audry system developed by Bell Laboratory could recognize 10 English letters, which was an initial system with speech recognition. In the early 1990s, many technology companies started developing practical speech recognition systems in the application, and many human resources and material resources were invested in the research. In the late 1990s, the accuracy of the speech recognition system was greatly optimized, such as the Via-Voice platform, Dragon platform, naturally speaking platform, and Nuance Voice platform. In recent years, the layout of the speech recognition industry has become a speeding-up field. For example, Apple, Google, Facebook, and Microsoft successfully acquired Phonetic arts, Skype, Cortana, and other technology companies to strengthen and develop speech recognition applications.[17]

The research of speech recognition in China originated in the 1950s. With the progress of related technologies, the research level of speech recognition is always upgrading rapidly, and it has gradually reached a practical stage. Intelligent voice input systems adopt distributed computing technologies, with the advantages of better robustness, higher flexibility, and better performance. It can effectively improve efficiency and reduce the intensity and workload in the clinic, especially for

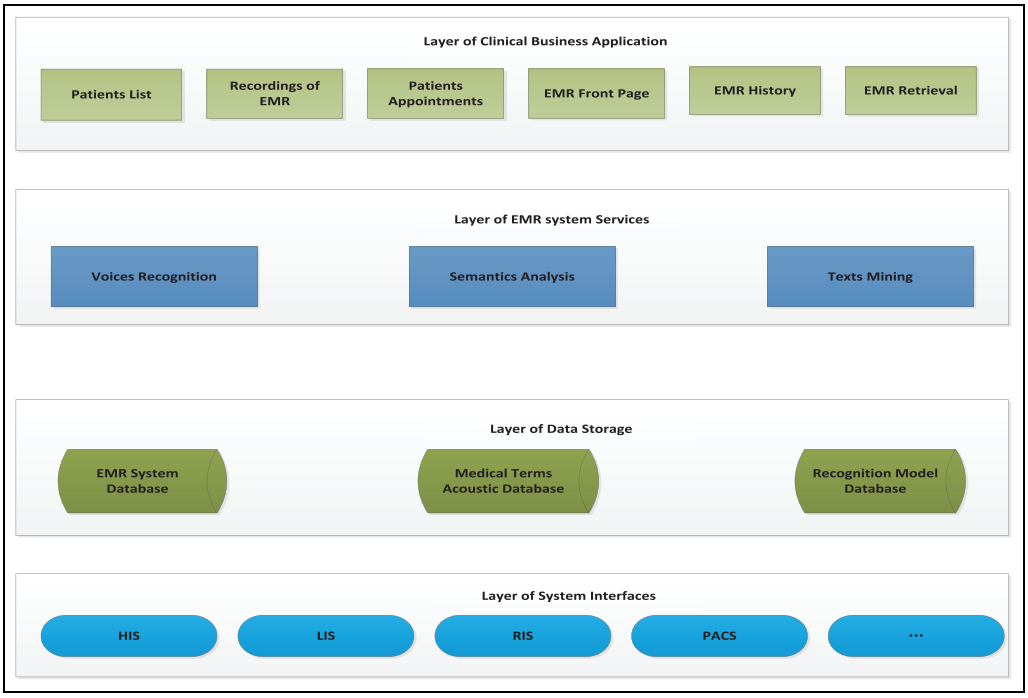

Figure 1. The architecture of functional description of speech recognition EMR.

the free long-term inputs of medical order information. For instance, Beijing Union Medical College Hospital has applied a medical voice recognition technology for the information input of EMR, which improves the work efficiency in the clinic, and more than 50% of doctors appreciate the voice recognition technology can shorten about 1 h per day and improve the work efficiency through an investigation in the clinic. Especially in recent years, with the progress of speech recognition technology, its paces are also accelerating in medical information application at home and abroad.[18,19] Such as Xingshulin, a domestic mobile medical company, has developed the first Chinese speech recognition engine of medical specialty, which is combined with its professional service product in the application of speech recognition of medical records. With the continuous research advances of medical information, speech recognition technology will profoundly impact the development of medical equipment interaction, information transmission, and data retrieval research.[20]

Speech recognition has entered a period of rapid growth in the past decade. Improving the capability of acoustic modeling and carrying out end-to-end joint optimization are a hot topic in speech recognition.[9–16] However, a few studies have been conducted to apply speech recognition technology to the medical field, a highly specialized field. The terms and concepts used in medical systems are very different from the words that people usually use. In patient–doctor communication, it is a very challenging problem to accurately extract information about the patient's condition and the doctor's diagnosis and translate it into the terms and concepts used by the medical system.

Methodology

The purpose of this article is to propose an online EMR system based on ISR technology. As a functional description of the EMR speech recognition, Figure 1 manifests the business functional parts of the system. The proposed system uses the input of voices as an alternative to the input approach of keyboards. The doctor dictates the medical record through the microphone. The system automatically generates the required EMR in real-time based on a speech recognition algorithm we proposed with high recognition accuracy, which effectively overcomes the shortcomings of other algorithms in medical scenarios. We describe the proposed system and algorithm in detail as the following sections.

Architectures of the online EMR system with speech recognition

To design an online EMR system with speech recognition, it is necessary to enhance the pattern RRs in terms of audio input device terminals, language knowledge databases, language recognition algorithm models, and the robustness of the algorithm. Thus, functionally, the system is composed of four plausible modules.

Build a language knowledge database. For the continuously processes of enriching and updating medical knowledge, the construction technology of knowledge bases with self-learning is employed for achieving the knowledge database. However, the knowledge atlas based on symbolic representation cannot accurately manifest clinical knowledge concepts; thus, this research intends to study the construction method of self-learning knowledge base, which mainly includes three aspects:

- Learning the multilingual embedded representation model of medical concepts automatically from a large number of medical books, medical literature, clinical guidelines, and other materials in combination with the characteristics of chapters.
- Discovering new concepts in medical literature based on deep attention model.
- Probabilistic representation and learning of disease diagnosis basis in medical literature and EMR data.

As we know, clinical data are extracted from present EMRs and doctors' daily languages by structurally organizing, processing, and restoring. As the basis of medical record ISR, it is very important to enhance the accuracy of medical record intelligent writing and recognition. Moreover, it is approached mainly through a machine learning medical knowledge database, such as medical reference books, professional teaching materials, databases of symptoms, diseases, examinations, drugs, clinical guides, case reports, and evidence-based literature.[21–23]

Establish a specialized and personalized language models. For clinical scenarios, speech recognition aims to facilitate effective communications between patients and doctors through semantic recognition, and subsequent extraction of key information of patients' voices and intention understandings. Meanwhile, it has to recognize the accents and dialects of patients from different regions accurately. For resolving the above problems in speech recognition, the following research works and development work are mainly carried out:

- Research the adaptive technology for speakers and accents and dialects.
- Research and develop the technology of customized large-scale language model in the medical field.
- Continuous construction of massive medical voice and language data resource database.

