Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 153 (2023) 106517

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Intelligent speech technologies for transcription, disease diagnosis, and medical equipment interactive control in smart hospitals: A review

Jun Zhang [a,*], Jingyue Wu [a], Yiyi Qiu [a], Aiguo Song [a], Weifeng Li [b], Xin Li [b], Yecheng Liu [c]

[a] The State Key Laboratory of Bioelectronics, School of Instrument Science and Engineering, Southeast University, Nanjing, 210096, China
[b] Department of Emergency Medicine, Guangdong Provincial People's Hospital, Guangdong Academy of Medical Sciences, Guangzhou, 510080, China
[c] Emergency Department, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Science and Peking Union Medical College, Beijing, 100730, China

ARTICLE INFO

Keywords:
Automatic speech recognition
Smart hospital
Machine learning
Transcription
Diagnosis
Human-computer interaction

ABSTRACT

The growing and aging of the world population have driven the shortage of medical resources in recent years, especially during the COVID-19 pandemic. Fortunately, the rapid development of robotics and artificial intelligence technologies help to adapt to the challenges in the healthcare field. Among them, intelligent speech technology (IST) has served doctors and patients to improve the efficiency of medical behavior and alleviate the medical burden. However, problems like noise interference in complex medical scenarios and pronunciation differences between patients and healthy people hamper the broad application of IST in hospitals. In recent years, technologies such as machine learning have developed rapidly in intelligent speech recognition, which is expected to solve these problems. This paper first introduces IST's procedure and system architecture and analyzes its application in medical scenarios. Secondly, we review existing IST applications in smart hospitals in detail, including electronic medical documentation, disease diagnosis and evaluation, and human-medical equipment interaction. In addition, we elaborate on an application case of IST in the early recognition, diagnosis, rehabilitation training, evaluation, and daily care of stroke patients. Finally, we discuss IST's limitations, challenges, and future directions in the medical field. Furthermore, we propose a novel medical voice analysis system architecture that employs active hardware, active software, and human-computer interaction to realize intelligent and evolvable speech recognition. This comprehensive review and the proposed architecture offer directions for future studies on IST and its applications in smart hospitals.

1. Introduction

The average lifespan of humans is increasing with the improvement of living standards and medical technology, leading to a rapidly aging population. The world's population aged 60 and over is expected to increase to 22% by 2050 [1], which poses numerous challenges to the healthcare system [2]. The aggravation of aging has caused an increase in healthcare costs and shortages in human and material resources. In addition, the unbalanced distribution of medical resources worldwide and the lack of advanced medical technology and equipment in underdeveloped areas, make some sudden diseases not treated timely and effectively [3]. Moreover, some early symptoms are often imperceptible, resulting in the aggravation of the diseases and the delay in the best treatment.

With the development of robotics and artificial intelligence (AI) technologies, machines can achieve more efficient and accurate disease diagnosis and assessment in some cases and replace nurses to assist patients in their lives, which alleviate the problem of insufficient medical resources. For example, intelligent image processing methods based on deep learning (DL) have been applied to processing X-ray, CT, ultrasound, and facial images for diagnosing diseases such as COVID-19 detection [4–6], paralysis assessment [7,8], and autism screening [9]. In addition, intelligent speech technology (IST) plays a critical role in smart hospitals because language is the most natural mean of communication between doctors and patients and contains much information, such as patients' identity, age, emotion, and even symptoms of diseases [10].

IST refers to the use of machine learning (ML) methods to process human vocal signals to obtain information and realize human-machine communication. In recent years, speech signals research has developed

* Corresponding author.
E-mail address: j.zhang@seu.edu.cn (J. Zhang).

https://doi.org/10.1016/j.compbiomed.2022.106517
Received 17 September 2022; Received in revised form 23 December 2022; Accepted 31 December 2022
Available online 5 January 2023

rapidly with ML. IST contains many research areas, such as Automatic Speech Recognition (ASR) [11], Voiceprint Recognition, and Speech Synthesis. After years of development, IST has made significant progress and has gradually been applied in social life. For example, Apple's Siri, Google and Baidu's speech-based search services, and smart speakers [12] have all entered people's lives and provided convenience for us.

There are many review articles on speech technologies in medical applications, such as medical reporting [13], clinical documentation [14], speech impairment assessment [15], and speech therapy [16], healthcare [10,17]. However, we still require the review of state-of-the-art IST applications in smart hospitals. The smart hospital is the key to significantly improving the efficiency of medical behavior, alleviating the medical burden, and strengthening the robustness of the medical system in response to public health events such as the COVID-19 pandemic. Therefore, the application of IST in smart hospitals and smart healthcare needs to be reviewed for further development.

As shown in Fig. 1, in addition to applying it in daily life, IST is a crucial part of smart hospitals to process vocal signals produced by healthy people and patients. It is gradually applied in medical and rehabilitation scenarios [17,18]. For example, IST can be used as a transcription tool to help doctors to record patient information such as personal information and chief complaints. It can also interactively guide patients to seek medical services. Moreover, IST can be an auxiliary tool for doctors to diagnose diseases preliminarily. At the same time, speech can identify patients' emotional states to help doctors communicate better with them. Furthermore, IST combined with robotics, Internet of Things (IoT) technology, and 5G communication technology can support identifying and monitoring early symptoms of diseases, healthcare for the elderly, telemedicine, etc.

This paper mainly introduces the latest research progress and applications of IST in the healthcare field, summarizes and analyzes the existing research from the perspective of technical realization, and proposes the current challenges and future development directions. The rest of the paper is organized as follows. Section 2 gives the search methodology. Section 3 introduces the typical flow of intelligent speech signal processing, the system architecture of ASR, and an overview of the IST in applications of medical scenarios. The applications of IST in electronic medical documentation, disease diagnosis and evaluation, and human-medical equipment interaction are reviewed in Sections 4, 5, and 6, respectively. Section 7 presents a case study of IST in stroke patients' early recognition, diagnosis, rehabilitation training, evaluation, and daily care. The limitations of current speech technologies in the applications of smart hospitals and future directions are proposed in Section 8. Finally, we conclude this work in Section 9.

2. Search methodology

We performed the literature search on Web of Science and ProQuest. The literature search included all available English-language journal articles published in peer-reviewed journals up to July 2022 to ensure the quality of this review. Moreover, in order to target only papers related to IST and healthcare, the following keyword combinations are searched limiting in the title and abstract: (hospital OR medical) AND (intelligent OR smart OR technology) AND (speech). Only Review articles and Research articles are included.

Fig. 2 illustrates the article selection process. The initial search returned 3389 articles. 227 articles were retained after removing duplicates, non-English language, and irrelevant to healthcare by screening the titles and abstracts. Then, 187 articles were retained after screening the full text and excluded the studies irrelevant to IST, transcription, disease diagnosis, and human-medical equipment interaction. Finally, we included 173 articles after removing the less relevant articles and dataset papers. The 173 articles are classified by the year of publication, as shown in Fig. 3. We also included 28 articles about the methods and algorithms of IST. Furthermore, 15 web pages of medical equipment using IST were also included.

3. Overview of intelligent speech technologies

Speech technology generally includes collecting, coding, transmitting, and processing speech signals. However, the speech signals of the doctors and patients collected in the hospital's public areas contain

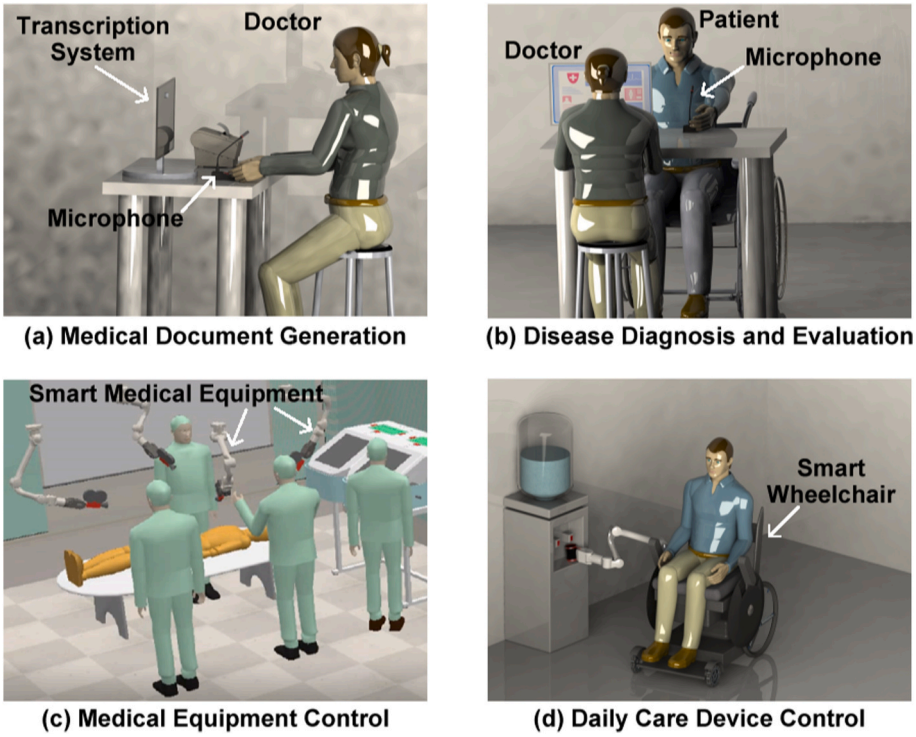

Fig. 1. Examples of the applications of intelligent speech technology (IST) in smart hospitals.

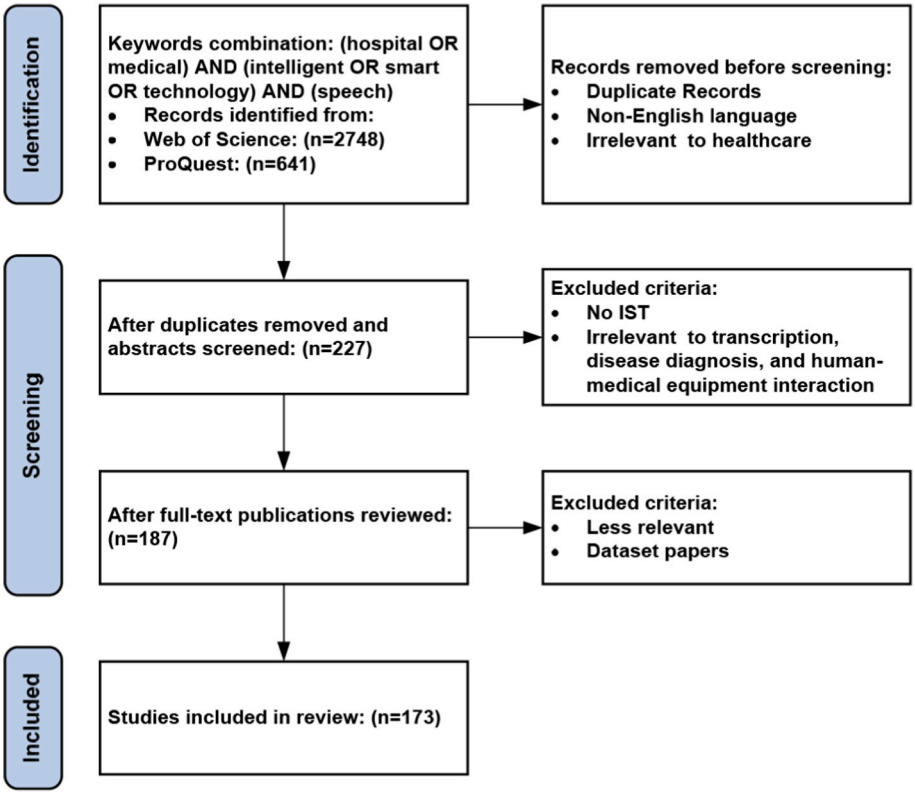

Fig. 2. Systematic Reviews selection process.

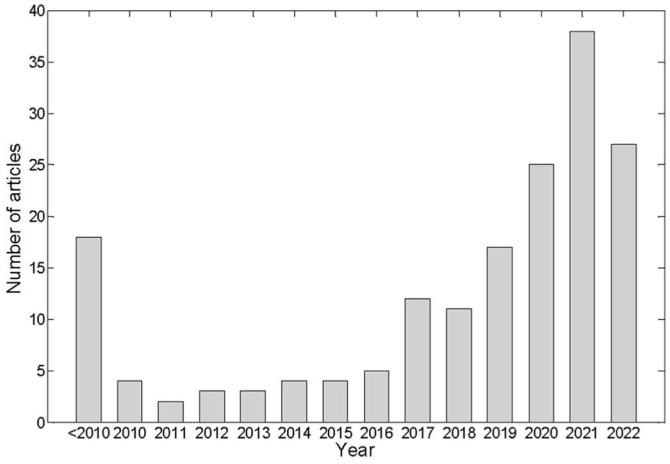

Fig. 3. The number of studies included in this review by the year of publication and its trend.

background noise. Moreover, some patients cannot speak and pronounce clearly due to illness or dialect. These issues bring challenges to the acquisition and processing of speech signals. We can upgrade the acquisition equipment for noise interference, such as using a microphone array, to suppress noise and acquire speech signals directionally [19,20]. In addition to noise suppression and collecting high-quality speech signals, the current research mainly focuses on their processing by state-of-the-art AI algorithms.

As shown in Fig. 4, speech signal processing mainly includes pre-processing, feature extraction, and recognition [21]. Among them, feature extraction and recognition are the critical steps of IST. Currently, the latest AI technologies are mainly used to improve the performance of feature extraction and recognition. Therefore, without loss of generality, this section first introduces the general flow of intelligent speech processing, presents the architecture of an ASR system, and then summarizes the application of IST in the medical field.

3.1. *Procedure of intelligent speech processing*

3.1.1. *Pre-processing*

The pre-processing of speech signals is the first step in IST. The speech signals are generally real-time audio streams and time sequences. There may be many invalid and silent segments in the speech signals that need to be segmented and filtered through the voice activity detection algorithm. Only the valid speech segments are retained for subsequent processing [22]. Hence, the speech signals are usually processed by pre-emphasis, framing, and windowing.

To improve the high-frequency resolution of the speech signals, they are usually pre-emphasized by using the first-order Finite Impulse Response high-pass digital filter [23]. The speech signals are time-varying signals. However, speech signals have short-term characteristics and can be treated as steady-state signals because the movement of the human muscles during speaking is slow. Therefore, the speech

Fig. 4. Typical processing flow of a speech system.

signals are needed to be divided into frames before processing and regarded as many short-term speech frames of equal length. Overlaps between adjacent speech frames are set during framing to ensure the short-term reliability of speech signal features and avoid feature mutation between adjacent speech signals.

Windowing is usually performed on each frame of the speech signals to reduce the error between the related speech segments and the original signals caused by the truncation of the voice signals. The commonly used window functions include the Rectangular window, Hanning window, and Hamming window [24]. We can obtain the speech signal required for feature extraction by processing each frame of the speech signals using these window functions with low-pass characteristics.

3.1.2. Feature extraction

The second step of IST is feature extraction, which is also crucial in determining the performance of the intelligent voice processing system. The feature extraction of speech signals aims to convert them into time-varying feature vector sequences through feature value extraction algorithms. The features of speech signals include time domain features, frequency domain features, and other transform domain features [26].

a) Time domain features

The common time domain features of speech signals include short-term amplitude, short-term energy, pitch period, pitch frequency, pitch, and zero-crossing rate. The short-term amplitude $M(i)$ is:

$$M(i) = \sum_{n=0}^{L-1} |y_i(n)|, 1 \le i \le N \tag{1}$$

The short-term energy $E(i)$ is:

$$E(i) = \sum_{n=0}^{L-1} y_i^2(n), 1 \le i \le N \tag{2}$$

where $y_i(n)$ refers to the amplitude of the n-th sample in the i-th frame of the speech signals, N is the total number of frames after framing, and L is the frame length. $M(i)$ and $E(i)$ are mainly used to distinguish the unvoiced and voiced segments in speech pronunciation. The difference between $M(i)$ and $E(i)$ is that the former has fewer fluctuations than the latter.

The pitch period is the vibration period of the vocal tract when a person makes a sound and is the reciprocal of the fundamental frequency F_0 [48], which can be estimated from the speech signal using pitch detection algorithms. Pitch represents the level of the sound frequency, which can be expressed by F_0 as

$$Pitch = 69 + 12 \times \log_2(F_0 / 440) \tag{3}$$

The zero-crossing rate $Z(i)$ refers to the number of changes of the sign of the sampled value in each frame of the speech signal:

$$Z(i) = \frac{1}{2} \sum_{n=0}^{L-1} |\mathrm{sgn}[y_i(n)] - \mathrm{sgn}[y_i(n-1)]| \tag{4}$$

where the symbolic function sgn[x] is:

$$\mathrm{sgn}[x] = \begin{cases} 1, x \ge 0 \\ -1, x < 0 \end{cases} \tag{5}$$

$Z(i)$ is also used to distinguish between unvoiced and voiced and is often combined with $E(i)$ for endpoint detection of speech segments, that is, the non-speech and speech segments. $Z(i)$ is more effective than $E(i)$ when there is considerable background noise.

b) Frequency domain features

The spectrum of the speech signal can be obtained by converting each frame of a time-domain speech signal to the frequency domain using the Fast Fourier Transform (FFT). The spectrum contains the frequency and amplitude information of the speech signal. The spectrum can only show the feature of one frame of the speech signal. Therefore, we can combine the spectrum of all speech frames to form a spectrogram to observe the frequency domain features of the whole speech signal. The spectrogram contains three kinds of information: frequency, time, and energy.

c) Other transform domain features

In addition to the characteristic parameters of speech signals commonly used in the time and frequency domains, researchers also use other characteristic parameters in the transform domain to improve the performance of the recognition. For example, the parameters in the transform domain can reflect the characteristics of people's vocal organs and auditory organs as speech features. Therefore, these feature parameters have a significant effect on speech signal recognition. Other domain features commonly used for speech signals include Mel Frequency Cepstral Coefficients (MFCC) [49,50], Discrete Wavelet Transform (DWT), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients, Perceptual Linear Prediction [51], and Line Spectral Frequency [26].

The above are common feature extraction methods in IST. However, in specific scenarios, we need to adjust the feature extraction method according to the type and characteristics of the collected signals and the performance of the speech recognition system. The extracted speech features are the input of speech recognition.

3.1.3. Recognition

Recognition based on the digital features of the speech signals is the final step in intelligent speech processing. There are many recognition algorithms. For example, Dynamic Time Warping (DTW) is a method for calculating the similarity of two temporal sequences. The similarity between the speech signal sample and the standard speech signal is obtained by comparing their feature sequences [52]. As shown in Fig. 5, DTW borrows the idea of dynamic programming, the minimum distance $D(i, j)$ between any time i and j of two sequences is

$$D(i,j) = Dist(i,j) + \min[D(i-1,j), D(i,j-1), D(i-1,j-1)] \tag{6}$$

where $Dist(i, j)$ is the relative distance between two speech signals at

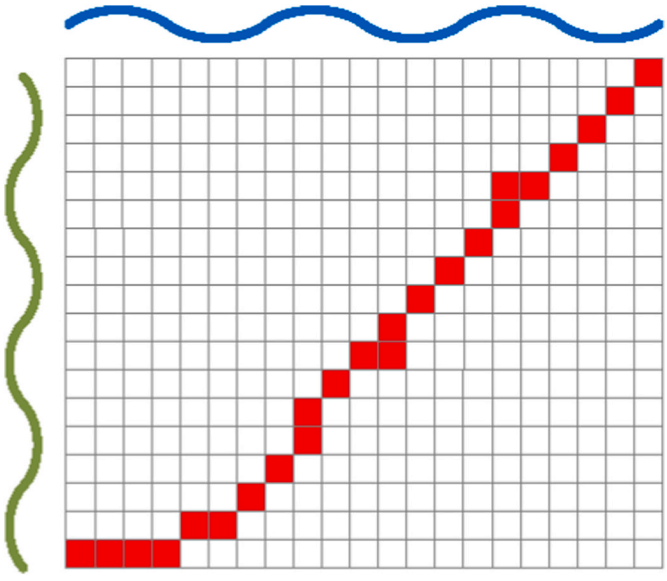

Fig. 5. Schematic diagram of the shortest path of dynamic time warping (DTW) algorithm.

times i and j, respectively. The distance generally is Euclidean distance. DTW requires less data and does not need pre-training, which is easy to implement and apply, and plays a vital role in small sample scenarios.

ML is the mainstream algorithm used in the current intelligent speech recognition. It utilizes the knowledge of probability and statistics and a dataset to train a model containing the mapping relationship between input and output to realize the feature recognition of speech signals. Table 1 shows the commonly used ML algorithms in medical speech signal processing. The traditional ML algorithms include Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Support Vector Machine (SVM), etc. The DL algorithms include Deep Neural Network (DNN), Convolutional Neural Network (CNN), and the Long Short-Term Memory (LSTM) algorithm in the Recurrent Neural Network (RNN), etc. Some of the algorithms are briefly introduced as follows.

Gaussian model is a one-dimensional variable Gaussian distribution

$$x \sim N\left(\mu, \delta^2\right). \tag{7}$$

As shown in Fig. 6, GMM refers to the superposition of multiple Gaussian models, and its variables are multi-dimensional vectors [53]. Then, the mixed Gaussian distribution $p(x)$ is generally represented by the mean and covariance matrix of the variables

$$p(x) = \frac{1}{(2\pi)^{D/2}|\sum|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \doteq N\left(x, \mu, \sum\right) \tag{8}$$

where the multidimensional variables x=(x_1, x_2, x_3, …, x_D), the covariance matrix is $\Sigma = \mathrm{E}[(x-\mu)(x-\mu)^{\mathrm{T}}]$, $\mu = \mathrm{E}(x)$. This model is usually trained using an Expectation-Maximization algorithm to obtain the maximum expectation on the training set.

Markov chains represent the transition relationship of states. As shown in Fig. 7, HMM adds the mapping from observations and states based on Markov chains [54]. a_{ij} is the probability of transitioning from the current state to the next state

$$a_{ij} = P(q_t = j | q_{t-1} = i) \quad i, j = 1, 2, 3, ..., N \tag{9}$$

b_t is the probability that the current state maps to the observed value

$$b_t = P(o_t | q_t) \quad t = 1, 2, 3, ..., N \tag{10}$$

We can use this model to establish the mapping relationship between the observation value and the actual state sequences. Then, the internal state with the highest probability can be found as the model's output, with input speech features as the observation value.

As illustrated in Fig. 8, the basic idea of SVM is to find an optimal hyperplane in a high-dimensional space for the segmentation of the binary classification problem. The hyperplane should ensure the minimum error rate of the classification [55]. The hyperplane in the high-dimensional space can be expressed as

$$W^T X + b = 0 \tag{11}$$

The training process of SVM is to find more suitable parameters W and b so that the hyperplane can better divide different categories.

In recent years, DL has considerably improved the performance of intelligent speech processing. As shown in Fig. 9(a), the basic unit of a neural network is a neuron. In addition to an input layer and an output layer, a DNN has multiple hidden layers. Each layer contains numerous neurons, fully connected between adjacent layers to form a network [56]. The output vector $\boldsymbol{v}^l$ of layer l is

$$\boldsymbol{v}^l = f\left(\boldsymbol{z}^l\right) = f\left(\boldsymbol{W}^l \boldsymbol{v}^{l-1} + \boldsymbol{b}^l\right), 0 < l < L \tag{12}$$

which is also the input vector of the next layer, where W^l and b^l are the weight matrix and the bias coefficient matrix of layer l, respectively. $\boldsymbol{v}^l \in \mathbf{R}^{Nl\times 1}$, $\boldsymbol{W}^l \in \mathbf{R}^{Nl\times Nl-1}$, $\boldsymbol{b}^l \in \mathbf{R}^{Nl\times 1}$, Nl is the number of the neurons in layer l. Therefore, by adjusting the model's W^l and b^l through the training data, we can establish connections among neurons in the current and previous layers and finally obtain the mapping relationship between the input and output.

