Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art

Siddique Latif, Junaid Qadir, Adnan Qayyum, Muhammad Usama, and Shahzad Younis

(Methodological Review)

***Abstract*—Speech technology is not appropriately explored even though modern advances in speech technology—especially those driven by deep learning (DL) technology—offer unprecedented opportunities for transforming the healthcare industry. In this paper, we have focused on the enormous potential of speech technology for revolutionising the healthcare domain. More specifically, we review the state-of-the-art approaches in automatic speech recognition (ASR), speech synthesis or text to speech (TTS), and health detection and monitoring using speech signals. We also present a comprehensive overview of various challenges hindering the growth of speech-based services in healthcare. To make speech-based healthcare solutions more prevalent, we discuss open issues and suggest some possible research directions aimed at fully leveraging the advantages of other technologies for making speech-based healthcare solutions more effective.**

***Index Terms*—Deep learning, automatic speech recognition (ASR), speech synthesis, healthcare, speech biomarkers, remote monitoring.**

I. Introduction

THE current healthcare system is unable to provide universal access to all patients and facing several problems. These problems include: (1) the increasing portion of ageing population, which is expected that the number of people aged 65 or older will rise from 524 million in 2010 to 1.5 billion in 2050 [1]; (2) the increasing burden of chronic diseases, which is expected to globally grow from 46% to 57% by 2020 [2]; (3) the lack of human resources (i.e., doctors and nurses) and healthcare facilities, especially in developing nations[1]; (4) the expensive provision of high-quality care [4]; and (5) the absence of data-driven patient-centred clinical methods, due to which people are being assessed on population averages [5]. To address these challenges, technology-based health can be utilised to provide support to the healthcare system. Especially, speech processing has great potential to provide innovative solutions in healthcare to facilitate both patients and doctors.

Broadly speaking, human speech is the most natural mode of human communication. It provides information about linguistic content and para-linguistic states and traits. The linguistic content represents the intended message that the speaker wishes to convey or communicate. Para-linguistics content of speech provides a much rich array of information related to speakers' identity, gender, and age. Research efforts are exploring the intelligent modelling of speech signals for various important applications. Speech processing research is currently gaining interest to utilise computational paralinguistic analysis for the assessment of different health conditions. The prime reason to use speech for healthcare is that it can be easily available, collected, transmitted, and stored [6]. Most importantly, various physical and mental diseases cause changes in human speech, which are measurable with the help of speech technology.

Speech technology involving the processing and analysis of human speech is a major area of research these days. It encompasses various areas of research such as automatic speech recognition (ASR), speaker recognition/verification, text to speech (TTS) conversion, and identification of language, age, and gender using speech. Research on speech technology has endeavoured to empower machines to involve in verbal human-machine interactions (HCI). These days speech technology-based interfaces have become widely adopted worldwide in various routinely-used devices and applications with services such as Apple's Siri and Google Voice Search used by millions of users [7]. Researchers are now aiming to transform the current verbal HCI interfaces into the next generation medical companions that react with humans more naturally and monitor users' mental and physical health at their home, work, hospital or anywhere. In all these areas, deep learning (DL) is emerging as an essential component of state-of-the-art approaches.

Recent progress in speech processing along with other advanced technologies including the internet of things (IoT) and communications systems can fix the current dysfunctional healthcare system. In particular, recent breakthroughs in DL, the advent of IoT, and the advancement in communication systems have opened up various promising opportunities for healthcare

Manuscript received February 3, 2020; revised May 29, 2020; accepted June 27, 2020. Date of publication July 3, 2020; date of current version January 22, 2021. *(Corresponding author: Siddique Latif.)*

Siddique Latif is with the University of Southern Queensland (USQ), Springfield, QLD 4350, Australia, and also with the Distributed Sensing Systems Group, Data61, CSIRO, Pullenvale, QLD 4069, Australia (e-mail: siddique.latif@usq.edu.au).