Aiming to sort out the common knowledge of diseases, diagnosis, drugs, inspections, and lab tests, summarize keywords of common diseases and establish specialized language models to improve the recognition accuracy, and follow doctors' different speech speeds and pronunciation habits, it is necessary to build a specialized language model and a personalized acoustic models based on speech recognition methods, such as segmentation methods of Chinese words for distinguishing different groups for improving the system's agility and RR.

Construct speech recognition model. Speech recognition is mainly composed of three parts: acoustic model, language model, and decoder.[20] In addition, there are front-end processing and post-processing modules. With the rise of various deep neural networks and end-to-end technologies, acoustic model is a very popular direction in recent years.[13–15]

From the above description, we can see that ISR plays an important role in speech recognition models. Integrated with outpatient and inpatient workstations, and medical detection report system, ISR transcribes doctors' speech contents into text information and adds them into outpatient and inpatient medical records, examination reports, and other text input positions. It also supports editing compiling commands, such as insertions, modifications, and deletions, and it is compiled with some complex operations such as cursor movements, line breaking, and cancelations.

Apart from the main function of recognition, the ISR module in the online EMR is also responsible to the recordings acquisition, that is, to collect voices and synchronously convert them into texts stored in the location. Suppose the online collection of voices is not feasible due to network problems. In that case, the local audio records will be directly used to collect the voices (as shown in Figure 2), restored, and the available network will accomplish the voice conversion. For example, android systems provide many class libraries compiled with Java to facilitate the application programs to collect voices and other operations. Most of the relevant class libraries are declared in the media package of SDK in Android, including audio records. The main purpose of designing audio record class libraries aims to facilitate the program for managing audio resources and interacting with the voices collected by the platform.

Multi-accent adaptive technology for doctors and patients of the online EMR with ISR

At present, EMR in use is a typical semi-structured medical text in hospitals. The convenient further research and effective analysis of Chinese medical texts depend on the extraction and construction of effective information from EMRs, so an accurate Chinese

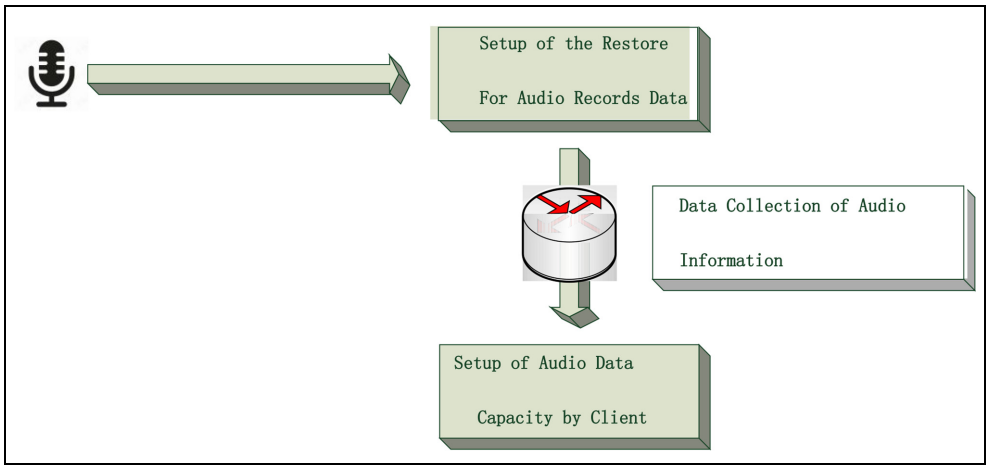

Figure 2. The architecture frame of audio record in ISR.

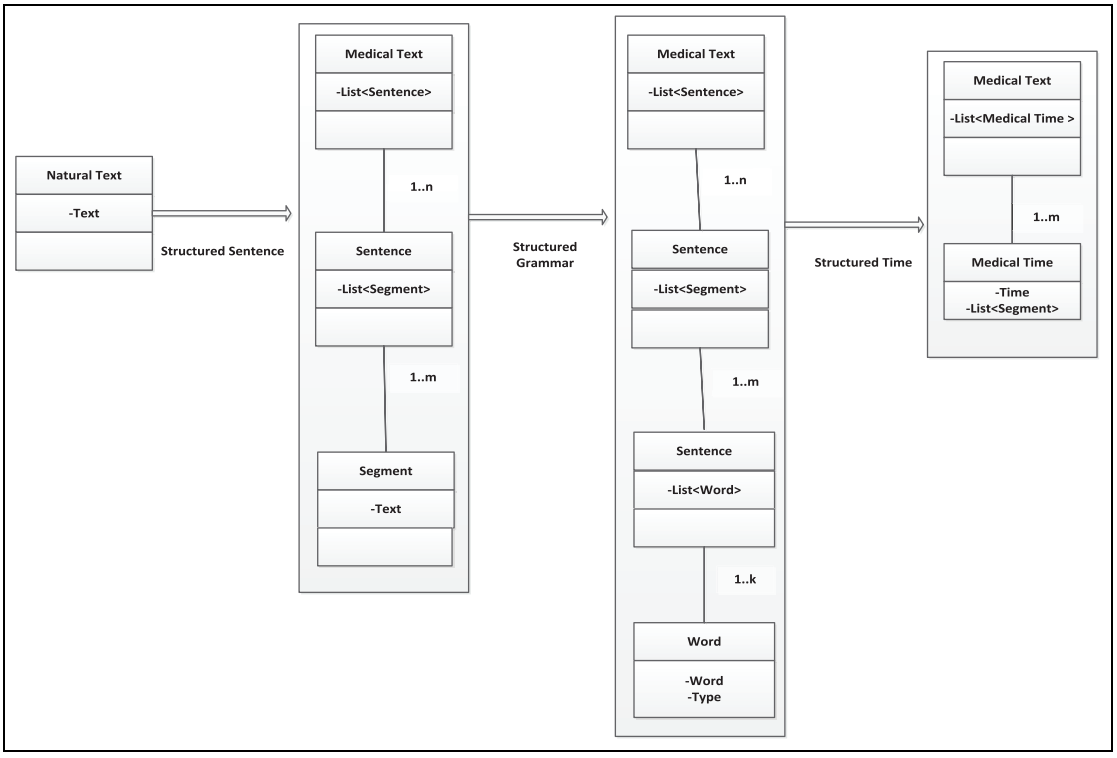

Figure 3. The flow chart of semantic analysis.

segmentation algorithm is required to achieve the goals. However, natural language processing technology based on common corpora is usually difficult to separate and label medical narrative texts effectively. As shown in Figure 3, it manifests the flow chart of semantic analysis. According to the characteristics and diversity of different patient or doctor groups in hospitals, it is considered to use the known training data for counting the dialect accent changes of doctors and patients from different regions, so as to achieve the construction of memory units based on the variation rules of dialect accent.