As shown in Fig. 9(b), CNN mainly consists of two components. One is a convolutional layer composed of filters to calculate the local feature maps. $(h_k)ij$ refers to the k-th output feature obtained by the input feature unit at position (i, j)

$$(h_k)_{ij} = (W_k \otimes q) + b_k \tag{13}$$

where q represents the input feature unit, W_k and b_k represent the k-th filter and bias, respectively, obtained from the training data. Another component of the CNN is the pooling layer, which can reduce the dimensionality of each feature and retain only the more critical features. Finally, the last layer of the CNN is usually a fully connected layer, which is utilized to implement regression or classification tasks [57].

As illustrated in Fig. 9(c), the characteristic of the RNN is that it will be affected by the previous input while processing the current input, which can better process the time sequences [58]. The state transition and output of the hidden layer are:

$$\begin{cases} s_t = f(U * x_t + W * s_{t-1}) \\ o_t = g(V * s_t) \end{cases} \tag{14}$$

where s_t and s_{t-1} are the states of the hidden layer at time t and time $t-1$, respectively, o_t is the output of the network, W is the weight matrix converting state $t-1$ to the input of state t, U and V are the weight matrices of input and output, respectively.

Typical RNN has the problem of vanishing gradient. Hence, researchers propose LSTM networks to solve this problem [59]. In addition to these recognition algorithms, many researchers have proposed other algorithms, such as Generative Adversarial Networks and Variational Auto Encoders, etc., which are less related to this paper and will not be repeated here.

The performance of speech recognition algorithms based on DL is far better than those based on traditional ML algorithms. Especially the performance of speech recognition has been dramatically improved by the end-to-end algorithm based on Attention and Transformer in recent years. However, due to insufficient pathological speech data, traditional ML algorithms are still primarily used in pathological speech recognition.

Table 1
Common machine learning algorithms for medical speech signal processing.

Algorithm	Characteristics	Ref.
GMM	The probability density function of observed data samples using a multivariate Gaussian mixture density.	[27–29]
HMM	The Markov process is a double stochastic process in which there is an unobservable Markov chain defined by a state transition matrix. Each state of the chain is associated with a discrete or a continuous output probability distribution.	[30–33]
SVM	Support vector machine (SVM) is a binary classifier with advantages in few-shot classification, such as pathological voice detection.	[34–37]
DNN	Consists of fully connected layers and is popular in learning a hierarchy of invariant and discriminative features. Features learned by DNNs are more generalized than the traditional hand-crafted features.	[38–40]
CNN	A convolutional layer is the main building block of CNNs. Designed for image recognition but also extended for speech technology. Using the spectrogram of speech signals to classify them.	[41–44]
LSTM	A type of recurrent neural network (RNN) architecture and well-suited to learn from experience to classify, process, and predict time series when there are very long-time lags of unknown size between important events.	[45–47]

3.2. Automatic speech recognition system architecture

As one of the representative ISTs, speech recognition plays a vital role in healthcare. As shown in Fig. 10, speech recognition has

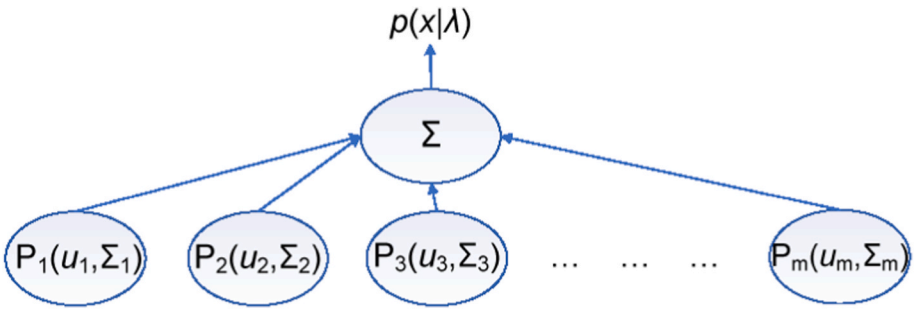

Fig. 6. Schematic diagram of Gaussian Mixture Model (GMM).

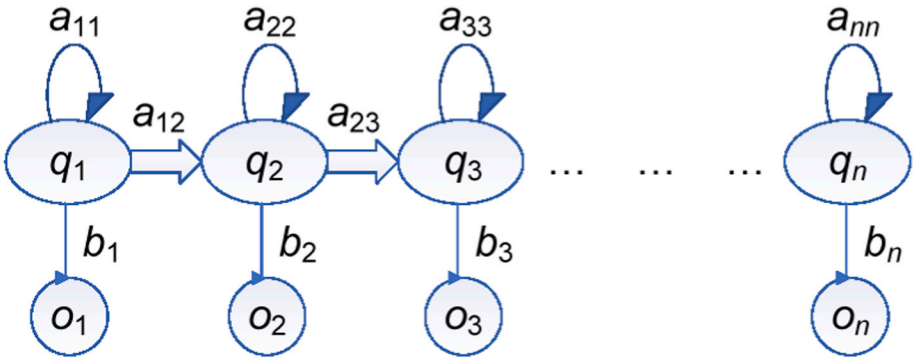

Fig. 7. Schematic diagram of Hidden Markov Model (HMM).

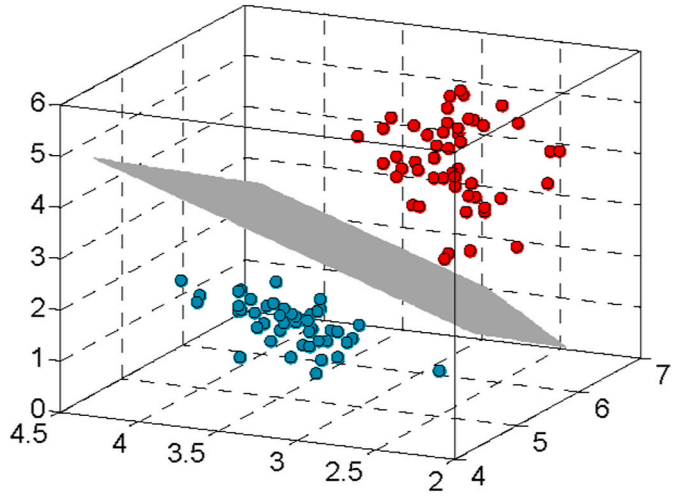

Fig. 8. Diagram of the hyperplane-based classification of Support Vector Machine (SVM).

developed over a long time, from the initial DTW algorithm to the later GMM-HMM algorithm and then to the algorithm combining DNN with GMM in recent years. They all have a process model. Pathological speech recognition is mainly based on traditional front-end and back-end architectures. The conventional architecture of a speech recognition system is briefly introduced as follows.

As shown in Fig. 11, a speech recognition system is generally divided into the front end and back end. The front end mainly completes speech signals' acquisition, pre-processing, and feature extraction. The back end realizes the recognition of the obtained speech feature sequences and gets the final recognition result. Unlike traditional architectures, the latest end-to-end speech recognition algorithms can directly convert speech signals into text or classification results, significantly improving speech recognition performance. The applications of these novel algorithms in medical speech recognition are attracting much attention. The state-of-the-art methods will be introduced in the following sections.

3.3. Intelligent speech technology in medical scenarios

Healthcare services and treatment are indispensable in human society. The health system's capacity is essential to people's life and health. However, a hospital can only treat a limited number of patients daily due to its capacity, which is more severe in densely populated and undeveloped areas. We can utilize medical resources more efficiently if the non-medical workload of doctors is reduced and their work efficiency is improved [60]. As the aging population increases, patients' timely treatment, rehabilitation, and daily care are essential for their health. Many studies are trying to apply IST in different medical scenarios, such as speech-based assistants, telemedicine, and health monitoring [61], to change the working ways of medical staff and improve the efficiency of the medical system.

This paper reviews the applications of IST in smart hospitals, mainly from three aspects. (1) Use IST to recognize the doctors' voices and reduce their time spent in non-medical related work, which was studied by researchers from an early stage [62,63]. (2) IST is also utilized to process the patients' speech signals to assist doctors in diagnosing and evaluating diseases [16]. This application has made significant breakthroughs in recent years with the development of ML and is also a hotspot of current research [64]. (3) IST is applied to medical equipment control to help doctors work efficiently [65,66]. These three aspects of applications are reviewed and summarized in the following three sections.

4. Speech recognition for electronic medical documentation

This section introduces the application of IST in electronic medical documentation, mainly including electronic medical record (EMR) transcription and electronic report generation. Then, we discuss some

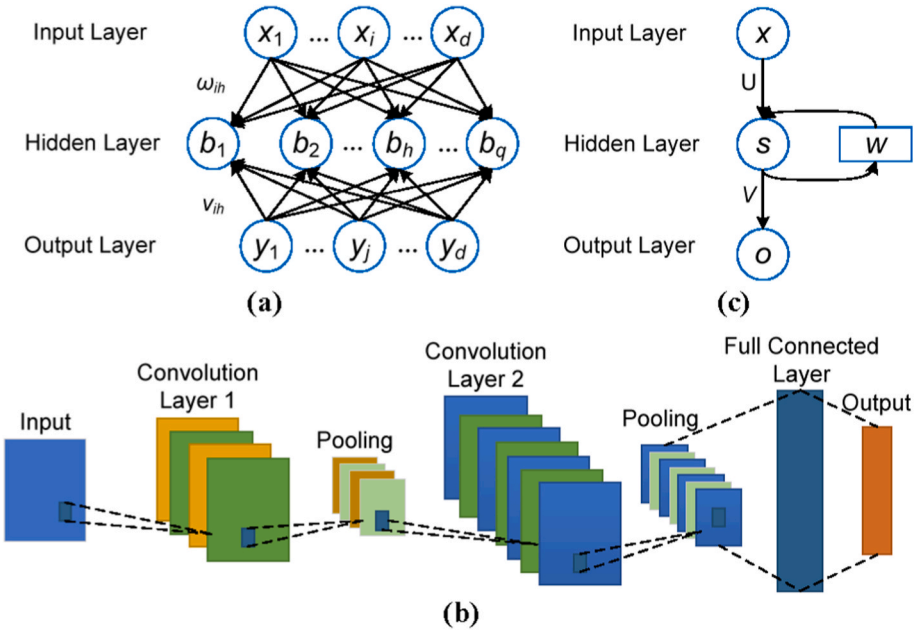

Fig. 9. Schematic diagram of several classic neural network models. (a) Deep Neural Network (DNN). (b) Convolutional Neural Network (CNN). (c) Recurrent Neural Network (RNN).

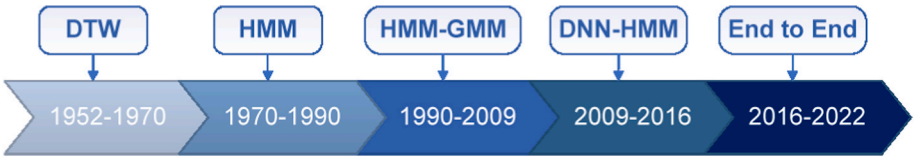

Fig. 10. Development process of the main technologies of speech recognition.

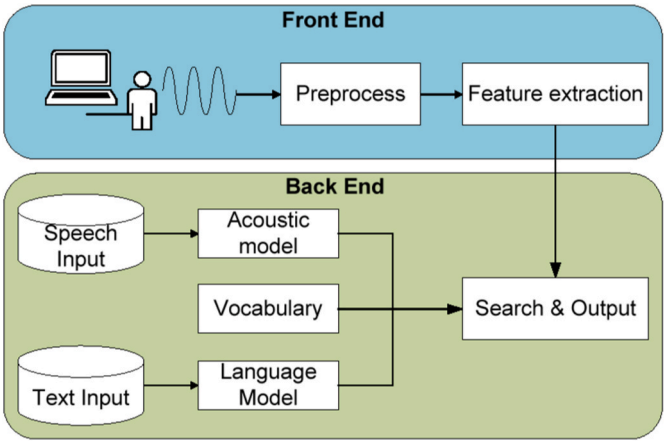

Fig. 11. Schematic diagram of the typical framework of speech recognition system with the front end and back end.

common issues of the existing medical transcription systems and typical solutions recently proposed. Finally, we present the critical indicators for evaluating the application effect of the transcription systems and their future directions.

Transcription refers to converting a speech signal into text using IST. The application of transcription in medical scenarios mainly refers to the generation of EMR and reports; that is, doctors' speech during diagnosis and pathological examination, the dialogue between doctors and patients, are all converted into text records. Transcription can reduce doctors' burden of manual document editing, allowing them to focus on medical work and thereby improving their work efficiency [67].

In addition, transcription significantly affects many aspects, such as doctors' enthusiasm for diagnosis, hospital treatment costs, and the treatment process [68]. Transcription also has some commercial value. Many companies have already developed some products. For example, Nuance designed an integrated healthcare system that could generate clinical records based on doctor-patient conversations [69]. A user survey has found that the system allowed physicians to devote more time to patients and their lives. Media Interface developed the Digital Patientendokumente product, which stored patient-related medical documents, nursing documents, and wills. This product allowed medical staff to review and sign patient documents quickly [70]. Unisound [71] and iFLYTEK [72] launched medical document entry systems, which effectively improved the work efficiency of medical staff. For instance, the entry system of iFLYTEK played an essential role in the fight against COVID-19.

4.1. Related studies and challenges

Using EMR transcription technology in medical scenarios has demonstrated apparent benefits. Table 2 shows that many researchers have used transcription technology to generate medical documents and investigated its application effects. They analyzed the accuracy, medical efficiency, and hospital cost of documentation by IST and proposed some problems and improvement methods.

Previous work has shown the effects and problems of transcription technology in medical document generation. For example, Ajami et al. investigated the previous medical transcription studies according to the usage scenario. Their results showed that the document generation performance was poor when the same vocabulary was used for different

Table 2
Examples of speech recognition technology in the application of electronic medical documentation.

Institute	Application scenario	Technical description	Application effect	Ref.
Zhejiang Provincial People's Hospital	Generate and extract pathological examination reports: 52h labeled pathological report recordings.	ASR system with Adaptive technology	Recognition rate = 77.87%; reduces labor costs; improves work efficiency and service quality	[81]
Western Paraná State University	Audios collected from 30 volunteers	Google API and Microsoft API integrated with the web	Reduces the time to elaborate reports in the radiology	[89]
University Hospital Mannheim	Lab test: 22 volunteers; Filed test: 2 male emergency physicians	IBM's Via-Voice Millennium Edition version 7.0	The overall recognition rate is about 85%. About 75% in emergency medical missions	[77]
Kerman University of Medical Sciences	Notes of hospitalized Patients from 2 groups of 35 nurses	Offline SR (Nevisa) Online SR (Speechtexter)	Users' technological literacy; Possibility of error report: handwritten < offline SR < online SR	[74]
University of North Carolina School of Medicine	6 radiologists dictated using speech-recognition software	PowerScribe 360 v4.0-SP2 reporting software	Near-significant increase in the rate of dictation errors; most errors are minor single incorrect words.	[79]
King Saud University	CENSREC-1 database: 422 utterances spoken by 110 speakers	Interlaced derivative pattern	99.78% and 97.30% accuracies using speeches recorded by microphone and smartphone	[18]
KPR Institute of Engineering and Technology	6660 medical speech transcription audio files and 1440 audio files from the RAVDESS dataset	Hybrid Speech Enhancement Algorithm	Minimum word error rates of 9.5% for medical speech and 7.6% for RAVDESS speech	[80]
Simon Fraser University	Co-occurrence statistics for 2700 anonymized magnetic resonance imaging reports	Dragon Naturally Speaking speech-recognition system; Bayes' theorem	Error detection rate as high as 96% in some cases	[83]
Graz University of Technology	239 clinical reports	Semantic and phonetic automatic reconstruction	Relative word error rate reduction of 7.74%	[25]
Zhejiang University	Radiology Information System Records	Synthetic method	About 3% superior to the traditional MAP + MLLR	[49]
Brigham and Women's Hospital	Records of 10 physicians who had used SR for at least 6 months	Morae usability software	Dictated notes have higher mean quality considering uncorrected errors and document time.	[75]

purposes. In addition, they found that although the use of speech recognition in the radiology report generation saved much time, the strict error checking in the later stage caused an increase in the overall turnaround time due to the high accuracy requirements of the report [73]. Peivandi et al. [74] and Poder et al. [13] also made a similar point that speech recognition accuracy was not as good as the accuracy of manual transcription. Although speech recognition has dramatically shortened the turnaround time of reports, doctors need to spend more time on dictation and correction due to the higher error rate of transcription [13].

Moreover, the advantages of electronic report generation are offset by the doctor's burden of verification and the risk of extra errors in the report. At the same time, previous studies have found considerable differences in the efficiency improvement of using transcription technology in different departments. By studying the previous work, Blackley et al. obtained some valuable and novel insights. For example, they found significant differences in the types and frequencies of words used when dictating and typing documents [75]. These differences may affect the quality of the documentation. They also found a lack of a unified and effective method for evaluating the impact of IST in medical scenarios [17].

The effects of transcription technology in medical scenarios include positive and negative aspects. The main advantages include reducing the turnaround time of most texts and quickly uploading the texts to the patient's electronic health record. Transcription also ensures the correctness of electronic documents in some scenarios that require multiple transcriptions and copies. In addition, transcription frees the doctors' eyes and hands, improves work efficiency in some scenarios, and brings them positive emotions [76]. Furthermore, in emergency medical missions, transcription technology can better meet the requirements for accurate time recordings of resuscitation than traditional methods [77]. Moreover, medical documents produced by transcription systems are more concise, standardized, and maintainable.

Negatively, there are potential recognition errors in the documents, resulting in the turnaround time not being shortened as expected in the scenarios with high accuracy requirements [78]. In addition, the delays in speech signal processing make doctors and patients lose patience with IST. Moreover, the background noise in public areas of hospitals, non-standard pronunciation, interruptions during speaking, and wearing surgical masks [79] will lead to decreased recognition accuracy and affect the mood of doctors and patients and their acceptance of IST [73].

4.2. Solutions for performance improvements

The critical question of medical transcription technology is continuous speech recognition. The current continuous speech recognition technology has high accuracy in most scenarios. However, improvements can be made in different processing stages of speech recognition to ensure accuracy and overcome the problems of IST in medical scenarios. Some improvement schemes have been proposed in several studies.

There are some methods to improve the adaptability of transcription systems. For the background noise problem, the microphone array combined with noise reduction algorithms can reduce the impact of the noise [19]. As shown in Fig. 12(a), Gnanamanickam et al. proposed a cascaded speech enhancement algorithm using HMM to optimize the algorithm of nonlinear spectral subtraction, which improved the effect of medical speech recognition [80]. For different department scenarios, Duan et al. added the noise of the corresponding department when training the acoustic model [81]. They combined the knowledge transfer technique to improve the adaptability of the acoustic model and its recognition performance in specific application scenarios.

Regarding acoustic models, Muhammad et al. proposed a feature extraction technique less affected by noise, the interlaced derivative pattern, which achieved higher accuracy and shorter recognition time in a cloud computing-based speech medical framework [18]. In terms of language models, according to the different types of generated medical documents and the various probabilities of lexical occurrences, training the corresponding language models in a targeted manner is a method to improve recognition accuracy. As shown in Fig. 12(b), to make the model more adaptable in different departments, Wu et al. introduced a simplified Maximum Likelihood Linear Regression (MLLR) into the incremental Maximum A Posteriori (MAP) process to enable the parameters to be continuously adjusted according to the speech and text [49]. Speech transcription technology has also been applied to some products. For example, Unisound developed a pathology entry system for the radiology department [71]. The system can free the doctors' hands and

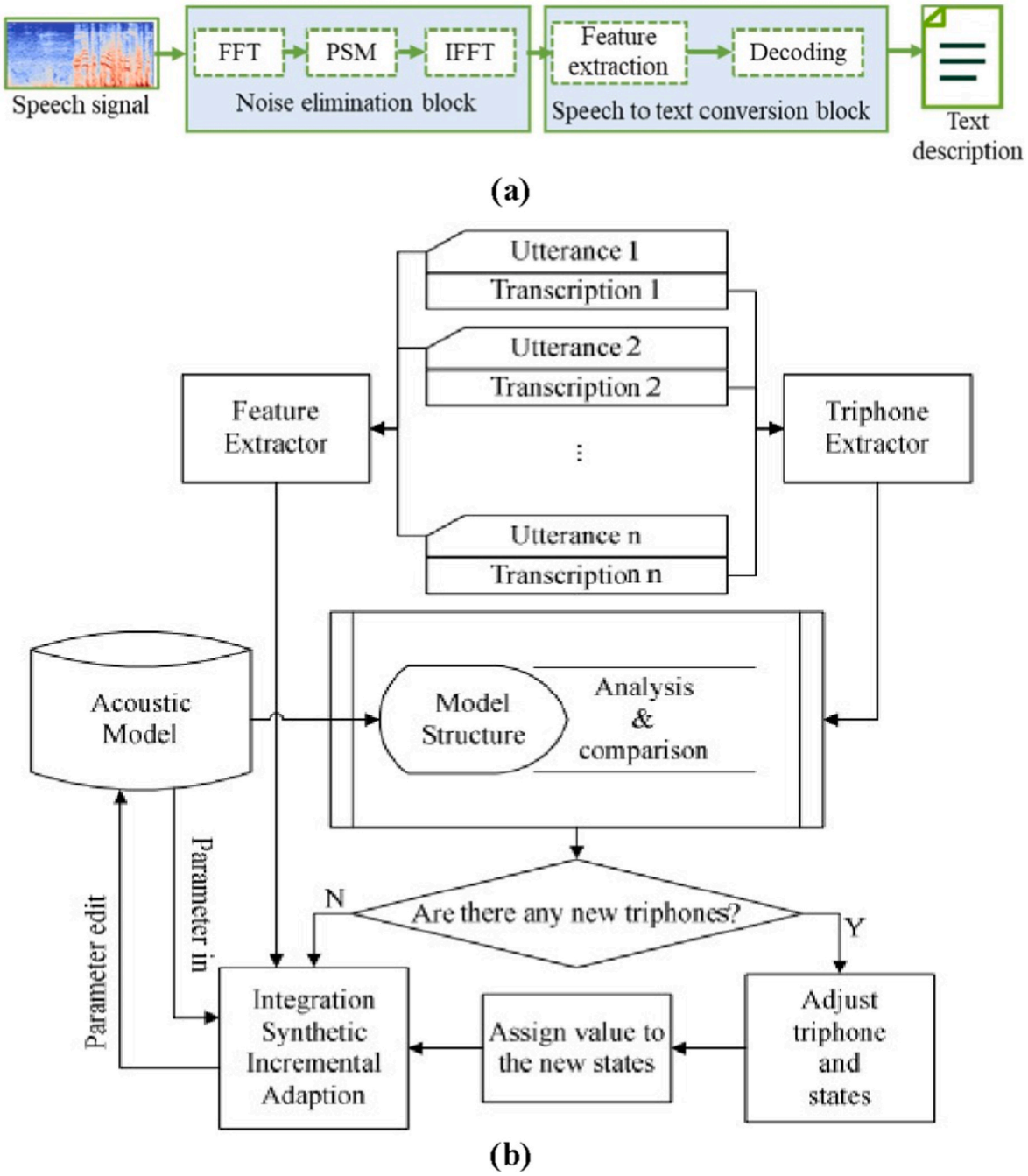

Fig. 12. Diagram of typical improvement scheme of transcription systems. (a) A hybrid enhancement algorithm for speech signal [80]. (b) An adaptation strategy for acoustic model [49].

allow them to enter the examination report while observing the image of the lesion. iFLYTEK also designed a medical document generation system for the dental department [82]. By wearing a small microphone, dentists can record information about the patient's condition during oral diagnosis.