Junaid Qadir, Adnan Qayyum, and Muhammad Usama are with Information Technology University, Lahore 54000, Pakistan (e-mail: junaid.qadir@itu.edu.pk; adnan.qayyum@itu.edu.pk; muhammad.usama@itu.edu.pk).

Shahzad Younis is with the National University of Sciences & Technology, Islamabad 44000, Pakistan (e-mail: muhammad.shahzad@seecs.edu.pk).

Digital Object Identifier 10.1109/RBME.2020.3006860

[1]It is anticipated that the world will have a shortage of 12.9 million healthcare workers by 2035 [3].

1937-3333 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

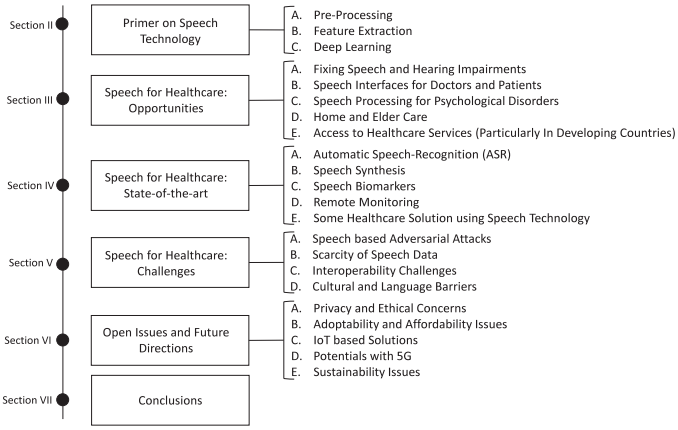

Fig. 1. Organisation of paper.

systems. It will create huge opportunities for speech technology to be utilised for remote diagnostics and monitoring, chronic disease management, and independent care for the elderly and much more. This paper aims to cover the state-of-the-art speech technology and its applications in the field of healthcare.

The **main contribution** of this paper is to highlight the substantial potential of speech technology for improving the state-of-the-art in healthcare. To the best of our knowledge, this is the first comprehensive paper that reviews the state-of-the-art research from the different speech-related fields—including automatic speech recognition (ASR), speech synthesis or text to speech (TTS), and speech biomarkers—to show their potential for healthcare. This work builds upon previous papers with a limited focus that aimed to demonstrate the potential of speech recognition for healthcare [8]–[13] or discuss the use of DL for health-related sub-challenges on publicly available datasets [14]. We feel that this paper is especially timely due to the recent advancements in speech technology and DL along with problems faced by an increasingly burdened healthcare system that is crying out for technological augmentation.

The organisation of the paper (see Fig. 1) is described next. In section II, a brief premier on speech technology is presented followed by a discussion of the potential opportunities of speech technology in healthcare in section III. Next, we cover the state-of-the-art works on speech technology for healthcare and some of its prominent healthcare solutions which is followed by challenges that are causing hurdles for speech technology to be utilised in healthcare in section V. Before concluding the paper, we discuss important open issues and future direction that can help researchers to make the use of speech technology more effective in section VI.

II. Primer on Speech Technology

Speech technology aims to enable machines to recognise, analyse, and understand human speech. The area has been developing for decades as a sub-field of signal processing and has seen much progress in the last decade or so due to the huge progress made under deep learning paradigms. Typically speech technology systems include three major components (as shown in Fig. 2) that include pre-processing, feature extraction, and ML algorithm(s) development.

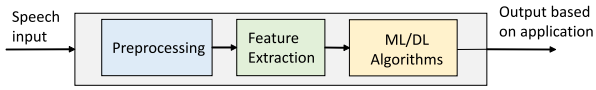

Fig. 2. Major components of speech technology based systems.