First, the acoustic part of the large-scale training data is decoded with the help of the general recognition model, and the decoding results and annotation are used to form a residual vector based on the phoneme level for representing the differences between the current speaker and the standard speaker, and then the part with large differences in the recognition of the speakers by the current model is selected as the residual attributes in the

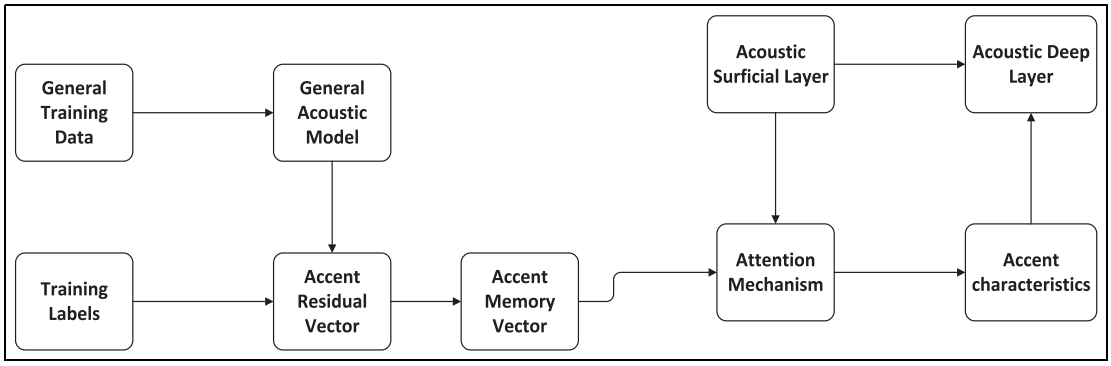

Figure 4. The recognition process of dialect accents.

dialect accent vector by setting thresholds. And the selected dialect accent residual attributes are clustered by clustering methods, and a fixed number of dialect accent memory vectors are formed for achieving the representation of dialect accent pronunciation. Considering that the characteristics of dialect accents in different regions are similar in some extent, after learning and refining the accent features with a small amount of data, the corresponding features are integrated into the acoustic modeling, and the dialect data are mapped to the space of standard pronunciation, so as to optimize the recognition effect of different dialect accents. In addition, because the attention mechanism is used to weight the clustered dialect features, for the regions with only a small amount of dialect data, the accent memory vector in the process of model training can also achieve ideal results in the corresponding dialect region by fine-tuning. And the recognition process of dialect accents is shown in Figure 4.

Construction of the training data knowledge base of medical recognition model

Medical terms and concepts are the most important contents in medical books and documents. They are usually recorded in the form of texts. Texts are the symbolic expressions of medical concepts, which often contain rich semantic connotation. Aiming at the term concepts in the medical field, the technologies of information extraction and disambiguation entities are adopted to automatically align the existing medical concepts and find out new term concepts. The details are as follows.

Entity extraction. In the medical field, there are many nested entities, such as:

> [left] location [thigh] body part [fatigue] symptom [element] ...

In the instance, "left" is the location word, "thigh" is the body part, "fatigue" is the symptom element, and these three subentities can form a top-level symptom entity, which is named as "left thigh fatigue."

The research adopts a multi-layer serialization annotation model based on BERT (Bidirectional Encoder Representation from Transformers) and self-attention mechanism, so as to realize the combined extraction of the nested and top-level entities. The model structure is described in Figure 5.

The elimination of entity ambiguity. However, the data sources of medical terms and concepts are diverse, which lead to the problems of conceptual duplication and ambiguous correlation between concepts. The same medical entity may have different statements in different data sources, so entity disambiguation (entity link) plays a very important step in medical term alignment and new concept discovery. Named entity recognition technology can identify the entity segments in texts, and the entity segments are often ambiguous or even unknown. Due to the ambiguity in texts, it is necessary to link the entity to the only entity in the objective world through entity linking technology. The total goal of entity linking task aims to correctly link the entity references extracted from the texts to the corresponding entity objects in the knowledge map. Through entity link technology, on one hand, it can eliminate the conceptual ambiguity of knowledge elements and redundant and wrong knowledge elements, so as to ensure the construction quality of knowledge base; on the other hand, new entities can be found actively for maintaining the real-time and completeness of the knowledge base. And the overall scheme is shown in Figure 6.

The large-scale integration language model in medics with medical scenarios integration

For balancing the recognition effects in general and medical application scenarios, the end-to-end adaptive method of the large-scale domain language model is inevitable to be employed. First, it is necessary to collect the corresponding text data combined with the

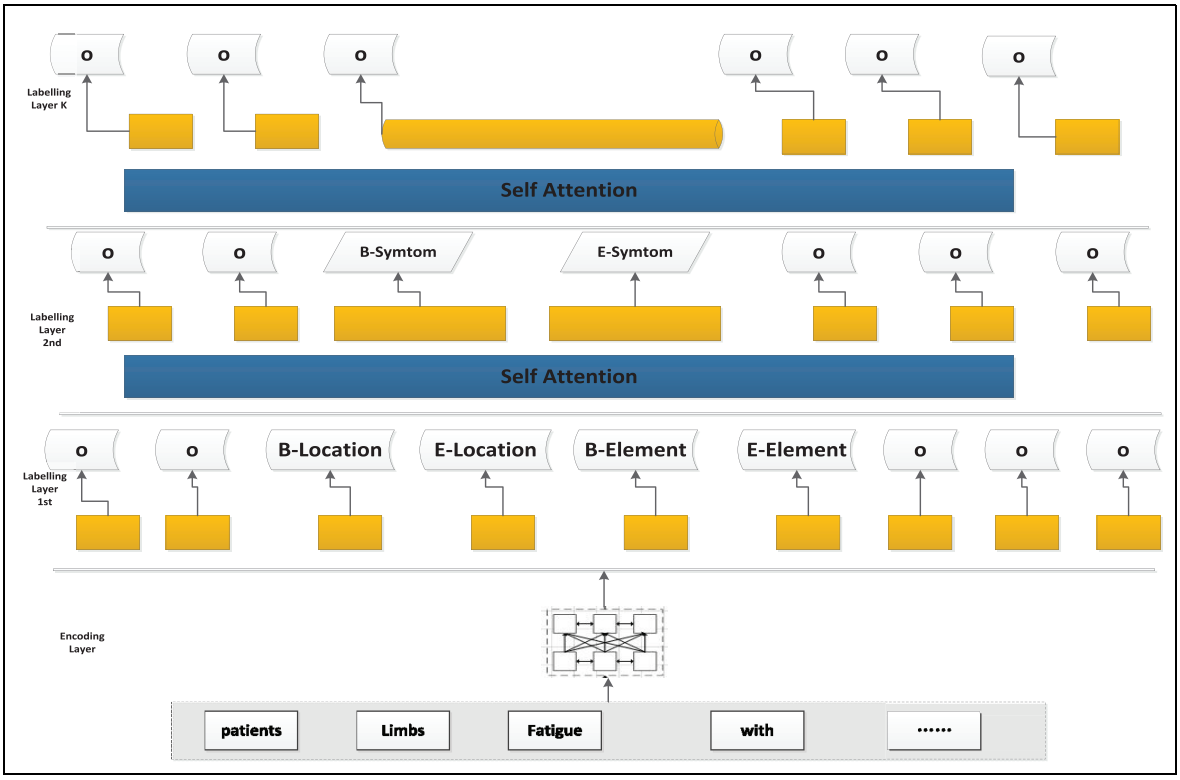

Figure 5. The formation construction of BERT model.