Researchers also proposed methods to improve the quality of reports generated by transcription systems. For example, correction reports in electronic documents usually cause the problem of massive waste of resources [62]. Voll et al. proposed a text error correction scheme in post-processing for different medical documents to address this problem [83]. After the radiology report was generated, the frequency of different words appearing in the context was used to correct the report and mark the keywords, which was convenient for manual proofreading to shorten the document generation time [83]. In addition, Klann et al. proposed that using the Key-Val method to structure the report could reduce errors and improve its quality [84].

Sharing and security of electronic medical documents are also important issues. As shown in Fig. 13(a), Muhammad et al. proposed an Internet-based cloud service architecture, which can realize unified management of electronic medical documents and facilitate communication between doctors and patients whenever and wherever possible. However, some scenarios have time delays and data security problems [18]. As shown in Fig. 13(b), Qin et al. proposed a hospital intelligence framework based on cloud computing and fog computing to alleviate the delay problem. The service nodes are deployed in the hospital, which can improve the quality of the voice transcription service [85] and

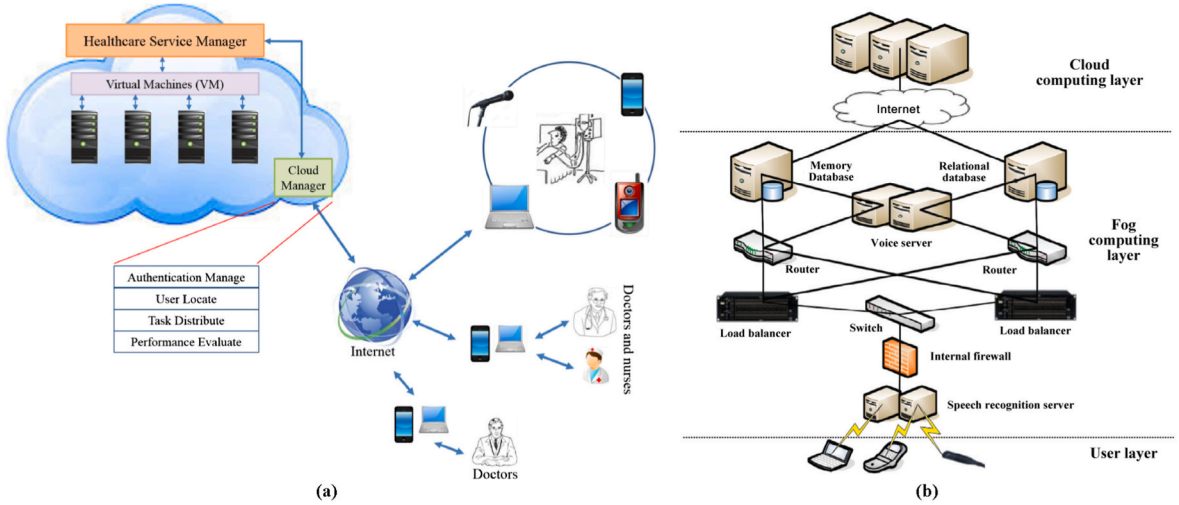

Fig. 13. Typical cloud computing-based voice medical frameworks. (a) A cloud-based framework for speech-enabled healthcare [18]. (b) A medical big data fog computing system [85].

ensure the security of the data. Singh et al. also presented an architecture similar to the one shown in Fig. 13(b). In addition, they proposed adding an IoT layer as a data source [86] so that guardians could obtain a real-time alert on students' overall emotions in response to their stressful situations.

4.3. Summary and discussion

Accuracy is a significant indicator for electronic document and report generation systems used in medical scenarios [87]. We also should pay more attention to the efficiency improvement of hospitals after using these systems [88]. Therefore, four key evaluation indicators shown in Table 3 can be referred to when we evaluate these systems. The four indicators reflect the primary concerns of doctors and patients in actual medical scenarios.

a) Report average turnaround time can measure the improvement of medical efficiency. Reducing this time is the primary purpose of applying transcription technology in medical scenarios.
b) The average number of critical errors in the generated medical documents can measure the reliability of the transcription system. Healthcare is related to the patient's health, so an error-prone transcription system is unacceptable.
c) The average word error rate of the generated documents will increase the time for medical staff to correct errors and affect the patient experience. We can quantitatively evaluate the above three indicators through the generated medical documents.
d) Questionnaires and other methods need to be adopted to assess the user experience of the medical staff and patients in different departments and scenarios to serve as a benchmark for improving the transcription system.

Table 3
Key evaluation metrics for transcription systems.

Indicators	Definition	Meaning	Ref.
Report average turnaround time	Average time from the start of report generation to patient accessibility	Turnaround time reduction reflects medical efficiency improvement brought by the transcription system.	[13]
Average number of critical errors	Number of medically misleading errors in generated documents	Reflects the reliability of the transcription system.	[87]
Average word error rate	Number of typos in generated documents	Reflects the effect of the document and influences the satisfaction of doctors.	[13] [87]
User experience of doctors and patients	Satisfaction of doctors and patients with all aspects of the generated documents	Improving work efficiency and user experience and reducing medical burden are goals of transcription systems.	[90] [91]

The interaction between the system and doctors should be considered a priority in the future development of medical transcription systems. Firstly, a more reasonable transcription process can be designed according to different departments so medical staff can use transcription tools efficiently after training. Secondly, we need to apply new speech recognition solutions in other fields to medical scenarios to enhance the reliability of the electronic medical documentation system. Thirdly, it can also start from the post-processing stage to improve the system's error correction capability and adaptability in generating different types of documents to provide doctors convenience [25].

5. Pathological voice recognition for diagnosis and evaluation

This section introduces the application of IST in disease diagnosis (disease unknown) and evaluation (disease known) using pathological voice. Then, we discuss data types, features, and recognition algorithms of pathological voices from a technical perspective. Finally, we present IST's future directions and trends in medical diagnosis. Since diseases can affect the patient's normal speech, cause them to cough and sneeze, and even make their breathing voice abnormal, we have investigated speech signals and other voice types in this section for disease diagnosis and evaluation.

5.1. Related studies and voice signal types

People express their feelings and thoughts by speaking. Speaking is accomplished through coordinated movements of the head, neck, and abdomen muscles. Individuals who cannot correctly coordinate these muscles will produce pathological speech [156]. Pathological speech-based disease diagnosis uses speech signal processing technologies to judge whether the patient suffers from certain diseases or to evaluate the patient's condition.

As shown in Table 4, many studies use speech technology to diagnose diseases that cause voice problems [157]. The diseases include Voice disorder [99], Acute decompensated heart failure [100], Alzheimer's Disease (AD) [104], Dysphonia [118], Parkinson's Disease (PD) [122, 125–128], Stroke [125,224], COVID-19 [130,132,135], Chronic

Table 4
Application of speech technology in pathological voice recognition and evaluation (otorhinolaryngology department).

Disease	Data sources	Voice type	Voice feature	Classifier	Effect	Ref.
Vocal Fold disorders	41 HP, 111 Ps	SV/a/	Jitter, RAP, Shimmer, APQ, MFCC Harmonic to Noise Ratio (HNR), SPI	ANN, GMM, HMM, SVM	Average classification rate in GMM reaches 95.2%	[92]
	KAY database: 53 HP, 94 Ps	SV/a/	Wavelet-packet coefficients, energy, and entropy, selected by algorithms	SVM, KNN	Best accuracy = 91%	[93]
	MEEI: 53 HP, 657 Ps	SV/a/	Features based on the phenomena of critical bandwidths	GMM	Best accuracy = 99.72%	[94]
Benign Vocal Fold Lesions	MEEI: 53 HP, 63 Ps; SVD: 869 HP, 108 Ps; Hospital Universitario Príncipe de Asturias (HUPA): 239 HP, 85 Ps; UEX-Voice: 30 HP, 84 Ps	SV/a/ and SS	MFCC, HNR, Energy, Normalized Noise Energy	Random-Forest (RF) and Multi-condition Training	Accuracies: about 95% in MEEI, 78% in HUPA, and 74% in SVD	[95]
Voice disorder	MEEI: 53 HP, 372 Ps SVD: 685 HP, 685 Ps VOICED: 58 HP, 150 Ps	SV/a/	Fundamental Frequency (F0), jitter, shimmer, HNR	Boosted Trees (BT), KNN, SVM, Decision Tree (DT), Naive Bayes (NB)	Best performance achieved by BT (AUC = 0.91)	[96]
	KAY: 213 Ps	SV/a/	Features are extracted through an adaptive wavelet filterbank	SVM	Sort six types of disorders successfully	[97]
	KAY: 57 HP, 653 Ps samples from Persian native speakers: 10 HP, 19 Ps	SV/a/	Same as above	SVM	Accuracy = 100% on both databases	[98]
	30 HP, 30 Ps	SV/a/	Daubechies' DWT, LPC	Least squares SVM	Accuracy >90%	[97]
	MEEI: 53 HP, 173 Ps	SV/a/ and SS	Linear Prediction Coefficients	GMM	Accuracy = 99.94% (voice disorder), Accuracy = 99.75% (running speech)	[101]
Dysphonia	Corpus Gesproken Nederlands corpus; EST speech database: 16 Ps; CHASING01 speech database: 5 Ps; Flemish COPAS pathological speech corpus: 122 HP, 197 Ps	SV/a/ and SS.	Gammatone filterbank features and bottleneck feature	Time-frequency CNN	Accuracy ≈89%	[144]
	TORGO Dataset: 8 HP, 7 Ps	SS	Mel-spectrogram	Transfer learning based CNN model	Accuracy = 97.73%,	[145]
	UA-Speech: 13 HP, 15 Ps	SS	Time- and frequency-domain glottal features and PCA-based glottal features	Multiclass-SVM	Best accuracy ≈ 69%	[146]
Pathological Voice	SVD: approximately 400 native Germans	SV/a/	Co-Occurrence Matrix	GMM	Accuracy reaches 99% only by voice	[102]
	MEEI: 53 HP SVD: 1500 Ps	SV/a/	Local binary pattern, MFCC	GMM, extreme learning machine	Best accuracy = 98.1%	[103]
	SVD	SV/a/ ,/i/,/u/	Multi-center and multi-threshold based ternary patterns and Features selected by Neighborhood Component Analysis	NB, KNN, DT, SVM, bagged tree, linear discriminant	Accuracy = 100%	[108]
	SVD: samples of speakers aged 15–60 years	SV/a/	Feature extracted from spectrograms by CNN	CNN, LSTM	Accuracy reaches 95.65%	[109]
Cyst Polyp Paralysis	SVD: 262 HP, 244 Ps MEEI: 53 HP, 95 Ps	SV/a/	spectrogram	CNN (VGG16 Net and Caffe-Net), SVM	Accuracy = 98.77% on SVD	[105]
	SVD: 686 HP, 1342 Ps	SV/a/ ,/i/,/u/ and SS	Spectro-temporal representation of the signal	Parallel CNN	Accuracy = 95.5%	[106]
Acute decompensated heart failure	1484 recordings from 40 patients	SS	time, frequency resolution, and linear versus perceptual (ear) mode	Similarity calculation and Cluster algorithm	94% of cases are tagged as different from the baseline	[100]
Common vocal diseases	FEMH data: 588 HP Phonotrauma data: 366 HP	SV/a/;	MFCC and medical record features	GMM and DNN, two stages DNN	Best accuracy = 87.26%	[107]

Application of speech technology in pathological voice recognition and evaluation (neurology department)

Disease	Data sources	Voice type	Voice feature	Classifier	Effect	Ref.
Parkinson's Disease (PD)	UCI Machine Learning repository: 8 HP, 23 Ps	SV	Features selected by the Relief algorithm	SVM and bacterial foraging algorithm	Best accuracy = 97.42%	[119]
	98 S	SV/a/, SS	OpenSMILE features, MPEG-7 features, etc.	RF	Best accuracy ≈80%	[120]
	UCI Machine Learning repository; Training: 20 HP, 20 Ps; Testing: 28 S	SV and SS	Wavelet Packet Transforms, MFCC, and the fusion	HMM, SVM	Best accuracy = 95.16%,	[121]
	Group 1: 28 PD Ps Group 2: 40 PD Ps	SS	Diadochokinetic sequences with repeated [pa], [ta], and [ka] syllables	Ordinal regression models	The [ka] model achieves agreements with human raters' perception	[122]
	Istanbul acoustic dataset (IAD) [123]: 74 PH, 188 Ps	SV/a/	MFCC, Wavelet and Tunable Q-Factor wavelet transform, Jitter, Shimmer, etc.	Three DTs.	Best accuracy = 94.12% on IAD and = 95% on SAD	[125]

(*continued on next page*)

Table 4 (*continued*)

Application of speech technology in pathological voice recognition and evaluation (neurology department)						
Disease	Data sources	Voice type	Voice feature	Classifier	Effect	Ref.
	Spanish acoustic dataset (SAD) [124]: 80 PH, 40 Ps Training: 392 HP, 106 Ps Testing: 80 HP, 40 Ps	SS	MFCC, Bark-band Energies (BBE) and F0, etc.	RF, SVM, LR, Multiple Instance Learning	The best model yielded 0.69/0.68/0.63/0.8 AUC for four languages	[126]
	Istanbul acoustic dataset: 74 HP, 188 Ps	SV/a/	MFCC, Deep Auto Encoder (DAE), SVM	LR, SVM, KNN, RF, GB, Stochastic Gradient Descent	Accuracy = 95.49%	[127]
	PC-GITA: 50 HP, 50 Ps SVD: 687 HP, 1355 Ps Vowels dataset: 1676 S	SV	Spectrogram	CNN	Best accuracy = 99%	[128]
Alzheimer's disease (AD)	50 HP, 20 Ps	SS	Fractal dimension and some features selected by algorithms	MLP, KNN	Best accuracy = 92.43% on AD	[104]
PD, Huntington's disease (HD), or dementia	8 HP, 7 Ps	SS	Pitch, Gammatone cepstral coefficients, MFCC, wavelet scattering transform	Bi-LSTM	Accuracy = 94.29%	[110]
Dementia	Two corpora recorded at the Hospital's memory clinic in Sheffield, UK; corpora 1: 30 Ps corpora 2: 12 Ps, 24 S	SS	44 features (20 conversation analysis based, 12 acoustic, and 12 lexical)	SVM	Accuracy = 90.9%	[111]
	DementiaBank Pitt Corpus [112]: 98 HP, 169 Ps PROMPT Database [113]: 72 HP, 91 Ps	SS	Combined Low-Level Descriptors (LLD) features extracted by openSMILE [114]	Gated CNN	Accuracy = 73.1% on Pitt Corpus and = 74.1 on PROMPT	[115]
Dysarthria	UA-Speech: 12 HP, 15 CP Ps MoSpeeDi: 20 HP, 20 Ps PC-GITA database [116]: 45 HP, 45 PD Ps	SS	Spectro-temporal subspace, MFCC, the frequency-dependent shape parameter	Grassmann Discriminant Analysis	Best accuracy = 96.3% on UA-Speech	[117]
	65 HP, 65 MS-positive Ps	SS	Seven features including Speech duration, vowel-to-recording ratio, etc.	SVM, RF, KNN, MLP, etc.	Accuracy = 82%	[118]
Distinguishing two kinds of dysarthria	174 HP, 76 Ps	SV and SS	Cepstral peak prominence	classification and regression tree; RF; Gradient Boosting Machine (GBM); XGBoost	Accuracy = 83%	[155]

Application of speech technology in pathological voice recognition and evaluation (respiratory department)						
Disease	Data sources	Voice type	Voice feature	Classifier	Effect	Ref.
COVID-19	130 HP, 69 Ps	SV/a/and cough	feature sets extracted with the openSMILE, open-source software, and Deep CNN, respectively	SVM and RF	Accuracy ≈80%	[129]
	Sonda Health COVID-19 2020 (SHC) dataset [130]: 44 HP, 22 Ps	SV and SS	Features (glottal, spectral, prosodic) extracted by COVAREP speech toolkit	DT	Feature-task combinations accuracy >80%	[131]
	Coswara: 490 HP, 54 Ps	SV/a/,/i/,/o/;	Fundamental, MFCC Frequency (F0), jitter, shimmer, HNR	SVM	Accuracy ≈ 97%	[132]
	DiCOVA Challenge dataset and COUGHVID: Training: 772 HP, 50 Ps Validation: 193 HP, 25 Ps Testing: 233 S	Cough	MFCC, Teager Energy Cepstral Coefficients TECC	Light GBM	The best result is 76.31%	[133]
	MSC-COVID-19 database: 260 S	SS	Mel spectrogram	SVM & Resnet	Assess patient status by sound is effective	[134]
	Integrated Portable Medical Assistant collected: 36 S	Cough and speech	Mel spectrogram, Local Ternary Pattern	SVM	Accuracy = 100%	[135]
	COUGHVID: more than 20,000 S Cambridge Dataset [136]: 660 HP, 204 Ps; Coswara: 1785 HP, 346 Ps	Cough	MFCC, spectral features, chroma features	Resnet and DNN	Sensitivity = 93%, specificity = 94%	[137]
	COUGHVID: 1010 Ps; Coswara: 400 Ps; Covid19-Cough: 682 Ps	Cough, breathing cycles, and SS	Mel-spectrograms and cochlea-grams, etc.	DCNN, Light GBM	AUC reaches 0.8	[138]
	Cambridge dataset: 330 HP, 195 Ps; Coswara: 1134 HP, 185 Ps; Virufy: 73 HP, 48 Ps; NoCoCODa: 73 Ps	Cough	audio features, including MFCC, Mel-Scaled Spectrogram, etc.	Extremely Randomized Trees, SVM, RF, MLP, KNN, etc.	AUC reaches 0.95	[139]
	Coswara: 1079 HP, 92 Ps Sarcos: 26 HP, 18 Ps	Cough	MFCC	LR, KNN, SVM, MLP, CNN, LSTM, Restnet50	AUC reaches 0.98	[140]
	Coswara, ComParE dataset, Sarcos dataset	Cough, breathing, sneeze, speech	Bottleneck feature	LR, SVM, KNN, MLP	AUC reaches 0.98	[141]
Chronic Obstructive Pulmonary Disease	25 HP, 30 Ps	respiratory sound signals	MFCC, LPC, etc.	SVM, KNN, LR, DT, etc.	Accuracies of SVM and LR are 100%	[142]
	429 respiratory sound samples	respiratory sound signals		SVM	Accuracy = 97.8% by HHT-MFCC-Energy	[143]

(*continued on next page*)

Table 4 (*continued*)

Application of speech technology in pathological voice recognition and evaluation (respiratory department)						
Disease	Data sources	Voice type	Voice feature	Classifier	Effect	Ref.
			MFCC; Hilbert-Huang Transform (HHT)-MFCC; HHT-MFCC-Energy			
Tuberculosis (TB)	21 HP, 17 Ps, cough recordings: 748	Cough	MFCC, Log spectral energy	LR	AUC reaches 0.95	[148]
	35 HP, 16 Ps, cough recordings:1358	Cough	MFCC, Log-filterbank energies, zero-crossing-rate, Kurtosis	LR, KNN, SVM, MLP, CNN	LR outperforms the other four classifiers, achieving an AUC of 0.86	[147]
	TASK, Sarcos, Brooklyn datasets: 21 HP, 17 Ps Wallacedene dataset: 16 Ps Coswara: 1079 HP, 92 Ps; ComParE: 398 HP, 199 Ps	Cough	MFCC	CNN, LSTM, Resnet50	Resnet50 AUC: 91.90% CNN AUC: 88.95% LSTM AUC: 88.84%	[149]

Application of speech technology in pathological voice recognition and evaluation. (Others)						
Disease	Data sources	Voice type	Voice feature	Classifier	Effect	Ref.
Juvenile Idiopathic Arthritis	5 HP, 3 Ps	Knee Acoustical	Spectral, MFCC, or band power feature	Gradient Boosted Trees, neural network	Accuracy = 92.3% using GBT, Accuracy = 72.9% using neural network	[150]
Stress	6 categories of emotions, namely: Surprise, Fear, Neutral, Anger, Sad, and Happy	SS (facial expressions, content of speech)	Mel scaled spectrogram	Multinomial Naïve Bayes, Bi-LSTM, CNN	Assess students' stress by facial expressions and speech is effective	[86]
Depression and Other Psychiatric Conditions	Gruop1: depression (DP) 27 S; Gruop2: other psychiatric conditions (OP) 12 S; Gruop3: normal controls (NC) 27 S	SS	Features extracted by openSMILE and Weka program [151]	Five multiclass classifier schemes of scikit-learn	Accuracy = 83.33%, sensitivity = 83.33%, and specificity = 91.67%	[152]
Depression	AVEC 2014 dataset: 84 S; TIMIT dataset	SS	TEO-CB-Auto-Env, Cepstral, Prosodic, Spectral, and Glottal, MFCC	Cosine similarity	Accuracy = 90%	[154]

SV=Sustained vowel, SS=Spontaneous speech, Ps = Patients, HP=Healthy People, S=Subjects.

Obstructive Pulmonary Disease [142,143], Aphasia [169,170,181], Tuberculosis (TB) [147–149], and organ lesions such as oral cancer [158], head and neck cancer [159], nodules, polyps, and Reinke's edema [95]. These studies are divided into four categories by diseases, including the otorhinolaryngology department, respiratory department, neurology department, and others, and are shown in the four sub-tables, respectively.

Most of the speech data used in these studies come from existing or small private datasets collected from medical institutions. For example, the frequently adopted pathological speech datasets include Parkinson's Telemonitoring Dataset [160], Saarbrucken Voice database (SVD) [161], Massachusetts Eye & Ear Infirmary (MEEI), TORGO [162], VOICED [163], University of California Irvine (UCI) Machine Learning repository [164], Universal Access Speech Database (UA-Speech) [165], Coswara database [166], the COUGHVID corpus [167], and Computational Paralinguistics ChallengE (ComParE) [190]. The above datasets contain pathological voices of many diseases, which provide convenience for IST-based diagnosis system research.

We can see from Table 4 that the accuracies of most studies are over 90% (sensitivity and specificity are not shown), which proves the feasibility of diagnosis through speech signals. Meanwhile, we can find that even if the same dataset is used to diagnose the same disease, there are significant differences between different studies. The main reasons include differences between identification methods and an implicit problem that different studies will screen the data in the same dataset. Therefore, it is not very meaningful to directly compare the recognition effects of different methods. Nevertheless, the trend and proposed methods can inspire our further research.

The studies surveyed have one or more recognition algorithms for processing pathological voices. As the statistical analysis results shown in Fig. 14, about 81% of the articles try ML methods, achieving satisfactory accuracy in disease diagnosis. In recent years, the proportion of DL methods has increased (42%), but ML methods are still the primary ones. A crucial reason is that the data is difficult to meet the needs of state-of-the-art end-to-end recognition methods. Therefore, some studies have tried solutions such as data augmentation [139,140,149,171] and transfer learning [141,145,149] to solve this problem. The details of the diagnosis systems in these studies, including data sources, voice type, voice feature, classifier, and effect, can be found in Table 4.