A. Pre-Processing

Pre-processing of speech signals is considered as an important step in designing robust and efficient systems for various applications. It usually involves noise suppression, silence removal, and channel equalisation, etc. The performance of speech-based systems can be improved with the use of these preprocessing techniques [15], [16]. It has been also validated that removing silence pauses and noise even helps ML/DL models in achieving better performance [17]. In the case of speech synthesis, text processing is even more complex, it involves text normalisation, tokenisation, sentence segmentation, etc [18]. It is equally important in speech synthesis to improve system performance.

B. Feature Extraction

The representation of speech signals into meaningful, informative, and a reasonably limited number of features is a crucial component for developing any system for speech-based applications. Despite the fact that there is no unique taxonomy of speech features, it is common to divide features into two types, i.e., *linguistic* and *acoustic* features. The extraction procedures for these two types of features are significantly different and their performance greatly depends upon the type of problem at hand.

Linguistic features represent information in spoken words. This usually includes specific words, their grammatical alterations or higher semantic and pragmatic markers [19]. A variety of techniques exists for the analysis of speech using linguistic features. For example, key-word spotting aims at the reliable detection of a particular word in a given speech [20], [21]. These words are chosen from daily life and considered sufficient to represent the speakers' states and related events.

Acoustic features are very popular and widely being used. These features are primarily extracted using the models of the human auditory system. Human hearing-related properties such as lower sensitivity at lower frequencies, spectral amplitude compression, nonlinear frequency scale, and large spectral integration are also considered in acoustic features [22].

Ongoing research on speech analysis has categorised the acoustic features into three categories: prosodic, spectral and temporal, and features related to the voice quality [23], [24]. Prosody refers to melody and rhythm of speech and prosodic features include the feature related to the length, tone, accent, stress, intonation, and few others [25], [26]. These features can be used to detect the irregularities in the rhythm and timing of speech. For instance, nonverbal speech cues such as interruptions natural turns, counting the number of interjections,

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

TABLE I
SUMMARY OF SOME POPULAR DL MODELS

Model	Characteristics	References
DNNs	Consists of fully connected layers and popular in learning a hierarchy of invariant and discriminative features. Representations learnt by DNNs are more generalised than the traditional hand-engineered features.	[35]
CNNs	Convolutional layer is main building block of CNNs. They were designed for image recognition but also extended for speech technology. CNNs can learn a high-level abstraction from input data.	[36] [37]
RNNs	RNNs are popular for temporal structure learning from sequential data like speech and text. They can learn temporal contexts from speech and perform better than standard DNNs	[38] [39]
AEs	Powerful unsupervised learning models that encode and learn representations form the data in sparse and compress representations.	[40]
VAEs	They are stochastic variational inference and learning model. Good for generating samples and also learning disentangled representations.	[41]
GANs	Game-theoretical model and very powerful for data generation. They are found robust to overfitting and already shown promising results in speech modelling and synthesis.	[42] [43]

and response times can help to identify irregular speech patterns [27]. In contrast to temporal features, spectral features are computed by converting the speech signal into the frequency domain. Most popular temporal and spectral features include energy [28], entropy [29], zero-crossing rate (ZCR) [30], spectral centroid, spectral roll-off, and spectral flux [31]. Similarly, voice quality-related features include jitter, shimmer, unvoiced rate, and harmonic-to-noise ratio (HNR), etc.

Recently, the speech community also started using raw speech instead of hand-engineered features. For these features, they use DL models to extract data-driven features related to the task at hand. Such features have shown promising results for different speech-related tasks including ASR [32], emotion detection [33], and speaker identification [34]. However, the performance of such methods needs to be explored and compared with various hand-engineered features for health-related tasks.