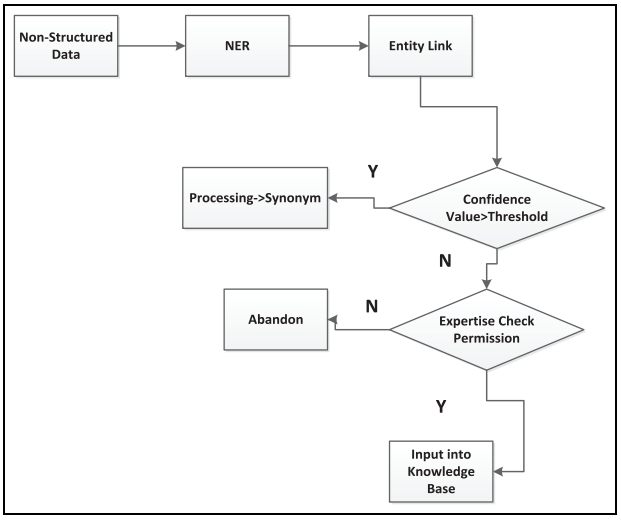

Figure 6. The workflow of the entity extraction scheme.

medical fields and training the domain language model; second, the general language model and medical domain model are self-learning integrated to ensure the general recognition effects of natural languages and dynamically optimize the effects in clinical scenarios. The technology needs to integrate the general natural language model and the medical term model through the gating mechanism. The gating parameters are dynamically learned in the process of model training. The decoder dynamically selects the former and the latter according to the relevant decoding parameters, which ensures a good recognition effect. The schematic diagram of the scheme is manifested in Figure 7.

Figure 7 shows the structure diagram of the general end-to-end (ED) model. The ED decoder end is a self-regressive decoding structure. The input parameters are the last decoding result ($yt - 1$) and the context vector ($ct - 1$) of the last decoding, which are used for self-regressive decoding through a one-way long and short-term memory (LSTM) module, and the hidden layer vector (s_t^d) of self-regressive decoding and the ED encoder end (H_{enc}) perform as the outputs of the attention operation (Att-ctx-dec) for obtaining the decoded context vector (ct), and the decoded context vector passes through a one-way LSTM module and a classification module to get the decoding results of this round. The ED model can be well used for sequence modeling without the loss of training data. If only the speech annotation data are used for the training of ED model, the advantages of the large amount of data of the language model cannot be well utilized. However, the shallow integration of the decoding results of ED model with the language model can rectify the ED results to a certain extent, there is an obvious deviation between the scenario covered by the language model trained by a large amount of data and the corpus of

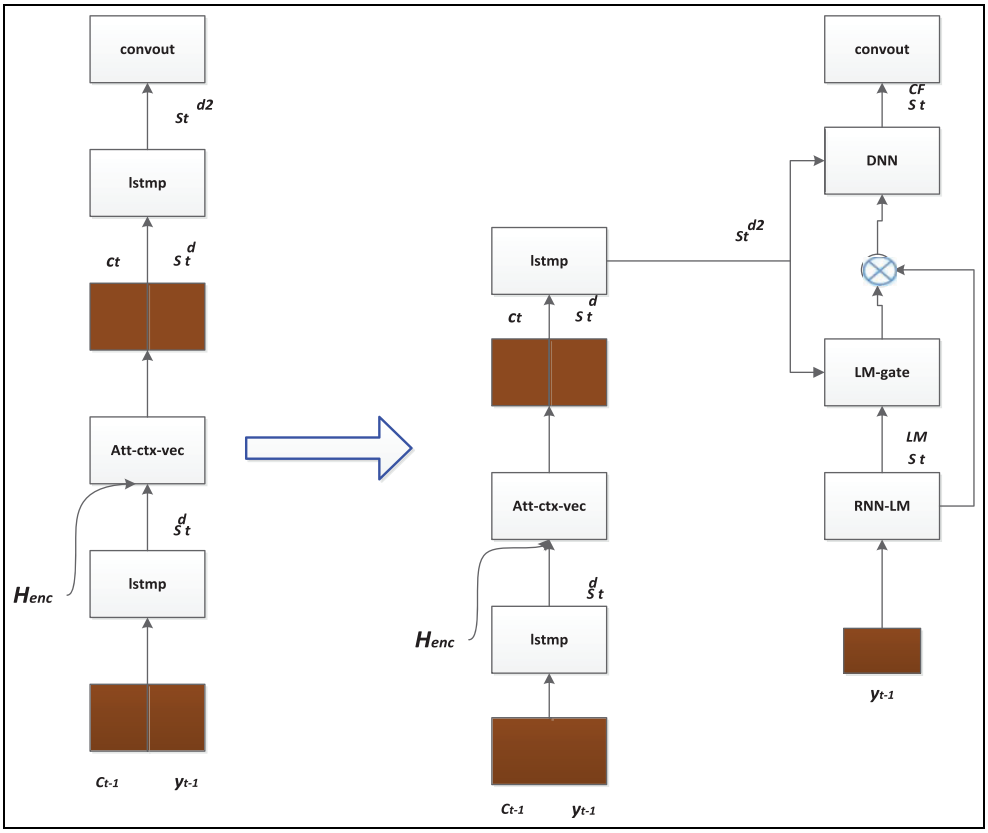

Figure 7. The manifestation of self-adaptive medical speech recognition.

medical terms trained by ED, which leads to a limited integration effect. Therefore, adjusting ED model and adopting the scheme of integration training between ED and medical terms language model can learn the large text corpus information of the language model in the structure of ED model, the process of fusion training between the encoder and recurrent neural network–based language model (RNN-LM). Through the integration training of the decoder of the model, as shown in Figure 6, the previous self-regressive, attention, and classification LSTM modules are consistent with the decoder of ED module. The integration training fuses the second LSTM output of the hidden layer with the hidden layer of the RNN-LM. And the input of the RNN language model is the decoding result of the last layer, and the RNN-LM is composed of two long-term and short-term memory modules. The hidden layer output of the RNN-LM calculates a gating (LM gate) together with s_t^{LM} and $s_t^{d_2}$. It is used to select the language model of medical terms. After selection, it is classified with $s_t^{d_2}$ to get the final decoding result. In this study, for obtaining a fit recognition effect in the application scene of clinical terms, a learning accumulation of the scenarios in different medical sections and departments is obviously employed for achieving a formulated network in connection with the speech recognition method, and in mutual communication and decoding. So, FSA (Finite State Automata), a directed acyclic graph structure, which is available to extract the summary of the language materials in some scenarios, and forms an FSA network, is manifested in Figure 8. With the aid of FSA network as a language model, an effective enhancement in recognition can be acquired by the mutual communication with the speech recognition model, and the workflow of FSA is figured out as in Figure 9.