For feature extraction, in addition to the common features in the time domain and frequency domain, some studies also try rare features [94,103,104,117,121,127,133,137,144] or use existing feature sets for research, such as OpenSMILE features [114,115,120,129,152], features extracted by the Weka program [152], and COVAREP speech toolkit [131], MPEG-7 features [120]. Moreover, some studies extract features by DL algorithms, dimension reduction algorithms [108,111,119,146], or heuristic algorithms [93,97,98].

The main types of voice data are sustained vowels (SV), spontaneous speech (SS) sentences, coughs, and breathing sounds. SV signals are generally processed by collecting the SV articulations of patients [92,96,101,106]. The SS processing-based method uses sentence-level features to collect patient speech or continuous pronunciation of a given text as experimental data [110]. Because the voice types of speech signals differ, their research also has apparent differences in feature extraction and recognition methods. In addition, some studies directly use general speech transcription systems to evaluate the condition of patients. Fig. 15 shows the statistical analysis results. 38% of articles adopted SV as the speech signal, which also obtained the highest average accuracy. 18% of articles used both SV and SS as the speech signal. Although more data types are utilized, there is no significant performance improvement. It shows how we can extract information from different types of voices and combine them effectively is also an issue. Other types of voice signals account for 28% because coughing, breathing, and sneezing are the main diagnostic signals in diagnosing respiratory-related diseases.

5.2. Conventional methods

The research using SV data as the object is generally carried out from the quality and frequency domain characteristic parameters of pathological voice signals. Wang et al. combined MFCC with six speech quality

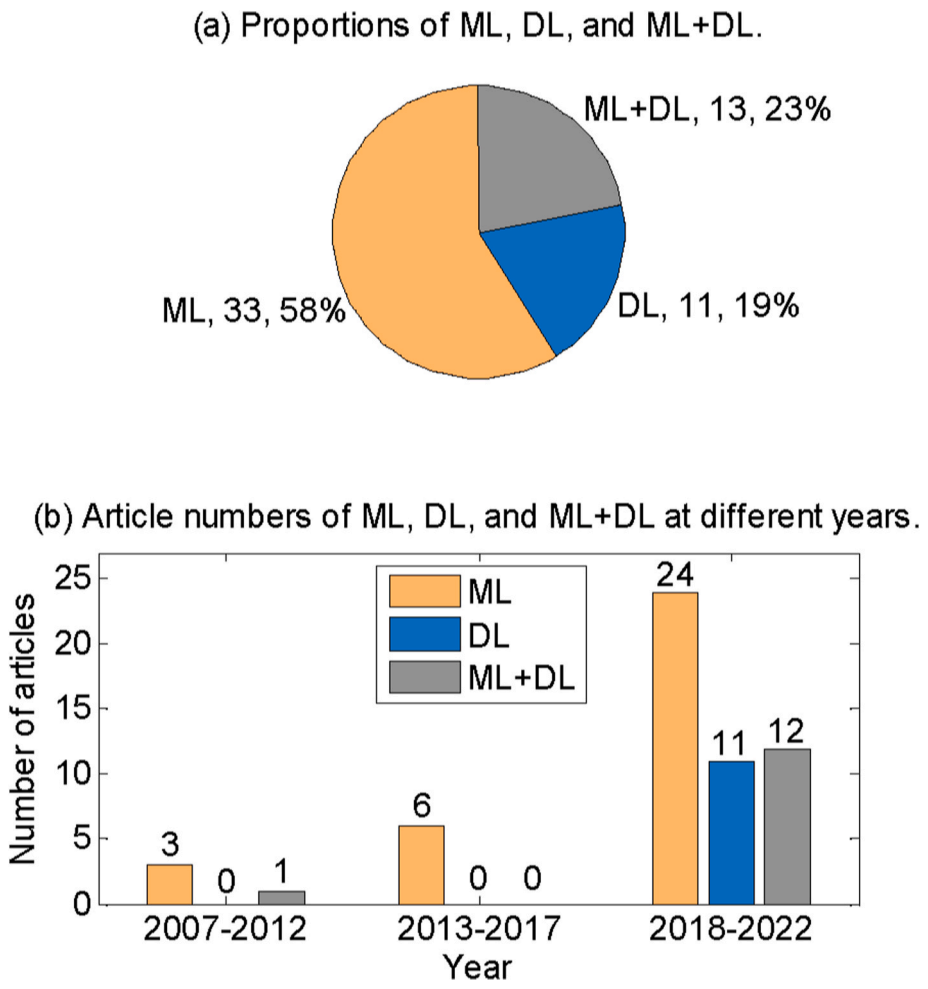

Fig. 14. Statistical analysis of traditional machine learning (ML), deep learning (DL), and ML + DL methods used in disease diagnosis.

features (jitter, shimmer, harmonic-to-noise ratio (HNR), soft phonation index (SPI), amplitude perturbation quotient (APQ), and relative average perturbation (RAP)) of the SV pronunciation/a/to recognize the pathological voice. They used HMM, GMM, SVM, and Artificial Neural Networks (ANN) to conduct two-class comparison experiments and found that the GMM method has the best classification accuracy, with an accuracy rate of 95.2% [92]. Similar research work was done by Verde et al. [96]. The difference between them is that Verde et al. extracted features that included the fundamental frequency F_0 of the speech signal. In addition, they used a boosted tree algorithm as a classifier to conduct an experimental study on data selected from three different databases. Ali et al. also adopted the patient's SV pronunciation/a/as the research object. They proposed features based on the phenomena of critical bandwidths and combined them with HMM to detect vocal cord disorders, with an accuracy rate of more than 95% [94]. Baird et al. extracted features such as pitch, intensity, and HNR from the SV in the Dusseldorf Anxiety Corpus to assess the anxiety of patients [172]. Their results verified the effectiveness of using speech-based features to predict anxiety and showed better recognition performance of higher-level anxiety.

For research based on sentence-level speech data, in addition to the quality and time-frequency domain characteristics, the prosodic characteristics of sentence data are also an effective breakthrough. Kim et al. adopted the speech signal parameters of phonemes, prosody, and speech quality as the features. They predicted the intelligibility of aphasia speech in the Korean database Quality-of-Life Technology using Support Vector Regression (SVR) [173]. They also proposed a structured sparse linear model containing phonological knowledge to predict the speech intelligibility of patients with dysarthria [174]. Martínez et al. assessed dysarthria intelligibility using i-vectors extracted by factor analysis from the supervector of universal GMM [175]. After being evaluated by SVR and Linear Prediction, the speeches in Wall Street Journal 1 and UA-Speech databases were divided into four levels: very low, low, mid, and high. Kadi et al. also used a set of prosodic features selected by linear discriminant analysis combined with SVM and GMM, respectively, to classify dysarthria speech of the Nemours database into four severity levels and got the best classification rate of 93% [176]. Kim et al. classified pathological voice using the features of abnormal changes in prosody, phonological quality, and pronunciation at the sentence level. The pathological speeches of the NKI CCRT Speech Corpus and the TORGO databases were classified into two categories (intelligible and incomprehensible), and posterior smoothing was performed after classification [177]. These studies all make use of the characteristics of prosody. However, different languages have different pronunciations in prosody, which means that compared with the model obtained by SVs, the model trained by this method has low generalization ability.

There are many other studies based on speech recognition technology [178]. As shown in Fig. 16, Liu et al. used speech recognition to extract features and then integrated traditional acoustic feature classification to assess the severity of the voice disorder [168]. Bhat et al. utilized a bidirectional LSTM network for binary classification of the speech intelligibility of dysarthria in the TORGO dataset [179]. They also compared the classification performances when using the features of MFCC, log filter banks, and i-vector. In addition, Dimauro et al. adopted Google's speech recognition system to convert patients' speech into text [180]. Their result showed that the PD group's recognition

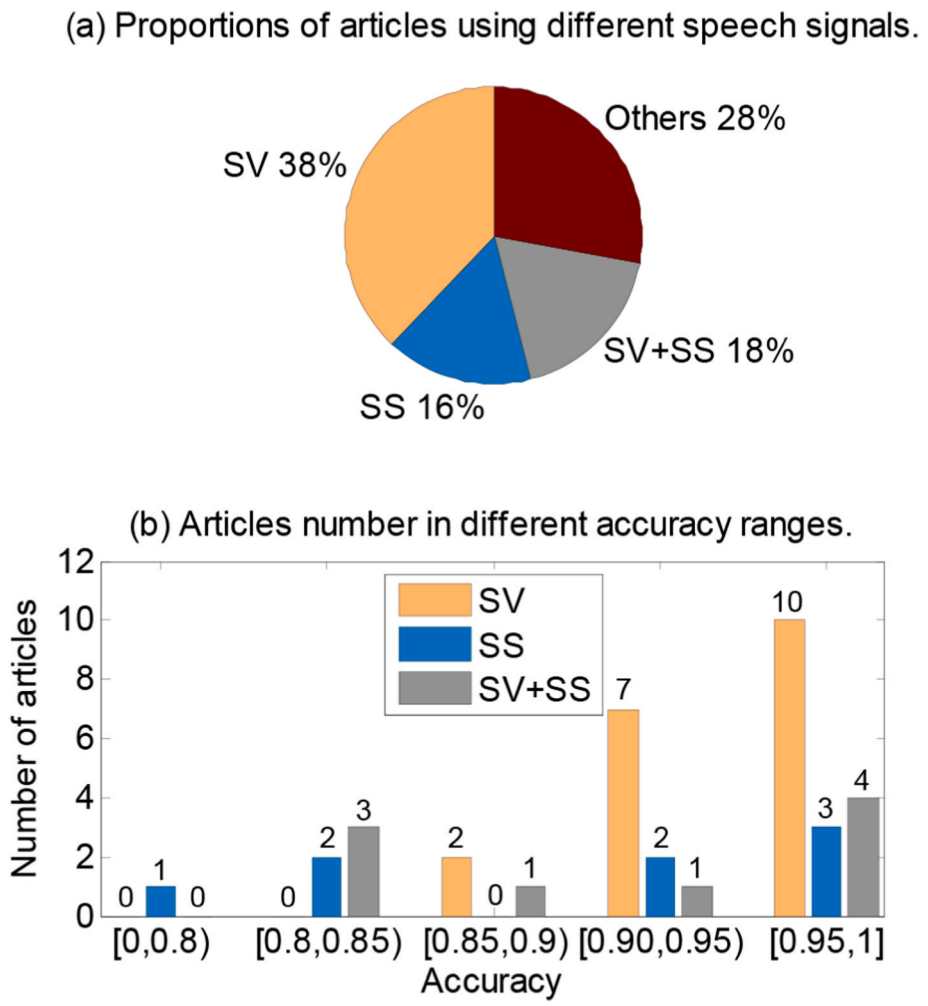

Fig. 15. Statistical analysis of articles using different voice signals of sustained vowels (SV), spontaneous speech (SS), SV + SS, and others.

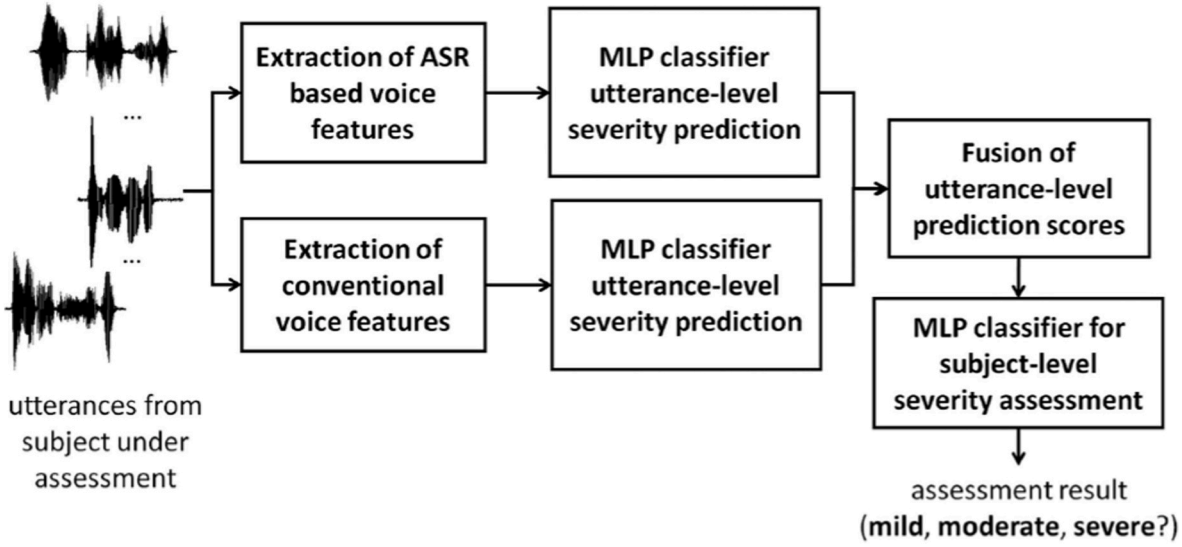

Fig. 16. Three-category voice disorder evaluation system based on Automatic speech recognition (ASR) [168].

error rate was almost always higher than that of the normal group.

5.3. State-of-the-art methods

In addition to the traditional identification methods, some new methods have also been designed in recent years. As shown in Fig. 17,

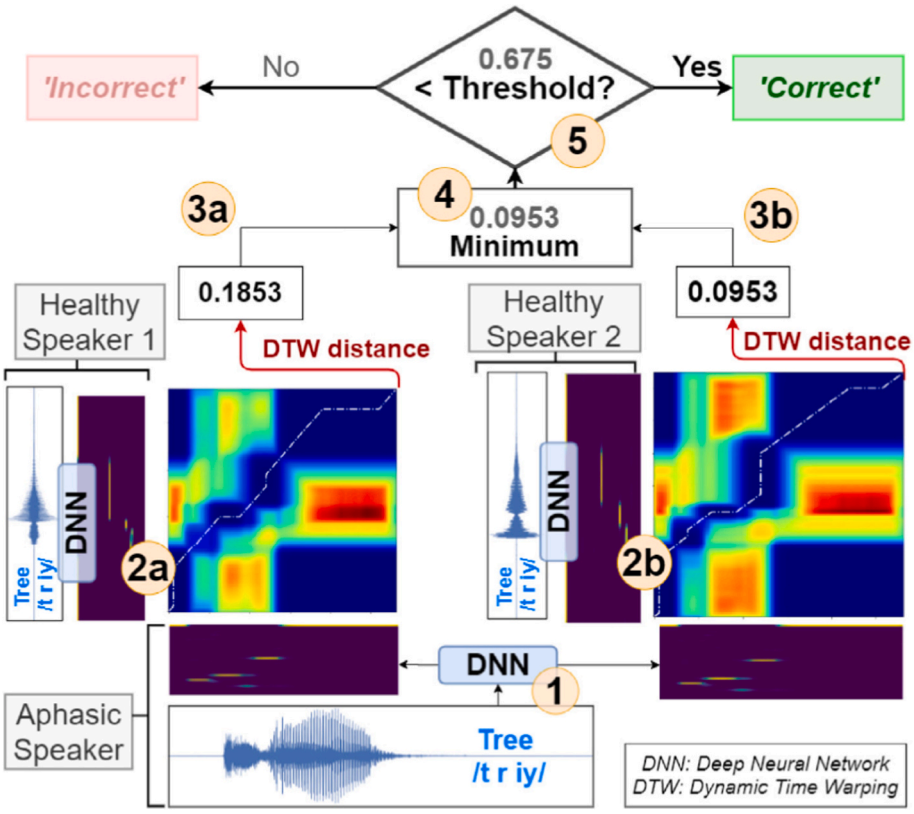

Fig. 17. NUVA: An utterance verification system for word naming of aphasic patients based on DNN and DTW [170].

Barbera et al. obtained the posterior probability of the patient's speech according to the acoustic model trained by a DNN network, compared it with the posterior probability of normal speech, and used the DTW algorithm to calculate the distance for classification [169,170]. The combination of DNN and vector matching method achieved a good result on the speech of word naming tests, which inspires us to integrate traditional methods with recent ones. For example, Lee et al. analyzed the distribution of frame-level posteriors produced by the DNN-HMM acoustic models [182]. They proposed an effective method for continuous speech utterances to extract dysphonia features from a specific set of discriminative phones with an ASR system.

Many studies also transformed the features of one-dimensional speech signals into two-dimensional features and used algorithms in the field of image recognition to investigate disease diagnosis. For example, Alhussein et al. converted pathological speech signals into spectrograms and then adopted CNNs for classification [105,106]. Qin et al. conducted a similar study, except that the input was a posterior probability map [181]. Muhammad et al. proposed to use the co-occurrence matrix feature combined with the GMM algorithm to classify pathological voices in the SVD database [102]. As shown in Fig. 18, in their recent study, Muhammad et al. utilized the LSTM algorithm to complete the recognition task [109]. They achieved an accuracy of 95% based on using CNN to fuse the spectrogram features of the voice and Electroglottograph (EGG) signals. Turning speech signal recognition into image recognition allows us to learn from the solutions in the field of image recognition to solve problems better. However, we also need to be careful in dealing with the problem of strict data alignment and the increase in computation.

As shown in Fig. 18, information fusion using multimodal data from different systems is also one of the main strategies used in speech-based

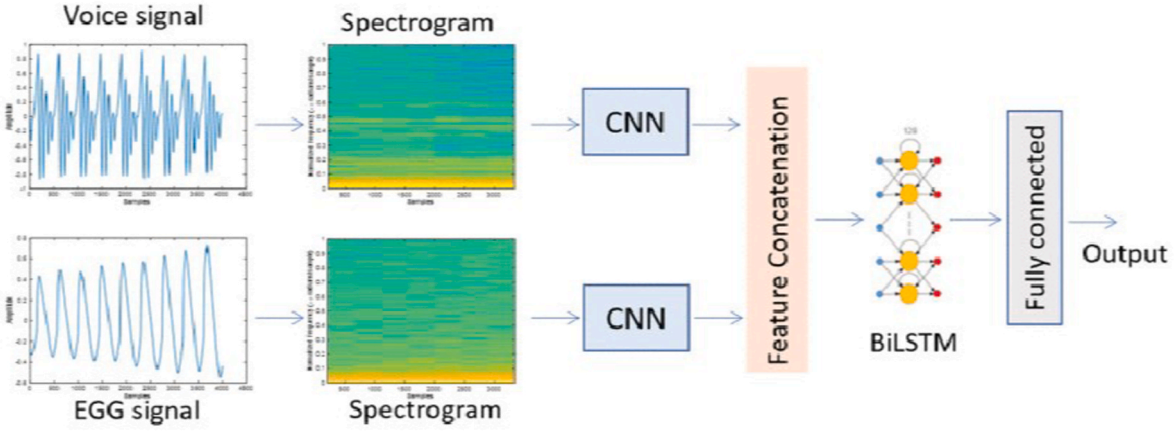

Fig. 18. A system architecture based on spectrograms [109].

disease diagnosis [102,109]. Fig. 19 shows the identification flow chart of the COVID-19 detection system [138] designed by Ponomarchuk et al. The patients' voice signals and symptom information are the system's input. First, the speech signal is processed by subsystem 1 based on Deep CNN and spectrogram and by subsystem 2 based on LightGBM and VGGish features to obtain the ensemble average class probabilities. Next, the symptom information is processed by Logistic Regression (LR) algorithm in subsystem 3 to obtain the class probabilities. Then, the final result of the weighted output probability is obtained based on the fusion of the results of the three subsystems. Botha et al. proposed a fused system combining the classifier based on objective clinical measurements and the classifier based on cough audio using LR, which improved sensitivity, specificity, and accuracy [148].

Similarly, Lauraitis et al. used information from three modalities of sound, finger tapping, and self-administered cognitive testing for symptom diagnosis [110]. The authors in Refs. [107,129,152,184] also conducted similar studies with multimodal data, and the recognition performances of their systems were higher than that with only one type of data. The COVID-19 detection system [131] designed by Stasak et al. used speech signals as the only input modality in their system. However,

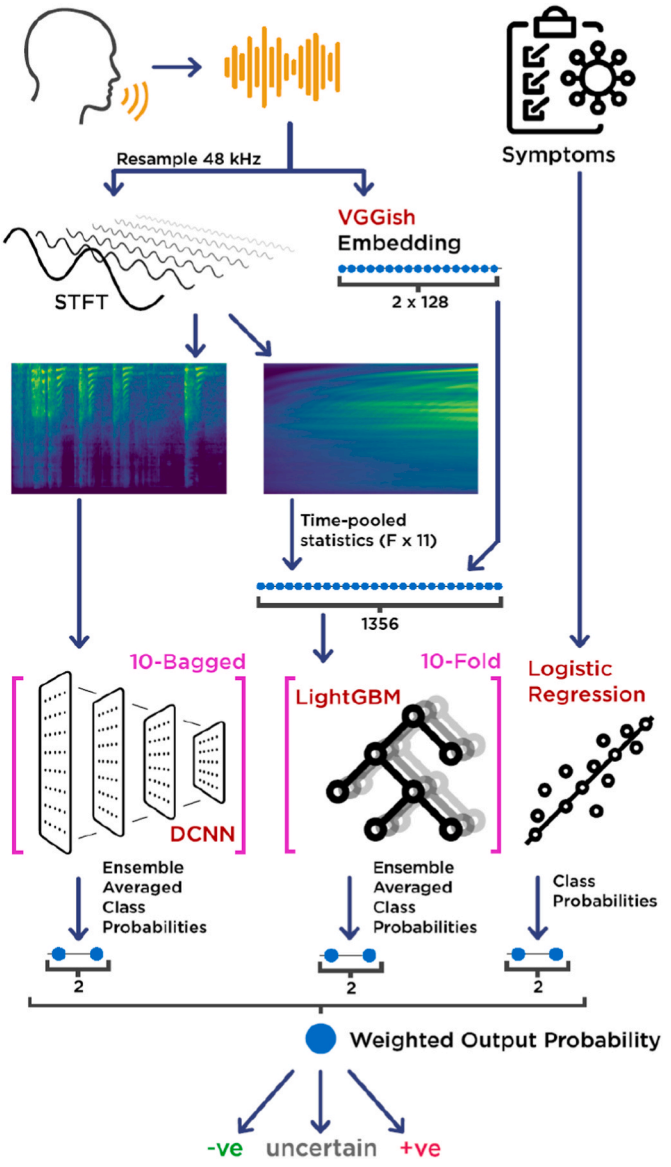

Fig. 19. Diagram of a COVID-19 detection system by fusion processing of speech and other signals [138].

the classification performance was effectively improved by adding a second-stage classifier to fuse the results of multiple first-stage classifiers. With the increase in data modalities, the amount of information, computation, and cost also increases, and the requirements for processing methods are also higher. For excellent performance, selecting several most effective modalities according to the experience of doctors may be a precondition for progress. As shown in Fig. 20, Chowdhury et al. also designed a complex ensemble-based system to detect COVID-19 [139]. The trained classifier layer is composed of 10 ML classifiers, which will be ranked by technique for order preference similarity to ideal solution and Entropy in Multi-Criteria Decision--Making blocks. At last, the features selected by Recursive Feature Elimination with Cross-Validation are fed into the best classifier. This method improves the diagnostic accuracy and adaptability of the whole system.

In addition to the above methods, some new attempts at pathological voice recognition exist.