C. Deep Learning

In speech technology, the hidden Markov model (HMM) and Gaussian mixture model (GMM) based models (GMM-HMM) have ruled for decades. A disruptive breakthrough happened in speech technology in the last decade due to DL and now DL models have become an essential component of ASR, TTS, and other speech processing and analysis tasks. This section is aimed to present an introductory and higher-level overview of DL. For an in-depth description, the interested readers are referred to classical resources [44]. However, we presented the summary of different DL architecture in Table I. DL is distinguished from legacy artificial neural networks (ANNs) in terms of having two or more layers between input and output layers. The basic component of a deep neural network (DNN) is the neuron unit. The neurons in each layer are fully connected with the neurons of the adjacent layers to create a network. The input signal is passed through the network with intermediate computation and approximate function $y = f(x;\theta)$ by learning the best value of the parameters θ. A multi-layer network or DNN creates a pipeline of non-linear transformations with the ability to learn intermediate representations, suitable for a given task at hand.

In 2006, an idea of the learning hierarchy of feature representations in different layers of the deep learning models was initiated by Hinton [45] and models like deep belief networks (DBN) [46] and stacked autoencoders (SAE) [47] were proposed. These deep typologies take the advantages of unsupervised layer-by-layer pre-training which is followed by fine-tuning of the entire network using back-propagation. These models were widely utilised in speech technology and still being considered for speech modelling [48]. More recently, research in speech technology has been focused on end-to-end learning paradigms from raw speech using convolutional neural networks (CNNs). They can learn filterbank from raw speech and able to capture more generalised, discriminative, and contextual representation from raw waveform [33].

Convolutional neural networks (CNNs) were originated from image processing for processing data in grid-like topology. They are also extended for natural language processing (NLP) and speech processing. The building block of CNNs is a convolutional layer that consists of multiple filters and it computes local feature maps from the input. The convolutional operation in CNNs can be defined as:

$$(h_k)_{ij} = (W_k \otimes q) + b_k, \tag{1}$$

where $(h_k)_{ij}$ is the $(i,j)^{th}$ element for the k^{th} output feature map, q represents the input feature maps, and W_k and b_k denote the k^{th} filter and bias, respectively. The symbol $\otimes$ represents the 2D convolution operation. The second component of CNNs is the pooling layer to facilitates nonlinear sub-sampling operations is to reduce the dimension of each feature maps while retaining the most important features. Finally, fully connected layers are used to achieve the required prediction for regression or classification tasks. In speech processing, it is very common to use CNN in conjunction with recurrent neural networks (RNNs).

Recurrent neural networks (RNNs) define a special DL architecture that uses recurrent connections within layers with the capability of processing previously processed inputs. In contrast to hidden Markov models (HMMs), RNNs have stronger representational memory [38] and are better suited for modelling sequences structures like speech. For an input sequence $x(t) = (x_1, \ldots..., x_T)$ at time step t, RNNs calculate the hidden state h_t by using the previous hidden state h_{t-1} and produce a output vector sequence $y_t = (y_1, \ldots..., y_T)$. The equations for standard RNNs are given below:

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \tag{2}$$

$$y_t = (W_{xh}x_t + b_y), \tag{3}$$

where W terms denote the weight matrices, b represents bias vector, and H defines the hidden layer function. Simple RNNs face vanishing gradient problem and fail to model the long-term temporal contingencies. To deal with this problem, multiple specialised RNN architectures were proposed. These include long short-term memory (LSTM) [49] and gated recurrent units (GRUs) [38] with gating mechanism to add and forget the memory selectively. Bidirectional RNNs [50] were also proposed