As we know, clinics are a professional and expertise scenario in application. It is necessary to carry out text data collection and resource construction, such as medical terms dictionary, including multiple information such as symptoms, diseases, drugs, diagnosis, lab examinations, surgeries, and medical units; base of rules, including 20 rule bases, such as statement expansion, unlisted words, special symbols and word segmentation differences, and audio data need to be collected and labeled. There are two sorts of data collection: one is to collect real data in combination with the data of actual application scenarios in hospital, and this type of data model training is totally benefit; the other type is to design the texts of medical terms, and design the recording scheme for the recording environment, language, acquisition channels, recording corpus, age ranges, and

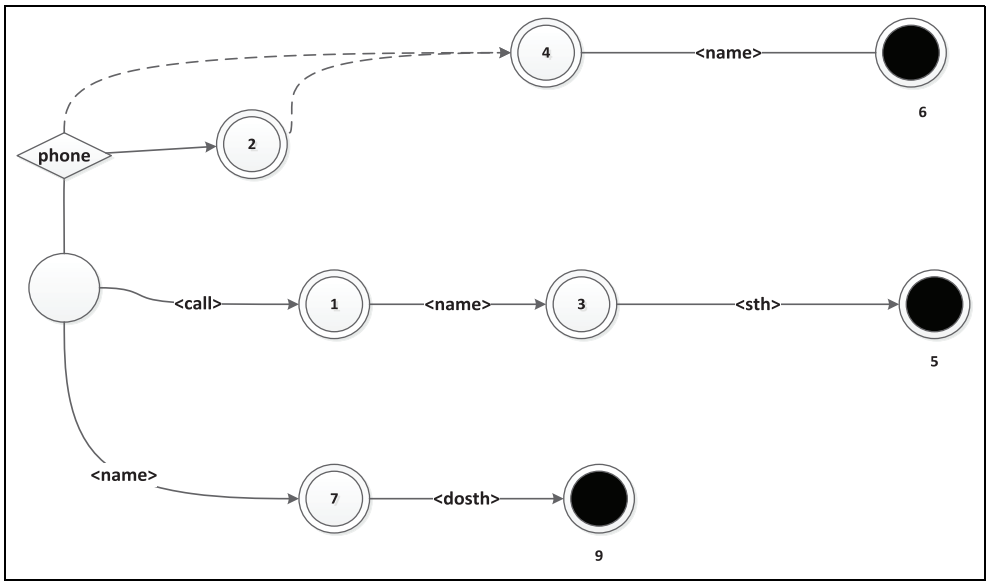

Figure 8. The network of FSA.

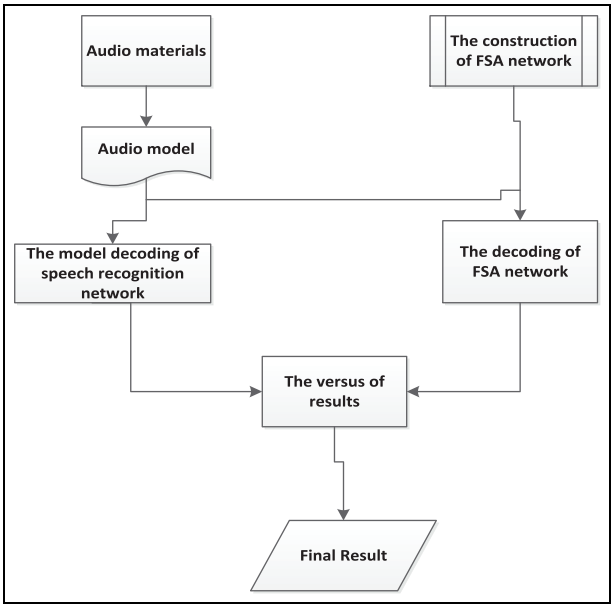

Figure 9. The process of FSA in the scenario recognition.

other factors. In the actual recording process, let the recorder will simulate the actual application scenarios as much as possible. Figure 10 refers to the process of data collection.

Related work

The design and development of proposed EMR system with speech recognition system started in 2020, and it is applied in different clinical sections in outpatient in Shanghai East Hospital, with the daily clinical patients as the study objects for observing the application results of the speech recognition EMR system. And the application departments in outpatient are including Respiratory, Digestive, Cardiovascular, Endocrinology, Neurology, Obstetrics and Gynecology, Pediatrics, Ophthalmology, Traditional Chinese Medicine, Stomatology, Orthopedics, Emergency, Hematology, ENT Department, Ultrasound, Pathology, and Medical imaging. And the proposed speech recognition EMR system is developed based on the database of ORACLE 12c.

Results and discussion

In section "The large-scale integration language model in medics with medical scenarios integration," it is mentioned that FSA network is employed to assist in the proposed speech recognition EMR system. Compared without FSA, with the aid of FSA network as a language model, a more effective enhancement in recognition can be gained by the mutual communication with the speech recognition model, and based on the data sets of Traditional Chinese Medicine, Stomatology, Ultrasound, Pathology, and Medical imaging. We can see the classification results with the aids of FSA are better than those of without FSA, and results in Table 1.

In this study, we use a technology with multiple accents adaption in the online EMR via speech recognition, for obtaining a better performance in speech recognition compared with the other speech recognition models. Based on the different clinical testing data set listed in Table 2, we can see that the multi-accent adaptive technology (Group A) plays a better average