Pahar et al. adopted speech, cough, and breath signals and rare bottleneck feature for pathological voice recognition [141]. In addition, they utilized transfer learning for training the model on the cough sounds of patients without COVID-19. Then the recognition system was tested with multiple pathological voice datasets and multiple classifiers, which verified the feasibility of this scheme [141]. Later, they adopted three DL classifiers, Restnet50, CNN, and LSTM, to classify TB, Covid-19, and health by cough. Finally, to make DL based-approaches achieve excellent performance and robustness, they adopted a synthetic minority over-sampling technique and transfer learning to address the issues of the class imbalance and insufficiency of their dataset, respectively [149].

Moreover, Harimoorthy et al. proposed an adaptive linear kernel SVM algorithm with higher prediction accuracy than traditional ML algorithms such as KNN, Random Forest (RF), Adaptive Weighted Probabilistic, and other k-SVMs [185]. Kambhampati et al. also proposed a fundamental heart sound segmentation algorithm based on sparse signal decomposition. They tested the algorithm's performance using various ML algorithms (hidden semi-Markov model, multilayer perceptron (MLP), SVM, and KNN) on real-time phonocardiogram (PCG) and PCG in a standard database. The results showed that their algorithm outperformed traditional heart sound segmentation algorithms [186].

Furthermore, Saeedi et al. used a genetic algorithm to find the filter bank parameters for feature extraction. They achieved an accuracy of 100% in classifying normal and pathological voices when the tests were performed on two databases [97,98]. Qian et al. [134] and Huang et al. [187] tried the popular end-to-end models in speech recognition and Transformer-based models for pathological speech signals processing. Their recognition results were very consistent with the evaluation scales of the patients. Fig. 21 is the framework diagram of pathological speech-based diagnosis designed by Wahengbam et al. [183]. First, a deep pathological denoiser (DPD) block is obtained by training the silence and noise features using CNN and has an inverse STFT operation to revert the spectrum of the voice signal to the time domain. The DPD block is the first step of the group decision analogy. Then, the three kinds of features of the denoised pathological speech obtained from the wavelet transform of Amor, Bump, and Morse are sent to three decision-making subsystems, respectively. Each subsystem uses multiple 3D convolutional network models for predictions. Finally, the fusion and decision-making are performed using the proposed group decision analogy strategy, and the accuracy was increased from 80.59% to 97.7% [183].

The studies mentioned above have made innovations in the procedures of speech technology and brought us many inspirations. Innovations in data include using different types of voice signals, integrating data of multiple modalities such as SVs, continuous speech, cough, breath, finger tapping, EGG, and disease symptoms, and utilizing transfer learning to train models to avoid the problem of insufficient data. For feature selection, these studies try genetic algorithms, DL

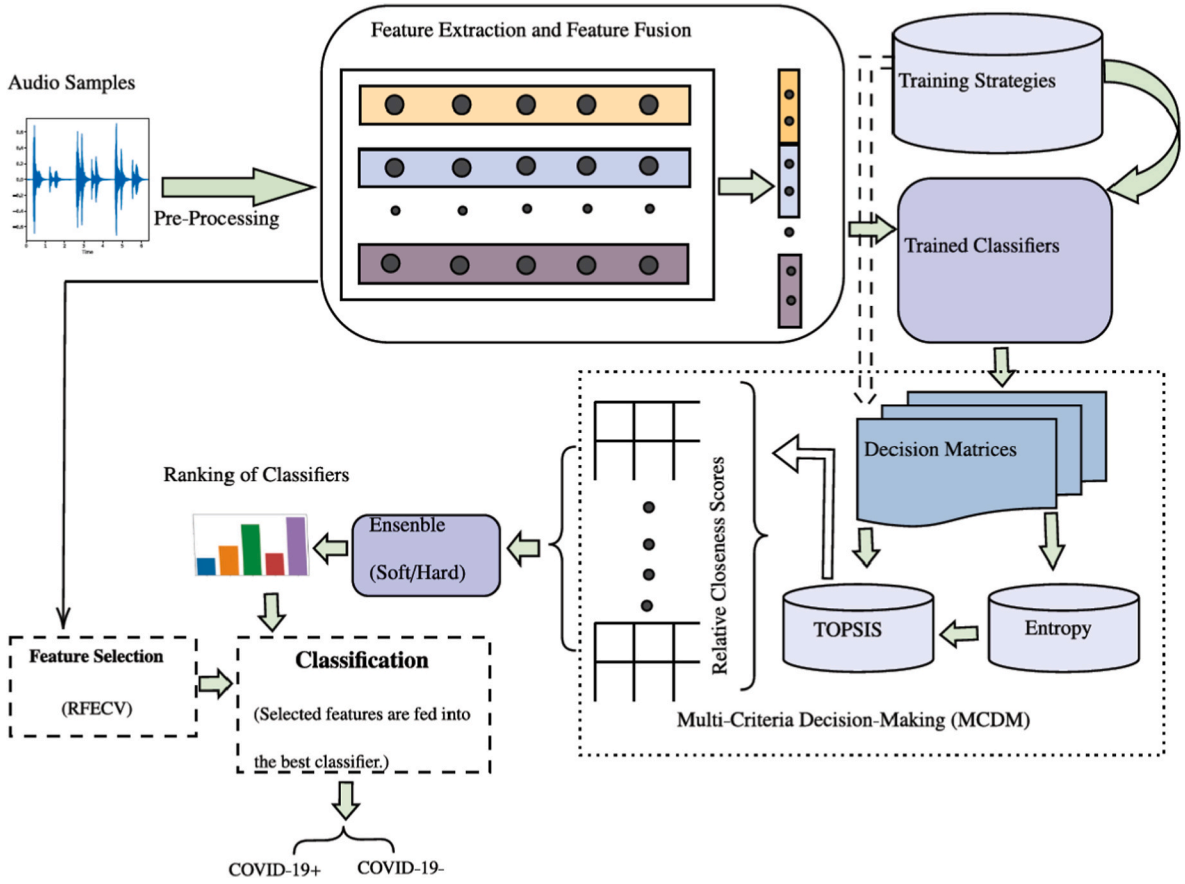

Fig. 20. An overview of an ensemble-based COVID-19 detection system [139].

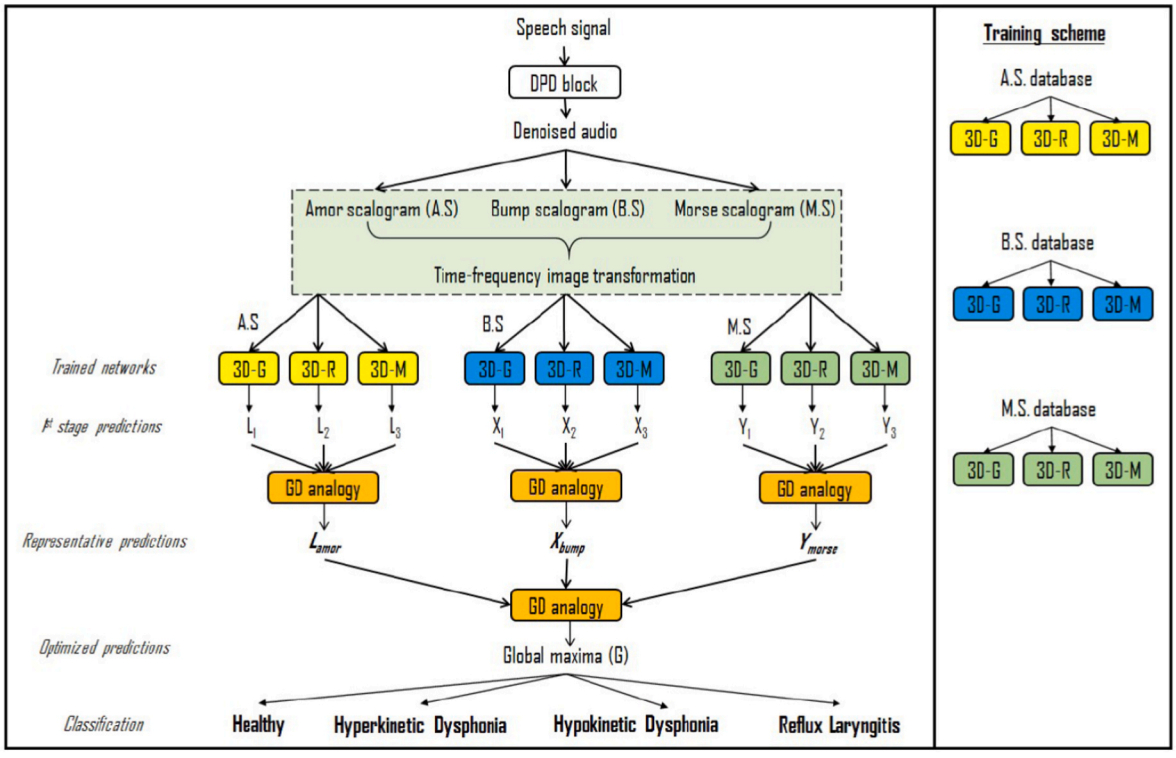

Fig. 21. Graphical workflow of group decision analogy showing the multiclass pathology identification framework [183].

algorithms, recursive feature reduction methods, fractal dimension approach, etc. In terms of classifiers, in addition to improving ML, these studies also try group decision strategy, end-to-end models, etc.

5.4. Summary and discussion

SV and SS are the two main types of speech signals used for pathological speech-based disease diagnosis and evaluation. In addition, cough is also considered indispensable in diagnosing respiratory diseases and is usually treated as an SV. The SV method mainly uses the abnormality of the patients' pronunciation as the basis for judgment, which is relatively less complicated in the process of experiment and application. However, the SV method ignores that the patient's speech differs from that of healthy people. On the other hand, SS utilizes the entire sentence as the judgment basis and can more accurately identify the obvious abnormal speech of the patient. However, the training of the algorithm model and the procedure of this method are relatively complicated.

In addition, Bhosale et al. [188] and Casanova et al. [189] also used cough sounds to diagnose COVID-19. Moreover, Gosztolya et al. utilized SS to distinguish schizophrenia from bipolar disorder [190]. No matter what kind of speech data is adopted, researchers try to find more effective speech signal features based on different speech signals for disease diagnosis and evaluation and choose a more suitable recognition algorithm according to the actual effect [15].

Using voice technology for disease diagnosis and assessment can effectively reduce the burden on doctors and improve the efficiency of medical resources. The traditional diagnosis methods rely on medical instruments combined with the doctors' experience. However, the application of speech technology only depends on the patient's speech and a pre-training algorithm model incorporating medical experience, which is more objective than the traditional methods [192]. In addition, the combination of pathological voice recognition technology with the IoT [103], telemedicine technology, and other technologies [193] allows patients to diagnose anytime and anywhere, reducing medical costs dramatically. We can also integrate pathological voice recognition functions into wearable devices to monitor patients' health during daily activities [194,195] to diagnose a disease early and prevent its deterioration [196].

In the future, to achieve better diagnosis and evaluation results, in addition to exploring more effective features and recognition algorithms, it is crucial to design multimodal data fusion methods [115,184,191,197,198] and build richer pathological voice datasets.

6. Speech recognition for human-medical equipment interaction

Doctors need to operate various equipment in their work. In addition, patients often require equipment to assist in treatment and rehabilitation. Integrating voice technology into medical equipment can bring great convenience to doctors and patients in many medical scenarios [199]. For example, smart medicine boxes remind patients to take medicines on time, intelligent ward round systems help doctors collect patient information [200], and voice systems perform automatic postoperative follow-up visits [201,202]. This section discusses related studies on medical device control using IST and how they can help doctors and patients in different scenarios. Finally, we discuss the requirements and future directions for the application of voice technology in smart medical equipment and devices.

6.1. Doctor and patient assist

6.1.1. Doctor assist

With the rapid development of medical speech technology, many studies have attempted to use it to assist doctors in operating equipment. For example, intelligent minimally invasive surgical systems have been put into clinical use, and doctors can control the robotic arm to perform precise operations through voice [208]. Ren et al. tried to embed speech recognition in the laparoscopic holder [209]. The holder with the speech command recognition function can replace the assistant and give corresponding feedback according to the instructions of the chief surgeon [209]. In addition, Tao et al. proposed an intelligent interactive operating room to solve the problem that the attending doctor must be in a sterile and non-contact environment and cannot view the lesion image in time during the operation [210]. The doctor can remotely control the display instrument using speech commands to locate and observe the image of the lesion quickly.

Furthermore, as shown in Fig. 22, Yoo et al. presented an intelligent voice assistant for the problem that the surgeon needs an assistant to check information during surgery continuously [211]. The voice assistant could recognize the proofreading speech of the attending doctor and compare it with the pre-input surgical information to ensure the smooth progress of the operation. Moreover, it also can remind the attending doctor of the length of the operation.

All these studies use IST to reduce the burden of inefficient labor on doctors and make the medical process more standardized and efficient.

6.1.2. Patient assist

In addition to using voice technology to assist doctors, many studies embed it in assistive devices for patients to help them have a better life quality. For example, intelligent wheelchairs integrated with voice technology are comprehensively studied. Li et al. designed a voice-controlled intelligent wheelchair that determines specific commands by comparing the appropriate distance of characteristic parameters [212]. As shown in Fig. 23(a), Atrash et al. added a computer, a display, a laser rangefinder, and an odometer to the wheelchair to realize an intelligent wheelchair that can navigate autonomously according to voice commands [203]. Al-Rousan et al. realized the movement direction control of an electric wheelchair using voice command recognition based on wavelets and neural networks [213]. Wang et al. developed an intelligent wheelchair that used a brain-computer interface and speech recognition for coordinated control for mentally ill patients with dysarthria [214].

Moreover, as shown in Fig. 23(b), Almutairi et al. proposed smart glasses that can navigate visually impaired patients to destinations based on Global Positioning System, Global System for Mobile communication, Google maps, and speech recognition [204]. The smart glasses designed by Punith et al. can also help a person with a visual disability to read printed notes, which works with Optical Character Recognition and Text to Speech technology [215].

Many studies also focus on using speech recognition technology as a communication method for deaf patients or patients with speech disorders. For instance, Jothi et al. proposed a knowledge-based system to analyze the unstructured words pronounced by the patient and transform them into meaningful text [216]. Balaji et al. attempted to help dysarthric persons overcome difficulties in interacting with others by mapping their distorted speech to normal or less severe dysarthric speech [217]. As shown in Fig. 23(c), Lee et al. designed an assistive agent system to help the hard of hearing person understand others. When the patient is talking to others, the assist device uses IST to recognize other people's speech as text and utilizes speech synthesis technology to convert the text into speech, helping the patient to communicate normally [205,218].

Furthermore, Fontan et al. experimentally found that using speech technology can improve the gain of hearing aids and maximize speech intelligibility and hearing comfort [206]. Akbarzadeh et al. employed reinforcement learning to personalize compression settings of hearing aids for patients to avoid loudness discomfort [219]. In addition, as shown in Fig. 23(d), LAPUL utilizes voice technology to make the preconditioning trainer easy to use [207]. All these studies use IST to help patients live and overcome the problems caused by diseases, which is conducive to the recovery of patients.

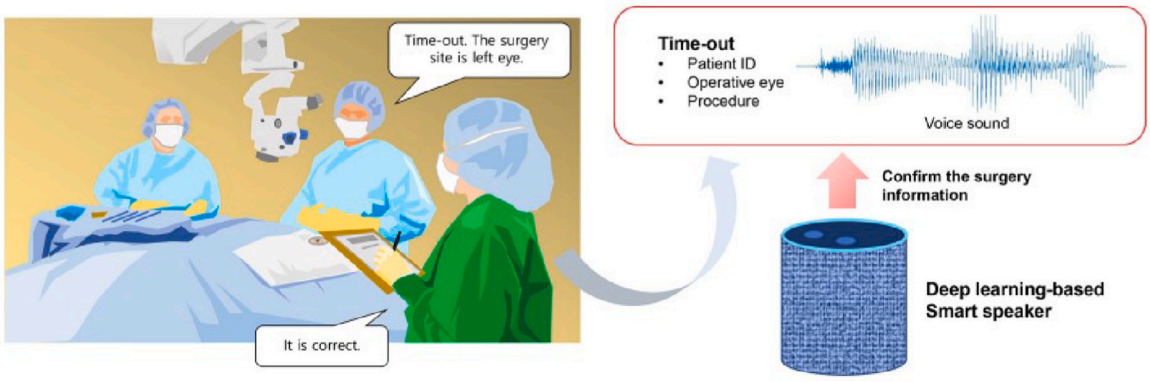

Fig. 22. A smart speaker to confirm surgical information in ophthalmic surgery [211].

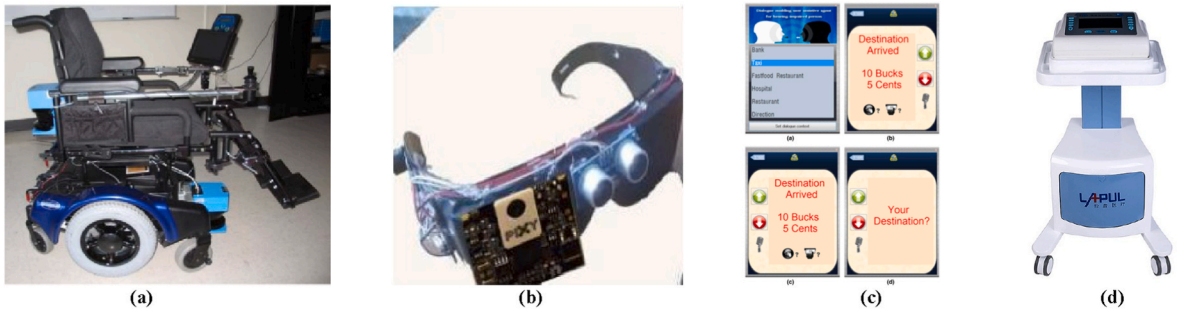

Fig. 23. Examples of smart devices integrated with voice technology for patients to live better. (a) Robotic wheelchair platform [203]. (b) Smart glasses prototype [204]. (c) A user interfaces for the hearing-impaired person [205]. (d) Preconditioning trainer [207].

6.2. Medical process optimization

In addition to assisting doctors and patients, IST will also bring changes to healthcare. Recently, COVID-19 has highlighted the importance of non-contact healthcare in the pandemic, and telemedicine is also an essential application of IST in healthcare [220]. For example, Amazon's Alexa Medical Dialogue systems have been widely used [221]. Nuance also has developed mature telemedicine solutions [222]. Liu et al. designed a low-cost cognitive tool to help children with autism exercise communication and color cognition skills [223]. As shown in Fig. 24, Bu et al. utilized virtual reality technology to design a system with functions of oral expression, auditory comprehension, cognition, and comprehensive application to help post-stroke aphasia patients to perform rehabilitation training [224]. They conducted a clinical trial using the system, and the subjects affirmed the rehabilitation training effect of the system's language skills. In addition, Jokić et al. proposed a contact-free cougher recognition approach using smartphone audio recordings and metric learning [153]. Pahar et al. also designed cough spotting and cougher identification methods for long-term personalized cough monitoring [225]. They also proposed an automatic non-invasive cough detection method based on audio and acceleration signals of a smartphone [225,226]. These non-contact cougher identification methods are helpful during the COVID-19 pandemic and promote the development of IST-based healthcare-monitoring technology.

Patient care is also a vital application scenario [231,232]. For example, Olami developed a smart speech-based hospital bed card [227]. Doctors can use the card to enter and manage patient information, allowing them to read it conveniently. Patients can also utilize the card to communicate with nursing stations easily. There are also studies using robots for patient care. With the help of speech technology, nursing robots can meet patients' needs according to their instructions [233]. Zorarobotics' designed an intelligent healthcare robot to help the elderly fight against loneliness and cognitive decline [229]. As shown in Fig. 25(a), Zhang et al. designed the Pepper rehabilitation medical robot for patients with cognitive and motor function decline disorders. This robot can interact with patients to help them practice language skills and remind them to take medicine [230]. Some studies have been applied in traditional medical scenarios, such as guiding robots for patient admission consultation [234]. Fig. 25(b) shows the guidance robot from Shen Zhou Yun Hai [228], which can provide consultation and guidance services for patients seeking medical treatment and reduce the burden of the consultation desk.

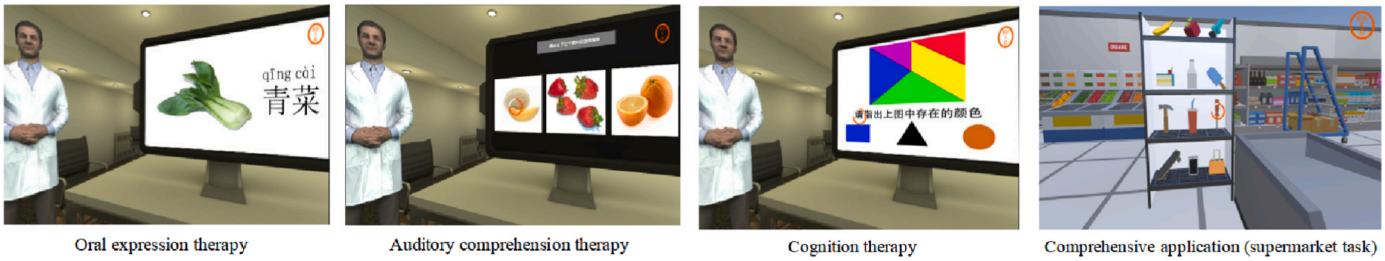

Fig. 24. Virtual reality and voice technologies for rehabilitation training [224].

Fig. 25. Examples of voice technology in intelligent healthcare. (a) Cognitive-motion rehabilitation medical robot [230]. (b) Guidance robot [228].

Optimizing the medical process makes the medical system more intelligent and efficient. People can entrust the work requiring less knowledge of medical experts to intelligent devices and improve the medical process through IST.

6.3. Summary and discussion

Embedding speech recognition, interaction, and synthesis into the devices can make them smarter. However, due to the particularity of the medical scenario, the research on speech recognition of intelligent equipment needs to pay attention to some issues.

a) A low misrecognition rate should be regarded as a prerequisite of the speech recognition system because misrecognition in medical scenarios costs high.
b) The methods for rejection recognition of a non-device user's voice also must be taken to ensure a low misrecognition rate.
c) The missed recognition rate will seriously affect the user experience.

Technically, speaker-dependent recognition can be used to ensure a low misrecognition rate. Speaker identification techniques [225] should be utilized to monitor the speaker who controls the equipment through voice commands in a noisy environment. In addition, the IST utilized for assisted control of medical equipment should ensure the highest priority of human intervention and the safety of medical operations.

In the future, smart medical devices can upload patient voices to the hospital's private cloud through IoT technology. The voice recognition model uses cloud computing technology to overcome the influence of different scenarios and language styles and improve speech recognition performance. In addition, Extended Reality technology integrated with voice technology will make telemedicine more realistic, effective, and acceptable [235].

7. Case study of intelligent speech technologies for stroke

Based on the above reviews of IST applied in the three medical scenarios, we conducted a case study for stroke recognition and rehabilitation assistance. In addition, we propose an IST application framework for stroke patients. In addition, we performed speech data collection and recognition experiments.

7.1. Speech technology for stroke patients

The medical system will become more intelligent with the development of IST, which enables the smart hospital to improve the efficiency of disease diagnosis, evaluation, surgery, and rehabilitation training. As depicted in Fig. 26, we take the medical process of a stroke patient treatment as an example to introduce the application of IST. IST combined with 5G communication technology connects hospitals and patients. Furthermore, early symptoms of stroke patients can be recognized by wearable devices integrated with speech recognition technology, such as smart wristbands, smartphones, smart glasses, and home smart monitoring devices, which are. These smart devices can give an early warning and make an emergency call after recognizing the symptoms.