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

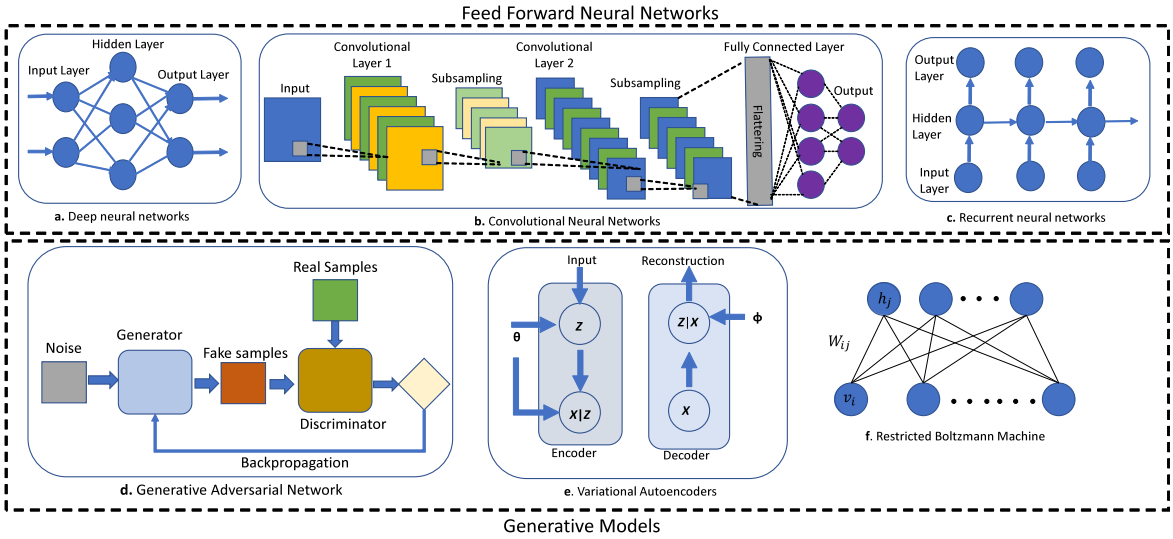

Fig. 3. Graphical illustration of different machine learning (ML)/deep learning (DL) models.

by passing the input sequence through two different recurrent hidden layers to enable both past and future modelling. These gated RNNs solved the issue of the vanishing or exploding gradient problems and can learn long-term contextual dependencies from the input sequence [51] and are widely used in speech technology [52], [53].

Generative models including variational autoencoders (VAEs) [54], [55], generative adversarial networks (GANs), and autoregressive generative models [56], [57] are being used in speech technology. GANs are becoming very popular in speech technology due to their ability to learn and generate data distributions. They include two neural networks—a generator, G, and a discriminator, D. Both these networks play a min-max adversarial game defined by the following optimisation:

$$\min_{G}\max_{D} \quad \mathrm{E}_x[\log(D(x))] + \mathrm{E}_y[\log(1 - D(G(y)))]. \tag{4}$$

For speech synthesis, autoregressive generative models like WaveNet [57] provide state-of-the-art results and are becoming increasingly popular. They are directed probabilistic models and can model joint distribution using the following chain-rule:

$$p(x) = \prod_{t=1}^{T} p(x_t|x_1, \ldots\ldots, x_{t-1}; \theta), \tag{5}$$

where x_t is the t^{th} variable of a waveform $x = \{x_1, ..., x_T\}$ and θ are the parameters of the autoregressive model. Some other popular autoregressive models include PixelRNN [56], and PixelCNN [58]. A graphical depiction of various ML/DL models can be seen in Fig. 3.

III. Speech for Healthcare: Opportunities and Applications

The current healthcare system is struggling to provide quality health services at an affordable price. The effectiveness of health services can be significantly enhanced by using the opportunities offered by speech technology [59]. Such opportunities are highlighted in Fig. 4 and this section provide a detailed discussion on these opportunities.

A. Fixing Speech and Hearing Impairments

Humans express their feelings, thoughts, and ideas by speaking. Speech is produced by the action of coordinated muscles in the head, neck, chest, and abdomen. The development of speech is a gradual process that involves years of practice. During this process, a human child learns how to regulate these muscles to produce understandable speech. Individuals that are unable to properly regulate these muscles face speech disorders. Disorders related to voice and language also affect human communication [60]. Hearing problems are also a cause of imperfect communications among humans. Individuals who do not hear someone with normal hearing thresholds of 25 dB or better in both ears face hearing impairments. According to the world health organisation (WHO), over 5% of the world's population (466 million people) has hearing loss and it is expected that this number will increase up to 900 million people by 2050 [61].