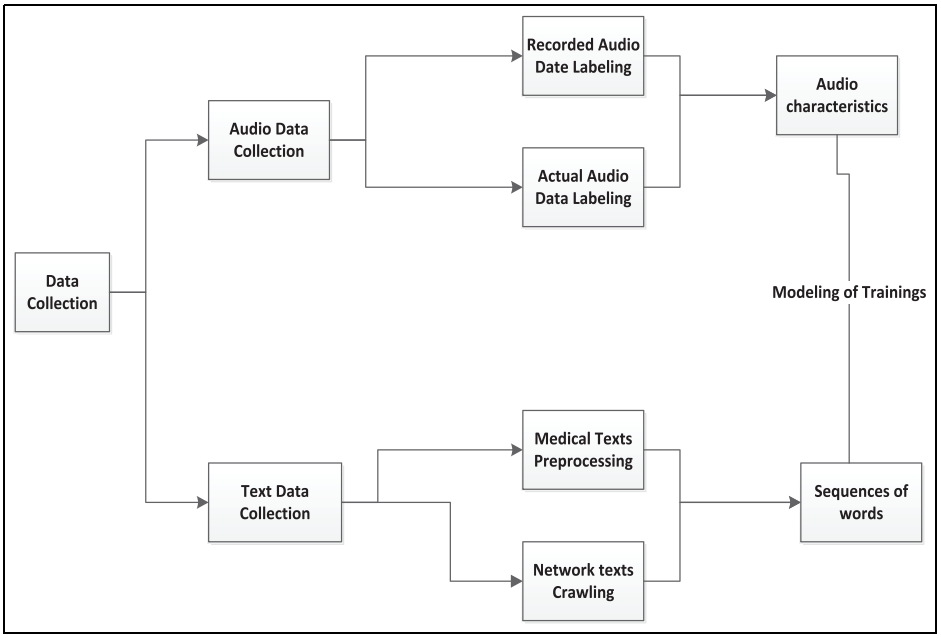

Figure 10. The frame of data construction in the scene of clinic.

Table 1. The comparison of classification accuracies (%) with FSA on the data sets.

Testing data sets	Words	With FSA	Without FSA
Ultrasound	10,000	96.72	96.23
Stomatology	10,000	97.29	96.37
Pathology	10,000	97.66	96.08
Medical imaging	10,000	98.01	96.86
Traditional Chinese medicine	10,000	92.13	95.95

FSA: finite state automata.

classification accuracy in data recognition on the data sets originating from the speech recognition application departments than the other three approaches in clinics.

Focusing on the average accuracy (%) of the four approaches in Table 2, a multiple comparison *T*-test is implemented for manifesting the differences of multi-accent adaptive technology employed in the proposed system, with the other three technical approaches, and the results are shown in Table 3.

In Tables 2 and 3, we can see the classification accuracies of multi-accent adaptive technology (Group

Table 2. The data recognition accuracy (%) of the multi-accent adaptive technology.

Data sets	Words	Multi-accent adaptive technology (Group A)	Integration of general training data with accents labeling (Group B)	Integration of training model with clinical scenario data (Group C)	End-to-end medical scenario model (Group D)
Respiratory	5000	97.09	97.08	97.06	97.10
Digestive	5000	96.28	96.25	96.24	95.38
Cardiovascular	5000	96.46	96.47	96.43	96.40
Endocrinology	5000	96.87	96.84	96.84	96.66
Neurology	5000	96.41	96.31	96.23	94.78
Obstetrics and gynecology	5000	96.95	96.94	96.97	96.96
Pediatrics	5000	96.29	96.21	96.13	95.24
Ophthalmology	5000	96.05	96.05	95.85	94.28
Rheumatism	5000	96.24	96.21	95.12	93.64
Nephrology	5000	96.13	96.13	96.11	94.94
Orthopedics	5000	96.12	96.12	96.22	93.66
Traditional Chinese medicine	5000	96.06	96.06	95.23	92.80
Emergency	5000	96.14	96.14	96.14	94.94
Hematology	5000	96.23	95.89	95.74	93.30
ENT department	5000	96.16	96.06	96.03	94.96
Average accuracy (%)	5000	96.37	96.31	96.15	95.0

ENT: ear nose throat.

Table 3. Multiple comparison *T*-test results for the accuracy (%) results of the four approaches ($n = 4$).

Groups	*p*-value	$\alpha' = 2\alpha/n(n-1)$ $\alpha = 0.05$
A vs B	0.3537	0.0083
A vs C	0.1098	
A vs D	0.0003	
B vs C	0.1748	
B vs D	0.0005	
C vs D	0.0023	

Table 4. The recognition accuracy (%) of the integration model on different sizes of testing data sets.

Different sizes	Scales_test1 (19,928 words)	
	12G	5G
Accuracy (%)	97.23	95.69

A) totally perform better than those of the other three approaches (Groups B, C, and D). The results of Group B and Group C (training data with accents labeling and training model with clinical scenario data) show no significance with multi-accent adaptive technology, it means the multi-accent adaptive technology used in the proposed speech recognition EMR system can replace the manual labeling of the accent data, and the training model endowed with medical scenario data in real application scenarios; multi-accent adaptive technology outperforms significantly than the technical approach of Group D (end-to-end scenario model). As a new advanced technology, end-to-end scenario model was first proposed in 2017, which greatly improves the performances of voice recognition, and solves the problem that the losses of the lengths of output sequences in voice recognition.[13–15]

As manifested in Figure 4, BERT is employed as the main model in the integration of speech recognition model with the medical scenario, which separates the training process into two steps. The first step is called pre-training step, that is, the model obtained by learning language feature representations by self-monitoring method, which is called large-scale pre-training language model. The pre-training model can learn a lot of grammar and semantic knowledge points from a large number of medical records, books, and texts, and these knowledge points are stored in hundreds of millions of parameters of the model. The second step is from the pre-training model to the training model of scene tasks in the medical field. This stage is called the fine-tuning fusion of models; based on different byte sizes of data sets as testing materials, we can see an upgrading recognition accuracy (%) with the increment of data sets, with the manifestation (Table 4).

Table 5. Statistics of speech recognition rate of the pathology section in hospital.

Weeks	Right recognition words	Total recognition words	RR (%)
First	9679	12,963	74.67
Second	11,932	14,371	83.03
Third	13,747	15,344	89.59
Fourth	16,259	16,986	95.72

RR: recognition rate.

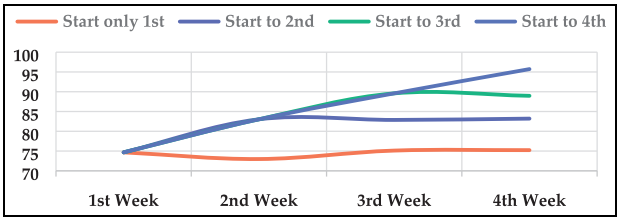

Figure 11. Online learning of accuracies of recognition rate (%) on different weeks.

The experiment of our proposed system in the medical ultrasound department lasted from October to November in 2020 for testing the continuous online study capability of the proposed speech recognition EMR. During the 4 weeks with the month, a total of 246,000 words from 308 patients have been input for recording tests. Two chief ultrasound physicians reviewed the tested electronic records. The accuracy of speech recognition is closely related to the availability of the system, so the RR is used as the evaluation index, and its calculation formula is shown in equation (1)[24]

$$R(\%) = \frac{\text{Right recognition words}}{\text{Total recognition words}} \quad (1)$$

The statistics based on the collected voice data of the online learning speech recognition system are made on the weekly recognition within 1 month after the system launch, as the results are listed in Table 5 and Figure 11. In the online learning of speech recognition, if every week is focused as a period unit, we can see the RR is raising step by step each week with the data learning function, which is also manifested in Figure 11. Before the speech recognition system launch, sorts of pathological reports within 1 year will be imported into the system for artificial intelligence training to ensure the initial RR of the pathological professional vocabularies. In the first week, the RR value is only 74.67%. Due to the speech recognition system has greatly changed doctors' working habits in the clinic, it

Table 6. Comparison of online learning module input and keyboard input, with other speech input methods.