In addition, there will be speech-based medical transcription systems in ambulances and the emergency department, which can help record the entire treatment process and complete documentation of the patient's information. Moreover, in the operating room, medical equipment understanding doctors' voice commands can help them to view patient lesions, proofread surgical information, record, and remind in real-time, which improves the standardization and efficiency of the surgical process. Furthermore, rehabilitation training is a vital treatment scenario for stroke patients. The patients can control smart wheelchairs, rehabilitation robots, and other equipment by using voice commands to help themselves in rehabilitation training and daily activity assistance at hospitals or homes, improving their quality of life. At the same time, voice technology is utilized to quickly evaluate the rehabilitation effect and record it in the patient's EMR system to help doctors adjust rehabilitation training strategies.

7.2. Data collection and speech recognition experiment

In this work, we conducted a pathological speech recognition experiment on stroke patients. A data collection system was developed for the pathological voice collection of stroke patients. The experimental protocol was approved by the Medical Research Ethics Committee of Guangdong Provincial People's Hospital (approval number: KY-Z-2021-431-02). Stroke patients and healthy people read the sentence "People's Republic of China" in the data collection experiments. The recorded audio was stored as.wav files. The audio signal had a sampling frequency of 16 kHz and a sampling accuracy of 16 bits. The hardware adopted in the data collection system was a laptop with the Ubuntu 20.04 operating system and a Hikivision microphone (Portable Speaker Phone (DS.

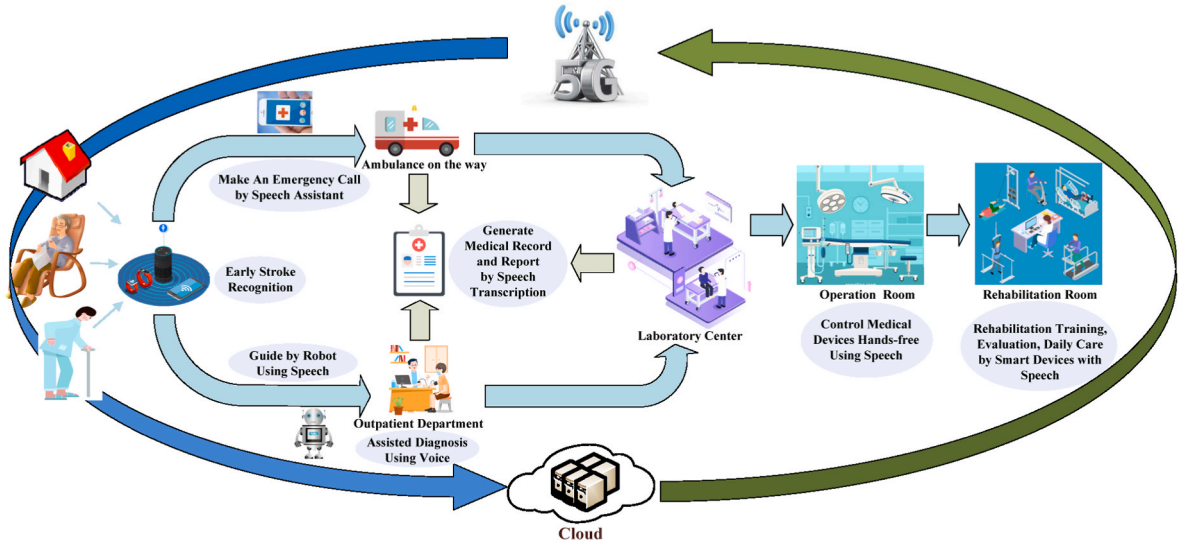

Fig. 26. Scenario of IST for early stroke recognition, rehabilitation assistance, and intelligent assessment.

65VA300B). Table 5 shows a dataset including 101 pathological sounds of stroke patients (male: 73, female: 27) and 101 healthy sounds of healthy people (male: 49, female: 52) used to validate stroke speech recognition.

The stroke speech recognition experiment was conducted under Anaconda. Hardware platform information: the CPU was Intel (R) i7-9700, and the Memory was 16 GB. Software environment: Python (3.7.13), librosa (0.9.0), sklearn (0.24.2). 22-dimensional features consisting of energy, perturbation, rhythms, 18-dim MFCC, and fundamental frequencies were extracted by librosa.

Finally, the features were utilized to perform five-fold Cross Validation using RF, KNN, linear kernel SVM, and Naïve Bayes. Table 6 shows the results. TP, FN, TN, FP, ACC, and F1 are the average of five rounds. RF algorithm achieves the best F1 and recognition accuracy of 0.87. Correspondingly, the prediction time is the longest, about 55.4 ms. Nevertheless, this time is acceptable for the actual applications. The results verify the feasibility of using speech technology to recognize the voice of stroke patients.

7.3. Voice assistants for stroke diagnosis

Voice assistants, such as Applications running on smartphones based on voice technology, can be used at home, in hospitals, and in community clinics. The voice assistants are helpful for the early recognition of stroke patients at home. Patients can diagnose by themselves or be assisted by family members. In addition, for the limited inspection equipment in community clinics, voice assistants can use the recognition model trained by the diagnostic data of large hospitals to better assist doctors in stroke diagnosis, which realizes the sharing of medical resources. Moreover, advanced voice assistants are also valuable in large hospitals. For example, with the development of AI technologies and the enrichment of speech datasets, voice assistants with more intelligent speech recognition algorithms can better adapt to complex environments and convert data accumulation into the accumulation of diagnostic experience. These voice assistants could achieve higher recognition accuracy than doctors in some cases and assist them in achieving the goal of smart hospitals.

Table 5
Subjects participated in the speech data collection experiments when they read "People's Republic of China.".

Subject	Male	Female	Total
Stroke Patients	74	27	101
Healthy People	49	52	101

Table 6
Speech recognition results in healthy people and stroke patients pronouncing "People's Republic of China.".

Recognition algorithm	TP	FN	TN	FP	ACC	F1	Time (ms)
RF	17.4	2.6	17.6	2.8	0.87	0.87	55.4
KNN	19.6	6.2	14	0.6	0.83	0.85	10.6
SVM (Linear Kernel)	15.2	6.2	14	5	0.72	0.73	1.5
Naïve Bayes	14.2	5.2	15	6	0.72	0.72	<1

8. Limitations and future directions

8.1. Limitations

Nowadays, speech technology is essential to traditional healthcare methods and systems. However, the application of speech technology in the medical system faces more challenges and needs to be continuously improved in the future. Therefore, we summarize some common issues in this section as potential future research directions.

8.1.1. Low adaptability and robustness

Most research on medical solutions based on voice technology is still under ideal conditions. However, the actual medical scenario is more complicated, and there will be more background noise, such as the sounds of the doctor's conversation and equipment beeping. Moreover, pronunciation differences from different doctors and the mixture of identifying results from multiple speakers are all potential factors that will degrade the performance of the speech recognition system. Medical application scenarios have high requirements for the adaptability and stability of the system, which is one considerable challenge to speech-based medical solutions.

8.1.2. Lacking high-quality pathological speech datasets

Although IST has excellent performances in some medical scenarios, pathological speech research lacks high-quality data for disease diagnosis and assessment of patients. Speech technologies are not transferable between different languages, thus slowing the study of pathological speech. Moreover, pathological speech datasets are even rarer due to patient discomfort and difficulty in speech data collection. The existing open-source pathological speech datasets are limited, small, and

collected by different criteria. These limitations make it challenging to use DL methods to obtain high-accuracy recognition models.

8.1.3. Medical voice data privacy protection

In the digital age, IST brings ethical and legal issues while playing an essential role in medical scenarios. The speech data may be stolen or used to extract personal information, such as users' identities, emotions, and other privacies. With the development of digital medical solutions, this problem will become prominent and challenge the popularization of speech-based medical solutions. Joint efforts from both legal and technical aspects should be made to protect the private speech data of patients and doctors. New privacy protection policies and techniques should be investigated comprehensively for the collection, transmission, storage, sharing, and use of medical speech data.

8.1.4. Conflict with traditional solutions

Applying IST in medical scenarios will inevitably change some traditional medical solutions. The resulting resistance of medical staff and patient acceptance are all issues that need to be considered. These problems also need to be resolved to promote the application of IST in smart hospitals.

8.1.5. Lacking unified and effective evaluation methods

The application of IST in medical scenarios must introduce performance indicators concerned in medical scenarios rather than evaluating the application effect only from the speech technical indicators. In the three scenarios described in this paper, medical transcription should focus on improving efficiency. Pathological speech recognition should emphasize the correct diagnosis rate and reduce misdiagnosis to diagnose patients' diseases safely. In medical equipment control, it is necessary to pay attention to the recognition effect in the actual scenario due to its complexity. While emphasizing the recognition rate, the recognition time, rejection rate, and rejection-recognition rate also need to be considered. Therefore, the speech-based medical solution should be evaluated in terms of its effectiveness in real-world medical scenarios, focusing on the actual experiences of doctors and patients.

8.2. Future directions

Speech technology has great potential in the field of healthcare. With the progress of research and clinical application testing, traditional healthcare systems will be reshaped. For example, as shown in Fig. 27, with the help of IoT technology, smart terminal devices such as microphone arrays, wearable devices, and medical assistive equipment can be connected to provide users with intelligent speech services such as transcription, diagnosis, and interaction. Meanwhile, these smart devices upload users' speech data into a database for storage over the network and use cloud computing technology to organize and analyze the speech to improve recognition accuracy further.

As a result, doctors and patients can utilize smart devices for initial diagnosis and telemedicine to improve the accessibility and quality of healthcare [236]. Intelligent sensors, devices, and robots can assist in monitoring patients and rehabilitation. Doctors can view patients' EMR and historical data collected by the sensors to analyze their health status systematically with speech-based auxiliary diagnosis systems. The intelligence of medical devices ensures the efficiency and standardization of surgical procedures. Researchers can use more comprehensive medical speech data to continuously optimize the recognition methods and improve the performance of the medical speech recognition systems. We can make efforts in the following four aspects to promote the development of smart hospitals and smart medicine to better serve patients, doctors, and researchers.

8.2.1. Improve the adaptability of IST

As illustrated in Fig. 27, improving the adaptability of speech processing-based smart equipment in different medical scenarios is a top priority because most current medical speech recognition systems are vulnerable to noise interference in practical application scenarios [19,20,183]. We can use a multi-channel microphone array and noise reduction algorithm to improve the quality of collected sound. Therefore, speech noise reduction algorithms and sound source identification algorithms related to microphone arrays [20] are popular research directions. Moreover, designing lightweight deep learning models [237] and selecting more critical features [147] will improve the real-time performance, adaptability, and robustness of the voice system in some medical scenarios, such as wearable devices-based patient monitoring. Furthermore, the ability to automatically switch personalized identification models based on the beeping of medical devices in different scenarios or patient-doctor interaction information is also worth exploring.

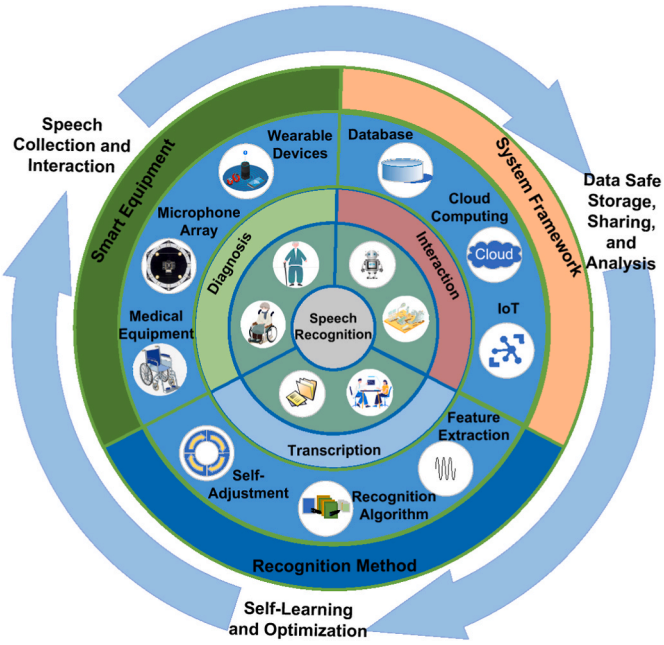

Fig. 27. Future intelligent healthcare speech technology in smart hospitals with smart equipment for speech data collection and human-computer interaction, novel system frameworks for medical voice data safe storage, sharing, and analysis, and intelligent recognition methods with self-learning and optimization capabilities.

8.2.2. Explore the potential of multimodal signals and DL methods

DL accounts for only a small part of the current pathological voice recognition methods because a large amount of data is unavailable [92,93,101,104,110,119–124,129,145,146]. However, researchers can continuously explore the potential of DL methods such as Attention-based LSTM [45,47], end-to-end models [134], and Transformer models [187] and try advanced recognition algorithms to improve the performance of IST for medical applications. Moreover, the fusion of voice signals with signals of other modalities such as electroacoustic gate signals, EMR, X-ray images, and ultrasound [4,5] will be more valuable for disease diagnosis in smart hospitals. For example, combining the chest X-ray images and cough sounds-based COVID-19 non-contact classification methods will minimize severity and mortality rates during the pandemic [5,6,188,189]. Furthermore, algorithms in other domains can be used in speech signal processing, such as AlexNet, VGGNet, GoogLeNet, and ResNet in image recognition can be adopted in the spectrum of speech signals. In addition, it is worth studying how to extract more effective speech signal features using DL with the help of medical experts and use few-shot learning methods to deal with the problem of limited pathological speech data.

8.2.3. Integrate with emerging technologies

As illustrated in Fig. 27, state-of-the-art information technologies such as IoT, 5G communication, cloud computing, virtual reality, and blockchain can facilitate voice-based medical solutions in smart hospitals and healthcare. IoT can provide solutions for distributed data collection and real-time monitoring. Audio coding, 5G, and cloud computing can reduce data transmission latency and computation delay, deal with big-data issues, and drive telemedicine services forward [10]. Virtual reality technology can make voice-assisted diagnosis and rehabilitation systems more fun and increase patient engagement. In addition, blockchain technology can protect users' privacy, facilitate the sharing of medical voice data, and promote the creation of open-source and high-quality voice datasets. Moreover, the voice analysis system should have self-learning and automatic optimization capabilities to obtain more intelligent and accurate recognition performance.

As shown in Fig. 28, we propose a novel medical voice analysis system architecture based on active perception. With active hardware, active software, and human-computer interaction, this framework realizes the active data collection and recognition of medical speech, as well as the closed-loop optimization of the recognition model to improve the intelligence of the medical system. Furthermore, the framework integrates knowledge reasoning and self-learning into speech-based systems, promoting the evolution of more powerful voice assistants.

8.2.4. Think from the points of view of doctors and patients

When applying speech technology to healthcare, it is necessary to clarify the auxiliary role of speech technology because there are uncertainties in speech recognition, such as noise interference, language differences, pronunciation differences, etc. We also need to think about problems from the perspective of patients and doctors. For example, a patient with multiple diseases should be treated by doctors first. At the same time, IST combined with image processing [6] and other technologies can assist doctors in their work to address the uncertainties and improve the robustness and acceptability of AI-based medical systems. In most cases, with techniques such as semantic reasoning and knowledge mapping, machines can better understand the intentions of the speaking of doctors and patients. Meanwhile, a friendlier and ergonomically excellent voice assistant system can be designed according to the specific application scenario to reduce the difficulty of using speech technology. In the face of resistance from traditional healthcare, speech technologies with customer-oriented and user-friendly interfaces and multimodal human-computer interaction capability can be more persuasive and welcomed.

Moreover, the methods for assessing the effectiveness of IST should be designed by integrating the objective results from transcription, disease diagnosis, and medical equipment control, as well as the questionnaire results of the subjective experience of doctors and patients. Furthermore, in addition to doctors, patients, and scientists, the government, medical industry, and legal departments need to work together to build standardized application procedures and assessment systems for IST used in smart hospitals to alleviate the shortage and imbalance of medical resources.

9. Conclusion

The COVID-19 pandemic has made us realize that the traditional medical system struggles to provide high-quality care due to a lack of staff and other medical resources. IST has brought unprecedented opportunities for health systems to address this issue. This paper first comprehensively reviews the application of IST in smart hospitals, including electronic medical document transcription, pathological voice recognition, and medical process optimization through human-medical equipment interaction. Then, we discuss how a speech-based healthcare system facilitates the early recognition, rehabilitation assistance, and intelligent assessment of stroke patients and introduce the diagnosis results of 101 stroke patients using their pathological speech data. The literature review shows that the study of IST in medical scenarios has attracted more and more scholars' attention and achieved promising results. State-of-the-art AI models, such as models based on Attention or Transformer, are applied to speech recognition. Moreover, the multimodal fusion of speech and other signals improves recognition accuracy and system robustness. However, these results are mainly from pilot projects or small datasets. Therefore, adequate research and validations are needed before clinical applications.

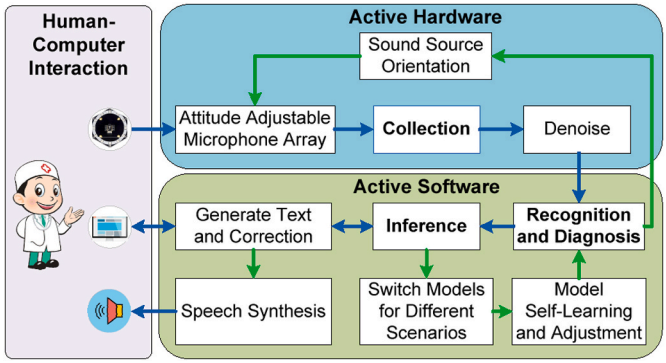

Fig. 28. Architecture diagram of a novel medical speech analysis system based on active perception. The system's intelligence is improved from three aspects, i.e., active hardware, active software, and human-computer interaction. Active hardware refers to the acquisition of voice information using a microphone array with active attitude adjustment ability. Active software refers to the application of reasoning and learning technology to voice recognition to continuously improve its accuracy. Finally, human-computer interaction refers to doctors and patients interacting with the voice system through voice, text, keyboard, mouse, wearable devices, etc., to realize a personalized intelligent voice system and improve the intelligence of the system.

Furthermore, we discuss some limitations to the development of IST in the medical field, such as the scarcity of available high-quality datasets, privacy issues, and lack of unified and effective evaluation methods. Finally, we present some future directions for medical speech technology. We also propose a novel active perception concept-based medical voice analysis system architecture, which employs active hardware, active software, and human-computer interaction to realize an intelligent and evolvable speech recognition system for smart hospitals.

The comprehensive review of the applications in smart hospitals provides helpful information for researchers on this topic. In addition, the summarized limitations and proposed future directions could give inspiration for future studies. Moreover, the case study of IST for stroke gives a reference for a full-process application of IST in various medical behaviors. Furthermore, the proposed active perception concept and the speech analysis system architecture can advance the IST applications in smart hospitals and offer an opportunity to apply IST in other scenarios with noise interference, such as airports, railway stations, and shopping malls.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2020AAA0109603, the National Natural Science Foundation of China under Grants 61873066 and 62173090, and the Zhi Shan Scholars Program of Southeast University under Grant 2242020R40096. We would like to thank the Reviewers for taking the

time and effort necessary to review the manuscript. We sincerely appreciate all valuable comments and suggestions that help us improve the quality of the manuscript.

References

[1] World Health Organization, 10 Facts on Ageing and Health, 2017 [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/10-facts-on-ageing-and-health.

[2] World Health Organization, Decade of Healthy Ageing: Baseline Report, 2020 [Online] Available: https://www.who.int/publications/i/item/9789240017900.

[3] H.I. Turkmen, M.E. Karsligil, Advanced computing solutions for analysis of laryngeal disorders, Med. Biol. Eng. Comput. 57 (2019) 2535–2552.

[4] Y.H. Bhosale, K.S. Patnaik, Application of deep learning techniques in diagnosis of Covid-19 (coronavirus): a systematic review, Neural Process. Lett. (2022) 1–53.

[5] Y.H. Bhosale, S. Zanwar, Z. Ahmed, M. Nakrani, D. Bhuyar, U. Shinde, Deep convolutional neural network based Covid-19 classification from radiology X-Ray images for IoT enabled devices, Int. Conf. Adv. Comput. Commun. Syst. (2022) 1398–1402.

[6] Y.H. Bhosale, K.S. Patnaik, PulDi-Covid, Chronic obstructive pulmonary (lung) diseases with COVID-19 classification using ensemble deep convolutional neural network from chest X-ray images to minimize severity and mortality rates, Biomed. Signal Proces. 81 (2023), 104445.

[7] M. Sajid, T. Shafique, M.J.A. Baig, I. Riaz, S. Amin, S. Manzoor, Automatic grading of palsy using asymmetrical facial features: a study complemented by new solutions, Symmetry 10 (2018) 242.

[8] Z. Guo, M. Shen, L. Duan, Y. Zhou, J. Xiang, H. Ding, S. Chen, O. Deussen, G. Dan, Deep assessment process: objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network, in: IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recogn., 2017, pp. 135–138.

[9] M.R. Kanhirakadavath, M.S.M. Chandran, Investigation of eye-tracking scan path as a biomarker for autism screening using machine learning algorithms, Diagnostics 12 (2022) 518.

[10] S. Latif, J. Qadir, A. Qayyum, M. Usama, S. Younis, Speech technology for healthcare: opportunities, challenges, and state of the art, IEEE Rev. Biomed. Eng. 14 (2021) 342–356.

[11] C.C. Chiu, T.N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R.J. Weiss, K. Rao, E. Gonina, et al., State-of-the-art speech recognition with sequence-to-sequence models, in: IEEE Int Conf Acoust Speech Signal Process Proc, IEEE, 2018, pp. 4774–4778.

[12] C. Herff, T. Schultz, Automatic speech recognition from neural signals: a focused review, Front. Neurosci. 10 (2016) 429.

[13] T.G. Poder, J.-F. Fisette, V. Dery, Speech recognition for medical dictation: overview in Quebec and systematic review, J. Med. Syst. 42 (2018) 89.

[14] S.V. Blackley, J. Huynh, L. Wang, Z. Korach, L. Zhou, Speech recognition for clinical documentation from 1990 to 2018: a systematic review, J. Am. Med. Inf. Assoc. 26 (2019) 324–338.

[15] V.L. Mamatha, A systematic review of machine learning based automatic speech assessment system to evaluate speech impairment, Int. Conf. Intell. Sustain. Syst. (2020) 175–185.

[16] N. Jamal, S. Shanta, F. Mahmud, M.N.A.H. Sha'abani, Automatic speech recognition (ASR) based approach for speech therapy of aphasic patients: a Review, AIP Conf. Proc. 2017 (1883), 020028.

[17] M. Johnson, S. Lapkin, V. Long, P. Sanchez, H. Suominen, J. Basilakis, L. Dawson, A systematic review of speech recognition technology in health care, Bmc. Med. Inform. Decis. 14 (2014) 94.

[18] G. Muhammad, Automatic speech recognition using interlaced derivative pattern for cloud based healthcare system, Cluster Comput. 18 (2015) 795–802.

[19] A. Ishfaque, B. Kim, Fly Ormia Ochracea inspired MEMS directional microphone: a review, IEEE Sensor. J. 18 (2018) 1778–1789.

[20] A. Rahaman, B. Kim, Microscale devices for biomimetic sound source localization: a review, J. Microelectromech. Syst. 31 (2022) 9–18.