Comparison	Keyboard input	Online learning module	Remarks
Record time	43 s	9 s	Average time from creation to finished edition
Check time	7 s	16 s	After finished edition, time cost of auditing reports
Transition time	46 min	26 min	Average time from specimen received to report published

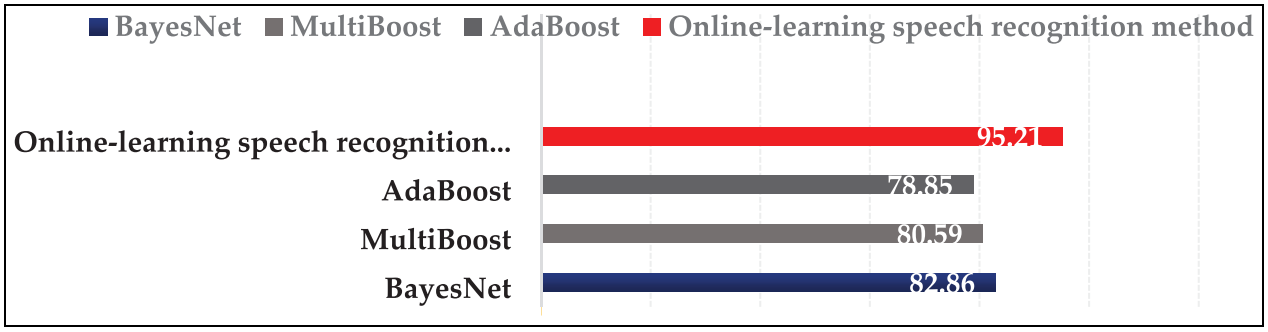

Figure 12. Comparison of online learning speech recognition module with other classification methods.

still runs at the initial stage. In addition, different doctors' accents also affect the accuracies of recognition. From the second week, the number of words and RR gradually increases, and in the fourth week, the RR value has an upgrade of 21.05% higher than that of in the first week, and it has reached 95.72%. It shows that after a period of adaptive learning, the speech recognition system can effectively overcome the personalized differences in pronunciation habits of doctors, or speaking accents, and the performance of recognition has been greatly improved. Obviously, with the help of online learning, the speech RR of the proposed module appears to upgrade step by step, following the accumulation of patients' data training.

The relevant indicators are statistically analyzed before and after the system launch, as shown in Table 6. From this table, we can see compared with the input mode of the keyboard, the doctor's voice entry report is more efficient, and the recording time is reduced obviously. Moreover, the report audit time of voice input is slightly longer than that of the keyboard. Because of the tires of doctors, it is feasible to produce changes in voice tones, raps, and other phenomena, which lead to a decrease in the accuracy of voice recognition. Therefore, it needs to spend more time checking and correcting the input reports in the audit stage. Generally speaking, the turnover time of a report through voice input is relatively shorter, which is due to the use of the speech recognition system to achieve the needs of pathological doctors while taking materials and inputting reports, it effectively reduces the time cost of taking materials, not only reducing the workload of doctors but also enhancing the timeliness of obtaining reports for patients.

In Table 6, we can see that the average time consumption of receive, input, and publication of electronic records by ISR is shortened to 56% (26/46), so ISR can perform better in electronic medical recordings than manual inputs through keyboards.

However, as we know, different approaches make differentiation in data classification or recognition inevitably.[25–28] For observing the performances of the feature selection methods in data classification, we adopt three classifiers of Boosting and Bayes methods as the testing classifiers by 10-fold cross-validation on the data samples of the tested patients, including AdaboostM1,[29] MultiBoostAB,[30] and BayesNet.[31] AdaboostM1 and MultiBoostAB are ensemble learning methods among the classifiers, so we use C4.5[32] as their basic unit classification classifiers. Moreover, the result is manifested in Figure 12. As a gradually upgraded approach, the online learning module performs better than ensemble learning methods in speech recognition.

Conclusion

Via technologies of speech recognition, we propose and construct an online EMR system, which achieve the medical information input and interaction between the

manual and computer points through the way of speech communication. And the proposed EMR speech recognition system combines a medicine term base, a systemic specialized speech model, and a personalized acoustic model, with a fault-tolerance mechanism for accent adjustment or rectification in speech recognition. For gaining an optimistic accuracy of speech recognition, an improved forward great matching algorithm has been proposed to achieve the enhanced speech recognition process during the information input and interaction. To accomplish the improved acoustic recognition, we build a medical speech data set availing audio and EMRs originating from the real environments in clinical and diagnosis services for patients. Based on this medical speech data set, the proposed algorithm obviously outperforms other machine learning algorithms in the speech recognition process. In this study, the proposed online EMR system has the following advantages over traditional EMR system:

First, our system shortens the time consumption of recording medical records for physicians, supplying clinicians more time to focus on the health conditions of their patients, as to implement and detail more effective treatment plans. Indirectly, the convenient interactions of speech recognition in EMR information inputs keep a good quality of medical records during the process of patients in hospital.
Second, the proposed system provides great conveniences for medical imaging staffs, it is known that medical imaging staffs have to review hundreds of images and issue reports in their daily works, and the voice input approach brings great conveniences to their daily work. Usually, it takes about 7 min to produce an examination report with traditional manual editing methods through computer keyboards, but the time consumption of data input in recordings can be shortened to about 3 min with the aid of the proposed intelligent voice recognition system.
Third, the adoption of the speech EMR system also optimizes the clinical inspection workflow. As an efficient way of medical record inputs, it speeds up the efficiencies of the test technicians, and a medical speech corpus for intelligent correction of text and homophone errors after speech recognition have also been built for improving clinical speech recognition accuracy and avoiding the pronunciation mistakes of medics in clinics.

Furthermore, compared to traditional EMR systems that rely on keyboard inputs, the proposed speech recognition system serves more efficient in clinics, and its higher speech recognition accuracy rate keep increases with a longer online time in utilization, which can form less recording mistakes and decrease the time consumption in modification of medical records, and because of the proposed speech recognition, EMR is developed on a medical term base as the knowledge database, so it has a good generalized application in the clinical scenarios of different hospital sections or departments.

Notwithstanding, this study has proposed an speech recognition technology approach for revolutionizing the EMR application in clinics in initial; the effects of the proposed speech recognition system still need to be in observation in more clinical scenarios, and we will continue to study the further enhancements of the speech recognition algorithm, enlargements of medical term base, and the upgrades of recognition models in the future.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the General Program of National Natural Science Foundation of China (NSFC; grant no. 61806147), the National Natural Science Foundation of China under grants nos 71690234 and 61573257, and the Science and Technology Commission of Shanghai Municipality under grant no. 19JG0500700.

ORCID iD

Jianwei Lu https://orcid.org/0000-0002-4283-4973

References

1. Fei W, Chuan H, Hao L, et al. Application research of medical scene intelligent speech recognition technology. *China Digit Med* 2019; 14(12): 19–21.
2. de bruijn B, Cherry C, Kiritchenko S, et al. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 2011; 18(5): 557–562.
3. Faravelon A and Verdier C. Towards a framework for privacy preserving medical data mining based on standard medical classifications. In: *International conference on electronic healthcare*, Casablanca, Morocco, 13–15 December 2010, pp.204–211. New York: *Springer*.
4. Hu Y, Lin WC, Tsai CF, et al. An efficient data preprocessing approach for large scale medical data mining. *Technol Health Care* 2015; 23(2): 153–160.
5. HIMSS Analytics. Electronic Medical Record Adoption Model (EMRAM), 20 January 2021, https://www.himss.org/what-we-do-solutions/digital-health-transformation/maturity-models/electronic-medical-record-adoption-model-emram
6. HIMSS Report. Electronic Clinical Quality Measure Reporting for the Inpatient Quality Reporting Program Letter, 3 August 2020, https://www.himss.org/resources/

electronic-clinical-quality-measure-reporting-inpatient-quality-reporting-program-letter
7. Yao Y and Huang Z. Bi-directional LSTM recurrent neural network for Chinese word segmentation. In: *International conference on neural information processing*, Kyoto, Japan, 16–21 October 2016, pp.345–353. New York: *IEEE*.
8. Xue S and Yan Z. Improving latency-controlled BLSTM acoustic models for online speech recognition. In: *2017 IEEE international conference on acoustics, speech and signal processing*, New Orleans, LA, 05–09 March 2017, pp.5340–5344. New York: IEEE.
9. Rabiner LR. A tutorial on hidden Markov models and select applications in speech recognition. *Proc IEEE* 1990; 77(2): 267–296.
10. Gers FA, Schmidhuber J and Cummins F. Learning to forget: continual prediction with LSTM. *Neural Comput* 2000; 12(10): 2451–2471.
11. Wigington C, Stewart S, Davis B, et al. Data augmentation for recognition of handwritten words and lines using a CNN-LSTM network. In: *2017 14th IAPR international conference on document analysis and recognition*, Kyoto, Japan, 9–17 November 2017, Vol 1, pp.639–645. New York: IEEE.
12. Kalchbrenner N, Grefenstette E and Blunsom P. A convolutional neural network for modelling sentences. *Arxiv [preprint]* 2014. DOI: 10.48550/arXiv.1404.2188.
13. Pundak G and Sainath TN. Lower frame rate neural network acoustic models. In: *Interspeech 2016*, San Francisco, CA, 8–12 September 2016, pp.22–26, https://www.isca-speech.org/archive/pdfs/interspeech_2016/pundak16_interspeech.pdf
14. W,Chan N, Jaitly QV, et al. Listen, attend and spell. *arXiv* [preprint]. Doi: 10.48550/arXiv.1508.01211.
15. Prabhavalkar R, Rao K, Sainath TN, et al. A comparison of sequence-to-sequence models for speech recognition. In: *Interspeech 2017*, Stockholm, Sweden, 20–24 August 2017, pp.939–943, https://www.isca-speech.org/archive_v0/Interspeech_2017/pdfs/0233.PDF
16. Snyder D, Garcia-Romero D, Sell G, et al. X-vectors: robust DNN embeddings for speaker recognition. In: *2018 IEEE international conference on acoustics, speech and signal Processing*, Calgary, AB, Canada, 15–20 April 2018, pp.5329–5333. New York: *IEEE*.
17. Liu B, Nie S, Zhang Y, et al. Boosting noise robustness of acoustic model via deep adversarial training. In: *2018 IEEE international conference on acoustics, speech and signal processing*, Calgary, AB, Canada, 15–20 April 2018, pp.5034–5038. New York: *IEEE*.
18. Zhou P, Yang W, Chen W, et al. Modality attention for end-to-end audio-visual speech recognition. In: *ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing*, Brighton, 12–17 May 2019, pp.3995–3999. New York: IEEE.
19. Chiu CC, Sainath TN, Wu Y, et al. State-of-the-art speech recognition with sequence-to-sequence models. In: *2018 IEEE international conference on acoustics, speech and signal processing*, Calgary, AB, Canada, 15–20 April 2018, pp.6565–6569. New York: *IEEE*.
20. Afouras T, Chung JS, Senior A, et al. Deep audio-visual speech recognition. *IEEE Trans Pattern Anal Mach Intell*. Epub ahead of print 21 December 2018. DOI: 10.1109/TPAMI.2018.2889052.
21. Braun S and Liu SU. Parameter uncertainty for end-to-end speech recognition. In: *ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing*, Brighton, 12–17 May 2019, pp.5636–5640. New York: IEEE.
22. Véniat T, Schwander O and Denoyer L. Stochastic adaptive neural architecture search for keyword spotting. In: *ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing*, Brighton, 12–17 May 2019, pp.2842–2846. New York: *IEEE*.
23. Lee J, Choi HS, Jeon CB, et al. Adversarially trained end-to-end Korean singing voice synthesis system. In: *20th annual conference of the international speech communication association*, *Graz*, Austria, 15–19 September 2019, pp.2588–2592, https://www.isca-speech.org/archive_v0/Interspeech_2019/pdfs/1722.pdf
24. Yinyin Y and Xudong W. Application of speech recognition technology in dental outpatient medical record system. *Henan Sci Technol* 2019(23): 36–38.
25. Qian H, Peng W, Feng Z, et al. R & D and application of medical intelligent speech recognition system. *China Digit Med* 2018; 13(10): 5–8.
26. Sheng D, Qinglong L, Zhengzhe Y, et al. Speech recognition system of intelligent customer service based on BLSTM network. *N Media Technol Netw* 2019; 8(2): 36–40.
27. Wang JH and Liu K. Feature-based fuzzy neural network approach for intrusion data classification. *J Comput Intell Electron Syst* 2012; 1: 99–103.
28. Quadri SA and Sidek O. Role of algorithm engineering in data fusion algorithms. *J Comput Intell Electron Syst* 2013; 2: 29–35.
29. Freund Y and Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997; 55: 119–139.
30. Webb GI. MultiBoosting: a technique for combining boosting and wagging. *Mach Learn* 2000; 40, 159–196.
31. Wang X, Qu H, Liu P, et al. A self-learning expert system for diagnosis in traditional Chinese medicine. *Expert Syst Appl* 2004;26: 557–566.
32. Quinlan JR. *C4.5: programs for machine learning*. 1st ed. Burlington, MA: Morgan Kaufmann, 1993, pp.17–25.