[21] A.M. Ahmad, S. Ismail, D.F. Samaon, Recurrent neural network with backpropagation through time for speech recognition, in: IEEE Int. Symp. Commun. Inf. Technol, . IEEE, 2004, pp. 98–102.

[22] A. Keerio, B.K. Mitra, P. Birch, R. Young, C. Chatwin, On preprocessing of speech signals, Int. J. Signal Process. 5 (2009) 216–222.

[23] M.A. Al-Alaoui, L. Al-Kanj, J. Azar, E. Yaacoub, Speech recognition using artificial neural networks and hidden Markov models, IEEE Multidiscip. Eng. Educ. Mag. 3 (2008) 77–86.

[24] A.M. Othman, M.H. Riadh, Speech recognition using scaly neural networks, Int. J. Electr. Comput. Eng. 2 (2008) 211–216.

[25] S. Petrik, C. Drexel, L. Fessler, J. Jancsary, A. Klein, G. Kubin, J. Matiasek, F. Pernkopf, H. Trost, Semantic and phonetic automatic reconstruction of medical dictations, Comput. Speech Lang 25 (2011) 363–385.

[26] S.A. Alim, N.K.A. Rashid, Some Commonly Used Speech Feature Extraction Algorithms, From Natural to Artificial Intelligence - Algorithms and Applications, IntechOpen London, UK, 2018.

[27] S. Chehrehsa, T.J. Moir, Speech enhancement using maximum A-posteriori and Gaussian mixture models for speech and noise periodogram estimation, Comput. Speech Lang 36 (2016) 58–71.

[28] E.P. Frigieri, P.H.S. Campos, A.P. Paiva, P.P. Balestrassi, J.R. Ferreira, C. A. Ynoguti, A mel-frequency cepstral coefficient-based approach for surface roughness diagnosis in hard turning using acoustic signals and Gaussian mixture models, Appl. Acoust. 113 (2016) 230–237.

[29] R.S.S. Kumari, S.S. Nidhyananthan, A. G, Fused Mel feature sets based text-independent speaker identification using Gaussian mixture model, Procedia Eng. 30 (2012) 319–326.

[30] R.M. Ghoniem, K. Shaalan, A novel Arabic text-independent speaker verification system based on fuzzy hidden markov model, Procedia Comput. Sci. 117 (2017) 274–286.

[31] I. Shahin, Novel third-order hidden Markov models for speaker identification in shouted talking environments, Eng. Appl. Artif. Intell. 35 (2014) 316–323.

[32] H. Zeinali, H. Sameti, L. Burget, J.H. Černocký, Text-dependent speaker verification based on i-vectors, neural networks and hidden markov models, Comput. Speech Lang 46 (2017) 53–71.

[33] J.D. Bryan, S.E. Levinson, Autoregressive hidden markov model and the speech signal, Procedia Comput. Sci. 61 (2015) 328–333.

[34] P.J. Papandrea, E.P. Frigieri, P.R. Maia, L.G. Oliveira, A.P. Paiva, Surface roughness diagnosis in hard turning using acoustic signals and support vector machine: a PCA-based approach, Appl. Acoust. 159 (2020), 107102.

[35] B.R. Das, S. Sahoo, C.S. Panda, S. Patnaik, Part of speech tagging in Odia using support vector machine, Procedia Comput. Sci. 48 (2015) 507–512.

[36] M. Matsumoto, J. Hori, Classification of silent speech using support vector machine and relevance vector machine, Appl. Soft Comput. 20 (2014) 95–102.

[37] S. Lahmiri, A. Shmuel, Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine, Biomed. Signal Proces. 49 (2019) 427–433.

[38] L. Badino, C. Canevari, L. Fadiga, G. Metta, Integrating articulatory data in deep neural network-based acoustic modeling, Comput. Speech Lang 36 (2016) 173–195.

[39] L.L. Chen, J.J. Chen, Deep neural network for automatic classification of pathological voice signals, J. Voice 36 (2022) 288, e15–288.e24.

[40] I. Hwang, H.M. Park, J.H. Chang, Ensemble of deep neural networks using acoustic environment classification for statistical model-based voice activity detection, Comput. Speech Lang 38 (2016) 1–12.

[41] I. Shahin, A.B. Nassif, N. Hindawi, Speaker identification in stressful talking environments based on convolutional neural network, Int. J. Speech Technol. 24 (2021) 1055–1066.

[42] D. Issa, M.F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomed. Signal Proces. 59 (2020), 101894.

[43] S. Farsiani, H. Izadkhah, S. Lotfi, An optimum end-to-end text-independent speaker identification system using convolutional neural network, Comput. Electr. Eng. 100 (2022), 107882.

[44] M. Hireš, M. Gazda, P. Drotár, N.D. Pah, M.A. Motin, D.K. Kumar, Convolutional neural network ensemble for Parkinson's disease detection from voice recordings, Comput. Biol. Med. 141 (2022), 105021.

[45] M. Fernández-Díaz, A. Gallardo-Antolín, An attention Long Short-Term Memory based system for automatic classification of speech intelligibility, Eng. Appl. Artif. Intell. 96 (2020), 103976.

[46] B. Lindemann, T. Müller, H. Vietz, Na Jazdi, M. Weyrich, A survey on long short-term memory networks for time series prediction, Procedia CIRP 99 (2021) 650–655.

[47] A. Gallardo-Antolín, J.M. Montero, On combining acoustic and modulation spectrograms in an attention LSTM-based system for speech intelligibility level classification, Neurocomputing 456 (2021) 49–60.

[48] Y. Cheng, H.C. Leung, Speaker verification using fundamental frequency, Int. Conf. Spok. Lang. Process. (1998) 1–4.

[49] F. Wu, X. Wang, Z. Ye, The speaker and content adaptation in radiology information system, Appl. Mech. Mater. 195–196 (2012) 859–863.

[50] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. 28 (1980) 357–366.

[51] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, J. Acoust. Soc. Am. 87 (4) (1990) 1738–1752.

[52] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, E. Keogh, Searching and mining trillions of time series subsequences under dynamic time warping, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM, 2012, pp. 262–270.

[53] D.A. Reynolds, Speaker identification and verification using Gaussian mixture speaker models, Speech Commun. 17 (1995) 91–108.

[54] R.L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, P, IEEE ASME Trans. Mechatron. 77 (1989) 257–286.

[55] K.R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, An introduction to kernel-based learning algorithms, IEEE Trans. Neural Network. 12 (2001) 181–201.

[56] L. Deng, G. Hinton, B. Kingsbury, New types of deep neural network learning for speech recognition and related applications: an overview, in: IEEE Int. Conf. Acoust. Speech Signal Process. Proc., IEEE, 2013, pp. 8599–8603.

[57] K. O'Shea, R. Nash, An Introduction to Convolutional Neural Networks, 2015 arXiv preprint arXiv:1511.08458.

[58] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, S. Khudanpur, Recurrent neural network based language model, Proc. Annu. Conf. Int. Speech. Commun. Assoc. (2010) 1045–1048.

[59] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.

[60] L. Dawson, M. Johnson, H. Suominen, J. Basilakis, P. Sanchez, D. Estival, L. Hanlen, A usability framework for speech recognition technologies in clinical handover: a pre-implementation study, J. Med. Syst. 38 (2014) 56.

[61] L. Sun, M. Li, Sports and health management using big data based on voice feature processing and internet of things, Sci. Program. 2021 (2021), 3271863.
[62] L.M. Debruijn, E. Verheijen, A. Hasman, F.L. Vannes, J. Arends, Speech interfacing for diagnosis reporting systems - an overview, Comput. Methods Progr. Biomed. 48 (1995) 151–156.
[63] U.K.M. Teichgraber, T. Ehrenstein, M. Lemke, H. Stobbe, N. Hosten, U. Keske, R. Felix, Automatic speech recognition for report generation in computed tomography, Rofo.-Fortschr. Rontg. 171 (1999) 396–399.
[64] S.K. Fager, Speech recognition as a practice tool for dysarthria, Semin. Speech Lang. 38 (2017) 220–228.
[65] J.A. Landau, K.H. Norwich, S.J. Evans, Automatic speech recognition-can it improve the man-machine interface in medical expert systems? Int. J. Bio-Inspired Comput. 24 (1989) 111–117.
[66] T. Giorgino, I. Azzini, C. Rognoni, S. Quaglini, M. Stefanelli, R. Gretter, D. Falavigna, Automated spoken dialogue system for hypertensive patient home management, Int. J. Med. Inf. 74 (2005) 159–167.
[67] J. Shagoury, Dr. multi-task: using speech to build up electronic medical records while caring for patients, in: A. Neustein (Ed.), Adv. In Speech Recogn.: Mob. Environ., Call Cent. And Clinics, Springer US, Boston, MA, 2010, pp. 247–273.
[68] M. Rozmus, Transcription makeover. Virginia's Rockingham Memorial Hospital improves its clinical documentation process by implementing advanced speech-recognition technologies, Health Manag. Technol. 31 (2010) 20–21.
[69] Nuance. Heading off the Physician Shortage: the Role Ambient Clinical Intelligence Must Play. [Online] Available: https://(whatsnext.nuance.com/healthcare/the-role-ai-must-play-in-heading-off-the-physician-shortage/).
[70] Digitale Patientendokumente, MediaInterface [Online] Available: https://www.mediainterface.de/wie_wir_sie_unterstuetzen/digitale_patientendokumente.
[71] Meishan Traditional Chinese Medicine Hospital launched Unisound intelligent medical voice system (in Chinese). [Online] Available: https://m.sohu.com/a/237842928_401562#read.
[72] iFLYTEK: Medical care, education, justice, government services... Ten anti-epidemic artifacts (in Chinese). [Online] Available: http://sh.people.com.cn/big5/n2/2020/0225/c396182-33828035.html.
[73] K.C. Yaa, P. Claude, W. Harrison, G. Edward, A. Shilo, L. Christoph, Electronic health record interactions through voice: a review, Appl. Clin. Inf. 9 (2018) 541–552.
[74] S. Peivandi, L. Ahmadian, J. Farokhzadian, Y. Jahani, Evaluation and comparison of errors on nursing notes created by online and offline speech recognition technology and handwritten: an interventional study, Bmc. Med. Inform. Decis. 22 (2022) 96.
[75] S.V. Blackley, V.D. Schubert, F.R. Goss, W. Al Assad, P.M. Garabedian, L. Zhou, Physician use of speech recognition versus typing in clinical documentation: a controlled observational study, Int. J. Med. Inf. 141 (2020), 104178.
[76] B. Heinzer, Essential elements of nursing notes and the transition to electronic health records: the migration from narrative charting will require creativity to include essential elements in EHRs, J. Healthc. Inf. Manag. 24 (2010) 53–59.
[77] J. Groschel, F. Philipp, S. Skonetzki, H. Genzwurker, T. Wetter, K. Ellinger, Automated speech recognition for time recording in out-of-hospital emergency medicine-an experimental approach, Resuscitation 60 (2004) 205–212.
[78] T. Hodgson, F. Magrabi, E. Coiera, Efficiency and safety of speech recognition for documentation in the electronic health record, J. Am. Med. Inf. Assoc. 24 (2017) 1127–1133.
[79] A. Femi-Abodunde, K. Olinger, L.M.B. Burke, T. Benefield, E.R. Lee, K. McGinty, B.M. Mervak, Radiology dictation errors with COVID-19 protective equipment: does wearing a surgical mask increase the dictation error rate? J. Digit. Imag. 34 (2021) 1294–1301.
[80] J. Gnanamanickam, Y. Natarajan, S.P.K. R, A hybrid speech enhancement algorithm for voice assistance application, Sensors 21 (2021) 7025.
[81] T. Duan, X.K. Xu, S.F. Chen, Q. Zhang, W.Y. Chen, X.M. Lu, X.L. He, Application of adaptive technology-based speech recognition system in 600 pathological grossing process, Chin. J. Pathol. 50 (2021) 1034–1038.
[82] iFLYTEK. Dental electronic medical record. [Online] Available: https://health.xfyun.cn/solutions/eHistory.
[83] K. Voll, S. Atkins, B. Forster, Improving the utility of speech recognition through error detection, J. Digit. Imag. 21 (2008) 371–377.
[84] J.G. Klann, P. Szolovits, An intelligent listening framework for capturing encounter notes from a doctor-patient dialog, BMC Med. Inf. Decis. Making 9 (2009) S3.
[85] B. Qin, Research on the application of intelligent speech recognition technology in medical big data fog computing system, J. Decis. Syst. (2021) 1–13.
[86] M. Singh, S. Bharti, H. Kaur, V. Arora, M. Saini, M. Kaur, J. Singh, A facial and vocal expression based comprehensive framework for real-time student stress monitoring in an IoT-Fog-Cloud environment, IEEE Access 10 (2022) 63177–63188.
[87] F.R. Goss, L. Zhou, S.G. Weiner, Incidence of speech recognition errors in the emergency department, Int. J. Med. Inf. 93 (2016) 70–73.
[88] K.P. Andriole, L.M. Prevedello, A. Dufault, P. Pezeshk, R. Bransfield, R. Hanson, P.M. Doubilet, S.E. Seltzer, R. Khorasani, Augmenting the impact of technology adoption with financial incentive to improve radiology report signature times, J. Am. Coll. Radiol. 7 (2010) 198–204.
[89] S.-H. Lee, J. Park, K. Yang, J. Min, J. Choi, Accuracy of cloud-based speech recognition open application programming interface for medical terms of Korean, J. Kor. Med. Sci. 37 (2022) e144.
[90] F.R. Goss, S.V. Blackley, C.A. Ortega, L.T. Kowalski, A.B. Landman, C.T. Lin, M. Meteer, S. Bakes, S.C. Gradwohl, D.W. Bates, et al., A clinician survey of using speech recognition for clinical documentation in the electronic health record, Int. J. Med. Inf. 130 (2019), 103938.
[91] J.A. Rodger, P.C. Pendharkar, A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application, Int. J. Hum. Comput. Stud. 60 (2004) 529–544.
[92] J. Wang, C. Jo, Vocal folds disorder detection using pattern recognition methods, in: Annu Int Conf IEEE Eng Med Biol Proc, IEEE, 2007, pp. 3253–3256.
[93] H.K. Heris, B.S. Aghazadeh, M. Nikkhah-Bahrami, Optimal feature selection for the assessment of vocal fold disorders, Comput. Biol. Med. 39 (2009) 860–868.
[94] Z. Ali, M.S. Hossain, G. Muhammad, A.K. Sangaiah, An intelligent healthcare system for detection and classification to discriminate vocal fold disorders, Future Generat. Comput. Syst. 85 (2018) 19–28.
[95] M. Madruga, Y. Campos-Roca, C.J. Perez, Multicondition training for noise-robust detection of benign vocal fold lesions from recorded speech, IEEE Access 9 (2020) 1707–1722.
[96] L. Verde, G. De Pietro, M. Alrashoud, A. Ghoneim, K.N. Al-Mutib, G. Sannino, Leveraging artificial intelligence to improve voice disorder identification through the use of a reliable mobile app, IEEE Access 7 (2019) 124048–124054.
[97] N.E. Saeedi, F. Almasganj, Wavelet adaptation for automatic voice disorders sorting, Comput. Biol. Med. 43 (2013) 699–704.
[98] N.E. Saeedi, F. Almasganj, F. Torabinejad, Support vector wavelet adaptation for pathological voice assessment, Comput. Biol. Med. 41 (2011) 822–828.
[99] E.S. Fonseca, R.C. Guido, P.R. Scalassara, C.D. Maciel, J.C. Pereira, Wavelet time-frequency analysis and least squares support vector machines for the identification of voice disorders, Comput. Biol. Med. 37 (2007) 571–578.
[100] O. Amir, W.T. Abraham, Z.S. Azzam, G. Berger, S.D. Anker, S.P. Pinney, D. Burkhoff, I.D. Shallom, C. Lotan, E.R. Edelman, Remote speech analysis in the evaluation of hospitalized patients with acute decompensated heart failure, JACC-Heart Fail. 10 (2022) 41–49.
[101] Z. Ali, G. Muhammad, M.F. Alhamid, An automatic health monitoring system for patients suffering from voice complications in smart cities, IEEE Access 5 (2017), 3900−3908.
[102] G. Muhammad, M.F. Alhamid, M.S. Hossain, A.S. Almogren, A.V. Vasilakos, Enhanced living by assessing voice pathology using a co-cccurrence matrix, Sensors 17 (2017) 267.
[103] G. Muhammad, S.K.M.M. Rahman, A. Alelaiwi, A. Alamri, Smart health solution integrating IoT and cloud: a case study of voice pathology monitoring, IEEE Commun. Mag. 55 (2017) 69–73.
[104] K. Lopez-de-Ipina, J. Sole-Casals, H. Eguiraun, J.B. Alonso, C.M. Travieso, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage, B. Beitia, Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: a fractal dimension approach, Comput. Speech Lang 30 (2015) 43–60.
[105] M. Alhussein, G. Muhammad, Voice pathology detection using deep learning on mobile healthcare framework, IEEE Access 6 (2018) 41034–41041.
[106] M. Alhussein, G. Muhammad, Automatic voice pathology monitoring using parallel deep models for smart healthcare, IEEE Access 7 (2019) 46474–46479.
[107] S.H. Fang, C.T. Wang, J.Y. Chen, Y. Tsao, F.C. Lin, Combining acoustic signals and medical records to improve pathological voice classification, APSIPA Trans. Signal Inf. Process. 8 (2019) e14.
[108] T. Tuncer, S. Dogan, F. Ozyurt, S.B. Belhaouari, H. Bensmail, Novel multi center and threshold ternary pattern based method for disease detection method using voice, IEEE Access 8 (2020) 84532–84540.
[109] G. Muhammad, M. Alhussein, Convergence of artificial intelligence and internet of things in smart healthcare: a case study of voice pathology detection, IEEE Access 9 (2021) 89198–89209.
[110] A. Lauraitis, R. Maskeliunas, R. Damasevicius, T. Krilavicius, A mobile application for smart computer-aided self-administered testing of cognition, speech, and motor impairment, Sensors 20 (2020) 3236.
[111] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, H. Christensen, Dementia detection using automatic analysis of conversations, Comput. Speech Lang 53 (2019) 65–79.
[112] J.T. Becker, F. Boiler, O.L. Lopez, J. Saxton, K.L. McGonigle, The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis, Arch. Neurol. 51 (1994) 585–594.
[113] T. Kishimoto, A. Takamiya, K.C. Liang, K. Funaki, M. Mimura, The project for objective measures using computational psychiatry technology (PROMPT): rationale, design, and methodology, Contemp. Clin. Trials 19 (2020), 100649.
[114] F. Eyben, M. Wöllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proc. ACM MM, ACM, 2010, pp. 1459–1462.
[115] M. Rodrigues Makiuchi, T. Warnita, N. Inoue, K. Shinoda, M. Yoshimura, M. Kitazawa, K. Funaki, Y. Eguchi, T. Kishimoto, Speech paralinguistic approach for detecting dementia using gated convolutional neural network, IEICE Trans. Info Syst. E104D (2021) 1930–1940.
[116] J. Orozco-Arroyave, J. Arias-Londoo, J. Vargas-Bonilla, M. González-Rátiva, E. Nth, New Spanish speech corpus database for the analysis of people suffering from Parkinsons disease, in: Int. Conf. Lang, Resourc. and Eval., 2014, pp. 342–347.
[117] P. Janbakhshi, I. Kodrasi, H. Bourlard, Subspace-based learning for automatic dysarthric speech detection, IEEE Signal Process. Lett. 28 (2021) 96–100.
[118] E. Svoboda, T. Bořil, J. Rusz, T. Tykalová, D. Horáková, C.R.G. Guttmann, K. B. Blagoev, H. Hatabu, V.I. Valtchinov, Assessing clinical utility of machine learning and artificial intelligence approaches to analyze speech recordings in multiple sclerosis: a pilot study, Comput. Biol. Med. 148 (2022), 105853.
[119] Z. Cai, J. Gu, H.-L. Chen, A new hybrid intelligent framework for predicting Parkinson's disease, IEEE Access 5 (2017) 17188–17200.

[120] E. Vaiciukynas, A. Verikas, A. Gelzinis, M. Bacauskiene, Detecting Parkinson's disease from sustained phonation and speech signals, PLoS One 12 (2017), e0185613.
[121] H. Kuresan, D. Samiappan, S. Masunda, Fusion of WPT and MFCC feature extraction in Parkinson's disease diagnosis, Technol. Health Care 27 (2019) 363–372.
[122] F. Karlsson, E. Schalling, K. Laakso, K. Johansson, L. Hartelius, Assessment of speech impairment in patients with Parkinson's disease from acoustic quantifications of oral diadochokinetic sequences, J. Acoust. Soc. Am. 147 (2020) 839–851.
[123] C.O. Sakar, G. Serbes, A. Gunduz, H.C. Tunc, H. Nizam, B.E. Sakar, M. Tutuncu, T. Aydin, M.E. Isenkul, H. Apaydin, A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform, Appl. Soft Comput. 74 (2019) 255–263.
[124] L. Naranjo, C. Pérez, Y. Campos-Roca, J. Martín, Addressing voice recording replications for Parkinson's disease detection, Expert Syst. Appl. 46 (2016) 286–292.
[125] M. Pramanik, R. Pradhan, P. Nandy, A.K. Bhoi, P. Barsocchi, Machine learning methods with decision forests for Parkinson's detection, Appl. Sci. 11 (2021) 581.
[126] C. Laganas, D. Iakovakis, S. Hadjidimitriou, V. Charisis, S.B. Dias, S. Bostantzopoulou, Z. Katsarou, L. Klingelhoefer, H. Reichmann, D. Trivedi, et al., Parkinson's disease detection based on running speech data from phone calls, IEEE Trans. Biomed. Eng. 69 (2022) 1573–1584.
[127] A. Rawat, S. Mishra, Y. Sharma, P. Khetarpal, High accuracy multilayer autoencoder trained classification method for diagnosis of Parkinson's disease using vocal signals, J. Inf. Optim. Sci. 43 (2022) 93–99.
[128] M. Hireš, M. Gazda, P. Drotár, N.D. Pah, M.A. Motin, D.K. Kumar, Convolutional neural network ensemble for Parkinson's disease detection from voice recordings, Comput. Biol. Med. 141 (2022), 105021.
[129] C. Shimon, G. Shafat, I. Dangoor, A. Ben-Shitrit, Artificial intelligence enabled preliminary diagnosis for COVID-19 from voice cues and questionnaires, J. Acoust. Soc. Am. 149 (2021) 1120–1124.
[130] Sonde Health. Sonda Health (SH). [Online] Available: https://www. sondehealth. com/sondeone-page.
[131] B. Stasak, Z.C. Huang, S. Razavi, D. Joachim, J. Epps, Automatic detection of COVID-19 based on short-duration acoustic smartphone speech analysis, J. Healthcare Inform. Res. 5 (2021) 201–217.
[132] L. Verde, G.D. Pietro, A. Ghoneim, M. Alrashoud, K.N. Al-Mutib, G. Sannino, Exploring the use of artificial intelligence techniques to detect the presence of coronavirus covid-19 through speech and voice analysis, IEEE Access 9 (2021), 65750–65757.
[133] M. Kamble, J. Gonzalez-Lopez, T. Grau, J. Espín López, L. Cascioli, Y.Q. Huang, A. Gomez-Alanis, J. Patino, R. Font, A. Peinado, et al., PANACEA cough sound-based diagnosis of COVID-19 for the DiCOVA 2021 Challenge, Proc. Annu. Conf. Int. Speech. Commun. Assoc. (2021) 4271–4275.
[134] K. Qian, M. Schmitt, H.Y. Zheng, T. Koike, B. Schuller, Computer audition for fighting the SARS-CoV-2 corona crisis-introducing the multitask speech corpus for COVID-19, IEEE Internet Things 8 (2021) 16035–16046.
[135] A.C. Villa-Parra, I. Criollo, C. Valadão, L. Silva, Y. Coelho, L. Lampier, L. Rangel, G. Sharma, D. Delisle-Rodríguez, J. Calle-Siguencia, Towards multimodal equipment to help in the diagnosis of COVID-19 using machine learning algorithms, Sensors 22 (2022) 4341.
[136] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, C. Mascolo, Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., ACM, 2021, pp. 3474–3484.
[137] M.J. Son, S.P. Lee, COVID-19 diagnosis from crowdsourced cough sound data, Appl. Sci. 12 (2022) 1795.
[138] A. Ponomarchuk, I. Burenko, E. Malkin, I. Nazarov, V. Kokh, M. Avetisian, L. Zhukov, Project achoo: a practical model and application for COVID-19 detection from recordings of breath, voice, and cough, IEEE J. Sel. Top. Signal. Process. 16 (2022) 175–187.
[139] N.K. Chowdhury, M.A. Kabir, M.M. Rahman, S.M.S. Islam, Machine learning for detecting COVID-19 from cough sounds: an ensemble-based MCDM method, Comput. Biol. Med. 145 (2022), 105405.
[140] M. Pahar, M. Klopper, R. Warren, T. Niesler, COVID-19 cough classification using machine learning and global smartphone recordings, Comput. Biol. Med. 135 (2021), 104572.
[141] M. Pahar, M. Klopper, R. Warren, T. Niesler, COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features, Comput. Biol. Med. 141 (2022), 105153.
[142] N.S. Haider, B.K. Singh, R. Periyasamy, A.K. Behera, Respiratory sound based classification of chronic obstructive pulmonary disease: a risk stratification approach in machine learning paradigm, J. Med. Syst. 43 (2019) 255.
[143] Z. Chang, P. Luo, B. Yang, X. Zhang, Respiratory sound recognition of chronic obstructive pulmonary disease patients based on HHT-MFCC and short-term energy, J. Comput. Appl. 41 (2021) 598–603.
[144] E. Yilmaz, V. Mitra, G. Sivaraman, H. Franco, Articulatory and bottleneck features for speaker-independent ASR of dysarthric speech, Comput. Speech Lang 58 (2019) 319–334.
[145] S.R. Mani Sekhar, G. Kashyap, A. Bhansali, A. Andrew, K. Singh, Dysarthric-speech detection using transfer learning with convolutional neural networks, ICT Express 8 (2021) 61–64.
[146] N.P. Narendra, P. Alku, Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features, Comput. Speech Lang 65 (2021), 101117.
[147] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, T. Niesler, Automatic cough classification for tuberculosis screening in a real-world environment, Physiol. Meas. 42 (2021), 105014.
[148] G.H.R. Botha, G. Theron, R.M. Warren, M. Klopper, K. Dheda, P.D. van Helden, T. R. Niesler, Detection of tuberculosis by automatic cough sound analysis, Physiol. Meas. 39 (2018), 045005.
[149] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, A. Diacon, T. Niesler, Automatic Tuberculosis and COVID-19 cough classification using deep learning, in: Int. Conf. Electr., Comput., Energy Technol., 2022, pp. 1–9.
[150] B. Semiz, S. Hersek, D.C. Whittingslow, L.A. Ponder, S. Prahalad, O.T. Inan, Using knee acoustical emissions for sensing joint health in patients with juvenile idiopathic arthritis: a pilot study, IEEE Sensor. J. 18 (2018) 9128–9136.
[151] G. Holmes, A. Donkin, I.H. Witten, WEKA: a machine learning workbench, in: Proc. Of ANZIIS, 1994, pp. 357–361.
[152] N. Klangpornkun, M. Ruangritchai, A. Munthuli, C. Onsuwan, K. Jaisin, K. Pattanaseri, J. Lortrakul, P. Thanakulakkarachai, T. Anansiripinyo, A. Amornlaksananon, Classification of depression and other psychiatric conditions using speech features extracted from a Thai psychiatric and verbal screening test, in: 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., IEEE, 2021, pp. 651–656.
[153] S. Jokić, D. Cleres, F. Rassouli, C. Steurer-Stey, M.A. Puhan, M. Brutsche, E. Fleisch, F. Barata, TripletCough: cougher identification and verification from contact-free smartphone-based audio recordings using metric learning, IEEE J. Bio.Med. Health 26 (2022) 2746–2757.
[154] S. Mobram, M. Vali, Depression detection based on linear and nonlinear speech features in I-vector/SVDA framework, Comput. Biol. Med. 149 (2022), 105926.
[155] H. Byeon, Comparing ensemble-based machine learning classifiers developed for distinguishing hypokinetic dysarthria from presbyphonia, Appl. Sci. 11 (2021) 2235.
[156] I.R. Titze, D.W. Martin, Principles of voice production, J. Acoust. Soc. Am. 104 (1998) 1148.
[157] B. Tracey, S. Patel, Y. Zhang, K. Chappie, D. Volfson, F. Parisi, C. Adans-Dester, F. Bertacchi, P. Bonato, P. Wacnik, Voice biomarkers of recovery from acute respiratory illness, IEEE J. Biomed. Health 26 (2022) 2787–2795.
[158] B. Halpern, R. van Son, M. Brekel, O. Scharenborg, Detecting and analysing spontaneous oral cancer speech in the wild, in: Proc. Annu. Conf. Int. Speech, . Commun. Assoc., 2020, pp. 4826–4830.
[159] S. Quintas, J. Mauclair, V. Woisard, J. Pinquier, Automatic prediction of speech intelligibility based on X-vectors in the context of head and neck cancer, in: Proc. Annu. Conf. Int. Speech, . Commun. Assoc., 2020, pp. 4976–4980.
[160] A. Tsanas, M. Little, P. McSharry, L. Ramig, Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests, IEEE Trans. Biomed. Eng. 57 (2009) 884–893.
[161] B. Woldert-Jokisz, Saarbruecken Voice Database, 2007 [Online]. Available:http://www.stimmdatenbank.coli.uni-aarland.de/help_en.php4.
[162] F. Rudzicz, A.K. Namasivayam, T. Wolff, The TORGO database of acoustic and articulatory speech from speakers with dysarthria, Comput. Humanit. 46 (2012) 523–541.
[163] Cesari Ugo, Giuseppe De Pietro, Elio Marcian, Ciro Nir, Giovanna Sannino, Laura Verde, A new database of healthy and pathological voices, Comput. Electr. Eng. 68 (2018) 310–321.
[164] B.E. Sakar, M.M. Isenkul, C.O. Sakar, A. Sertbas, Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings, IEEE J. Biomed. Health 17 (2013) 828–834.
[165] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, S. Frame, Dysarthric speech database for universal access research, Proc. Annu. Conf. Int. Speech. Commun. Assoc. (2008) 1741–1744.
[166] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S.R. Chetupalli, P.K. Ghosh, S. Ganapathy, Coswara–a database of breathing, cough, and voice sounds for COVID-19 diagnosis, in: Proc. Annu. Conf. Int. Speech, . Commun. Assoc., 2020, pp. 4811–4815.
[167] L. Orlandic, T. Teijeiro, D. Atienza, The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms, Sci. Data 8 (2021) 156.
[168] Y. Liu, T. Lee, T. Law, K.Y. Lee, Acoustical assessment of voice disorder with continuous speech using ASR posterior features, IEEE/ACM Trans. Audio. SPE. 27 (2019) 1047–1059.
[169] D.S. Barbera, M. Huckvale, V. Fleming, E. Upton, J. Crinion, An utterance verification system for word naming therapy in Aphasia, in: Proc. Annu. Conf. Int. Speech, Commun. Assoc., 2020, pp. 706–710.
[170] D.S. Barbera, M. Huckvale, V. Fleming, E. Upton, H. Coley-Fisher, C. Doogan, I. Shaw, W. Latham, A.P. Leff, J. Crinion, NUVA: a naming utterance verifier for aphasia treatment, Comput. Speech Lang 69 (2021), 101221.
[171] S. Jayalakshmy, G.F. Sudha, Conditional GAN based augmentation for predictive modeling of respiratory signals, Comput. Biol. Med. 138 (2021), 104930.
[172] A. Baird, N. Cummins, S. Schnieder, J. Krajewski, B.W. Schuller, An evaluation of the effect of anxiety on speech-computational prediction of anxiety from sustained vowels, Proc. Annu. Conf. Int. Speech. Commun. Assoc. (2020) 4951–4955.
[173] M.J. Kim, H. Kim, Combination of multiple speech dimensions for automatic assessment of dysarthric speech intelligibility, in: Proc. Annu. Conf. Int. Speech, Commun. Assoc., 2012, pp. 1322–1325.
[174] M.J. Kim, Y. Kim, H. Kim, Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model, IEEE/ACM Trans. Audio. SPE. 23 (2015) 694–704.

[175] D. Martínez, P. Green, H. Christensen, Dysarthria intelligibility assessment in a factor analysis total variability space, in: Proc. Annu. Conf. Int. Speech, Commun. Assoc., 2013, pp. 2132–2136.
[176] K.L. Kadi, S.A. Selouani, B. Boudraa, M. Boudraa, Discriminative prosodic features to assess the dysarthria severity levels, in: Proc. Of the World Congr, on Eng., 2013, pp. 2201–2205.
[177] J. Kim, N. Kumar, A. Tsiartas, M. Li, S.S. Narayanan, Automatic intelligibility classification of sentence-level pathological speech, Comput. Speech Lang 29 (2015) 132–144.
[178] P. Kayasith, T. Theeramunkong, N. Thubthong, Speech confusion index (O): a recognition rate indicator for dysarthric speakers, in: Adv. In Natural Lang. Process., Proc, vol. 4139, 2006, pp. 604–615.
[179] C. Bhat, H. Strik, Automatic assessment of sentence-level dysarthria intelligibility using BLSTM, IEEE J. Sel. Topics in Signal Proc. 14 (2020) 322–330.
[180] G. Dimauro, V. Di Nicola, V. Bevilacqua, D. Caivano, F. Girardi, Assessment of speech intelligibility in Parkinson's disease using a speech-to-text system, IEEE Access 5 (2017) 22199–22208.
[181] Y. Qin, T. Lee, A.P.H. Kong, Combining phone posteriorgrams from strong and weak recognizers for automatic speech assessment of people with aphasia, in: IEEE Int Conf Acoust Speech Signal Process Proc, IEEE, 2019, pp. 6420–6424.
[182] T. Lee, Y. Liu, Y.T. Yeung, T.K.T. Law, K.Y.S. Lee, Predicting severity of voice disorder from DNN-HMM acoustic posteriors, in: Proc. Annu. Conf. Int. Speech, . Commun. Assoc., 2016, pp. 97–101.
[183] K. Wahengbam, M.P. Singh, K. Nongmeikapam, A.D. Singh, A group decision optimization analogy-based deep learning architecture for multiclass pathology classification in a voice signal, IEEE Sensor. J. 21 (2021) 8100–8116.
[184] M. Dhanalakshmi, T. Nagarajan, P. Vijayalakshmi, Significant sensors and parameters in assessment of dysarthric speech, Sens. Rev. 41 (2021) 271–286.
[185] K. Harimoorthy, M. Thangavelu, Cloud-assisted Parkinson disease identification system for remote patient monitoring and diagnosis in the smart healthcare applications, Concurr. Comput.-Pract. Exp. 33 (2021) e6419.
[186] A.B. Kambhampati, B. Ramkumar, Automatic detection and classification of systolic and diastolic profiles of PCG corrupted due to limitations of electronic stethoscope recording, IEEE Sensor. J. 21 (2021) 5292–5302.
[187] Y.J. Huang, Y.T. Lin, C.C. Liu, L.E. Lee, S.H. Hung, J.K. Lo, L.C. Fu, Assessing schizophrenia patients through linguistic and acoustic features using deep learning techniques, IEEE Trans. Neural Syst. Rehabil. 30 (2022) 947–956.
[188] S. Bhosale, U. Tiwari, R. Chakraborty, S.K. Kopparapu, Contrastive learning of cough descriptors for automatic COVID-19 preliminary diagnosis, in: Proc. Annu. Conf. Int. Speech, . Commun. Assoc., 2021, pp. 946–950.
[189] E. Casanova, A. Candido, R.C. Fernandes, M. Finger, L.R.S. Gris, M.A. Ponti, D.P. P. Da Silva, Transfer learning and data augmentation techniques to the COVID-19 identification tasks in ComParE 2021, in: Proc. Annu. Conf. Int. Speech, Commun. Assoc., 2021, pp. 4301–4305.
[190] G. Gosztolya, A. Bagi, S. Szalóki, I. Szendi, I. Hoffmann, Making a distinction between schizophrenia and bipolar disorder based on temporal parameters in spontaneous speech, in: Proc. Annu. Conf. Int. Speech, Commun. Assoc., 2020, pp. 4566–4570.
[191] P. Jonell, P. Jonel, B. Mol, K. Hkansso, G.E. Hente, J. Beskow, Multimodal capture of patient behaviour for improved detection of early dementia: clinical feasibility and preliminary results, Front. Comput. Sci. 3 (2021), 642633.
[192] P. Harar, Z. Galaz, J.B. Alonso-Hernandez, J. Mekyska, R. Burget, Z. Smekal, Towards robust voice pathology detection, Neural Comput. Appl. 32 (2018) 15747–15757.
[193] K.A. Al Mamun, M. Alhussein, K. Sailunaz, M.S. Islam, Cloud based framework for Parkinson's disease diagnosis and monitoring system for remote healthcare applications, Future Generat. Comput. Syst. 66 (2017) 36–47.
[194] L. Jiang, B. Gao, J. Gu, Y.P. Chen, Z. Gao, X.L. Ma, K.M. Kendrick, W.L. Woo, Wearable long-term social sensing for mental wellbeing, IEEE Sensor. J. 19 (2019) 8532–8542.
[195] H. Nakamoto, Y. Katsuno, A. Yamamoto, K. Umehara, Y. Bessho, F. Kobayashi, A. Ishikawa, Wearable band-shaped device and detection algorithm for laryngeal elevation in mendelsohn maneuver, IEEE Sensor. J. 21 (2021) 14352–14359.
[196] B. Trinite, Epidemiology of voice disorders in Latvian school teachers, J. Voice 31 (2017) 508, e1−508.e9.
[197] J. Gandhi, A. Gadekar, T. Rajabally, P. Vinayakray-Jani, D. Ambawade, Detection of Parkinsons disease via a multi-modal approach, in: Int. Conf. On Comput. Commun., Netw. Technol., 2021, pp. 1–7.
[198] D. Pustina, H.B. Coslett, L. Ungar, O.K. Faseyitan, J.D. Medaglia, B. Avants, M. F. Schwartz, Enhanced estimations of post-stroke aphasia severity using stacked multimodal predictions, Hum. Brain Mapp. 38 (2017) 5603–5615.
[199] A. Schulte, R. Suarez-Ibarrola, D. Wegen, P.F. Pohlmann, E. Petersen, A. Miernik, Automatic speech recognition in the operating room – an essential contemporary tool or a redundant gadget? A survey evaluation among physicians in form of a qualitative study, Ann. Med. Surg. (Lond) 59 (2020) 81–85.
[200] H. Zhang, Q. Xiao, X. Liu, Construction and application of intelligent mobile ward round system based on speech recognition, Chin. J. Health Inform. And Manag. 17 (2020) 799–803.
[201] Z. Lin, Q. Li, Y. Xiang, Q. Wang, Application and analysis of artificial intelligence voice system in postoperative follow-up of children with congenital cataract, Eye Sci. 36 (2021) 23–29.
[202] Unisound. Intelligent follow-up system. [Online] Available: http://www.unisound.com/follow-up.html.
[203] A. Atrash, R. Kaplow, J. Villemure, R. West, H. Yamani, J. Pineau, Development and validation of a robust speech interface for improved human-robot interaction, Int. J. Soc. Robot 1 (2009) 345–356.
[204] K. Almutairi, S. Abdlerazek, H. Elbakry, A.I. Ebada, Development of smart healthcare system for visually impaired using speech recognition, Int. J. Adv. Comput. Sci. Appl. 11 (2020) 647–654.
[205] S. Lee, S. Kang, D.K. Han, H. Ko, Dialogue enabling speech-to-text user assistive agent system for hearing-impaired person, Med. Biol. Eng. Comput. 54 (2016) 915–926.
[206] L. Fontan, M. Le Coz, C. Azzopardi, M.A. Stone, C. Fuellgrabe, Improving hearing-aid gains based on automatic speech recognition, J. Acoust. Soc. Am. 148 (2020) EL227–EL233.
[207] LAPUL, Preconditioning trainer (BB-RIC-D2) [Online] Available: http://www.lapul.com/Product/7516845939.html.
[208] R. Valencia-Garcia, R. Martinez-Bejar, A. Gasparetto, An intelligent framework for simulating robot-assisted surgical operations, Expert Syst. Appl. 28 (2005) 425–433.
[209] K. Ren, Y. Wang, X. Chen, H. Cai, Speaker-dependent speech recognition algorithm for laparoscopic supporter control, Laser Optoelectron Prog 57 (2020), 181702.
[210] J.H. Tao, M.H. Yang, Z.L. Wang, X.J. Ban, L. Jie, Non-contact multi-channel natural interactive surgical environment under sterile conditions, J. Softw. 30 (2019) 2986–3004.
[211] T.K. Yoo, E. Oh, H.K. Kim, I.H. Ryu, I.S. Lee, J.S. Kim, J.K. Kim, Deep learning-based smart speaker to confirm surgical sites for cataract surgeries: a pilot study, PLoS One 15 (2020), e0231322.
[212] S.X. Li, H. Zhang, J.Z. Liu, W.Q. Yang, K.H. Zhu, Speech control system for intelligent wheelchair based on SPCE061A, Comput. Eng. 34 (2008) 248–250.
[213] M. Al-Rousan, K. Assaleh, A wavelet- and neural network-based voice system for a smart wheelchair control, J. Franklin Inst. 348 (2011) 90–100.
[214] H.T. Wang, Y.Q. Li, T.Y. Yu, Coordinated control of an intelligent wheelchair based on a brain-computer interface and speech recognition, J. Zhejiang Univ. Sci. Comput. & Electron. 15 (2014) 832–838.
[215] A. Punith, G. Manish, M.S. Sumanth, A. Vinay, R. Karthik, K. Jyothi, Design and implementation of a smart reader for blind and visually impaired people, AIP Conf. Proc. 2317 (2021), 060002.
[216] K.R. Jothi, V.L. Mamatha, B.B. Saravana, P. Yawalkar, Speech intelligence using machine learning for aphasia individual, Int. Conf. Comput. Intell. Knowl.Econ. (2019) 664–667.
[217] V. Balaji, G. Sadashivappa, Waveform analysis and feature extraction from speech data of dysarthric persons, in: Int. Conf. Signal Process. Integr, . Netw., 2019, pp. 955–960.
[218] S. Lee, S. Kang, H. Ko, J. Yoon, M. Keum, Dialogue enabling speech-to-text user assistive agent with auditory perceptual beamforming for hearing-impaired, in: IEEE Int. Conf. Consum. Electron., Jan., IEEE, 2013, pp. 360–361.
[219] S. Akbarzadeh, E. Lobarinas, N. Kehtarnavaz, Online personalization of compression in hearing aids via maximum likelihood inverse reinforcement learning, IEEE Access 10 (2022) 58537–58546.
[220] M. Gibson, J.D.O.F.F. Coffin, Recommendations for telemedicine reimbursement, J. Med. Pract. Manag.: J. Med. Pract. Manag. 36 (2021) 226–228.
[221] V.N. Bhatt, Alexa for Health Practitioners, North Dakota State Univ., Ann Arbor, 2020 [Online] Available: https://library.ndsu.edu/ir/handle/10365/31843.
[222] Nuance, Increased health risks and a new telehealth playing field [Online] Available: https://whatsnext.nuance.com/healthcare/increased-health-risks-and-a-new-telehealth-playing-field/.
[223] Y. Liu, S.P. Zuo, C.L. Hsu, Interactive cognitive training tool designed for autism spectrum disorder children, Sensor. Mater. 33 (2021) 405–413.
[224] X. Bu, P.H. Ng, Y. Tong, P.Q. Chen, R.R. Fan, Q.P. Tang, Q.Q. Cheng, S.S. Li, A. Sk Cheng, X.Y. Liu, A mobile-based virtual reality speech rehabilitation App for patients with Aphasia after stroke: development and pilot usability study, JMIR Serious Games 10 (2022), e30196 e30196.
[225] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, A. Diacon, T. Niesler, Wake-Cough: cough spotting and cougher identification for personalised long-term cough monitoring, in: European Signal Proces. Conf., 2022, pp. 185–189.
[226] M. Pahar, I. Miranda, A. Diacon, T. Niesler, Automatic non-invasive cough detection based on accelerometer and audio signals, J. Signal Process Syst. 94 (2022) 821–835.
[227] A.I. Olami, Innovation in smart healthcare: VIA's new smart voice hospital bed card, Available: http://www.cmia.info/news_detail.asp?id=14143.
[228] Zhou Shen, Hai Yun, Guidance robot [Online] Available: http://www.szyh-smart.com/hangye/yiyuan/index.html.
[229] zorarobotics. Healthcare Robots Equipped with the Zora ZBOS. [Online] Available: https://www.zorarobotics.be/use-cases.
[230] Y. Zhang, Y. Diao, S. Liang, C. Ye, Y. Zhou, G. Zhao, Cognitive-motion rehabilitation medical robot application design, Inf. Control 50 (2021) 740–747, 760.
[231] R. Rana, S. Latif, R. Gururajan, A. Gray, G. Mackenzie, G. Humphris, J. Dunn, Automated screening for distress: a perspective for the future, Eur. J. Cancer Care 28 (2019), e13033.
[232] A. Ismail, S. Abdlerazek, I.M. El-Henawy, Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping, Sustainability 12 (2020) 2403.
[233] B.S. Pranathi, A. Nair, C.S. Anushree, T.S. Chandar, Sahayantra - a patient assistance robot, in: Int. Conf. Comput., Commun. Netw. Technol., 2020, pp. 1–6.
[234] C. Hao, L. Fengyuan, Design and implementation of intelligent guidance service robot, Comput. Appl. and Softw. 37 (2020) 329–333.
[235] Y. Chen, Z.R. Zhou, M. Cao, M. Liu, Z.H. Lin, W.X. Yang, X. Yang, D. Dhaidhai, P. Xiong, Extended Reality (XR) and telehealth interventions for children or

adolescents with autism spectrum disorder: systematic review of qualitative and quantitative studies, Neurosci. Biobehav. Rev. 138 (2022), 104683.
[236] A.I. Albarrak, R. Mohammed, N. Almarshoud, L. Almujalli, R. Aljaeed, S. Altuwaijiri, T. Albohairy, Assessment of physician's knowledge, perception and willingness of telemedicine in Riyadh region, Saudi Arabia, J. Infect. Public Heal. 14 (2021) 97–102.
[237] Y.H. Bhosale, K.S. Patnaik, IoT deployable lightweight deep learning application for COVID-19 detection with lung diseases using RaspberryPi, Int. Conf. IoT Blockchain Technol. (2022) 1–6.