Speech technology can be utilised to assist individuals with a hearing problem or a voice, speech, or language disorder to communicate effectively [62]. ASR plays an important role in the applications of speech therapy, which require to decode the user utterances [63]. Similarly, speech synthesis provides can teach users how a word or sentence should be pronounced which reinforcing the correct pronunciation in the speech therapy activities. In this way, systems based on speech synthesis and ASR can be used to augment the quality of human communication in healthcare [64], [65]. Such assisting systems aim at improving the intelligibility of pathologic speech by producing speech similar to the voice of the speaker [63]. Speech technology can also be used for design interfaces for children's speech therapy [66]. Dysarthria, one of the major speech disorders that arise as a secondary condition among individuals with cerebral palsy, amyotrophic lateral sclerosis (ALS), stroke survivors,

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

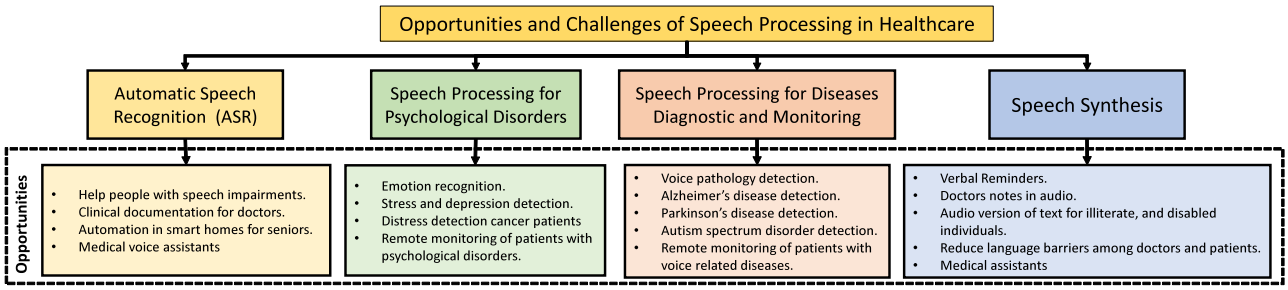

Fig. 4. Prominent opportunities of speech processing in healthcare.

multiple sclerosis, Alzheimer's, Parkinson's, and traumatic brain injury. It produces weakness or difficulty controlling the muscles involved in speech production. Studies demonstrate that speech technology-based systems can also help people with such disorders [67].

B. Speech Interfaces for Doctors and Patients

Speech interface describes an automated software that uses either simulated human speech or human speech to interact with humans. They provide services in hands-free, eyes-free, and keyboard-free situations. With 50% of a healthcare professional's time is spent on clinical documentation, overwhelmed clinicians often spend more time on clinical documentation compared to direct patient care [68]—a fact that can drain and demotivate clinicians. Speech interfaces can provide a convenient integration in healthcare to reduce the burden of medical doctors. The implementation of such interfaces can bring a lot of benefits to the healthcare industry, from time-saving to cost-reducing.

First of all, practitioners and nurses will be able to become more efficient in the process of transcripts [69]. Secondly, it will increase productivity in healthcare systems [70]. Another benefit is that it will also reduce the amount of time by a doctor to see the patients by using readily available information. Similarly, speech technology is a major mean of reducing the cost of traditional medical transcription in healthcare systems. More than 30% of institutions were able to save more than $1 million over the period of two or more years [70]. Voice-based assistants can maintain patients' electronic medical records (EMR) and provide relevant information when needed. In a recent controlled observational study [71], participants found that speech recognition in clinical documentation can saves time, increases efficiency, and allows to make more detailed notes with relevant details. Speech technology-based interfaces can help the patients during their hospital stay and after discharge. In particular, speech interfaces can facilitate patients recovering at home, especially when they have restricted mobility, through support for environment control (such as adjusting a room's temperature, controlling audio levels, requesting nursing assistance, and in decision support) [72], [73].

C. Speech Processing for Psychological Disorders

The term psychological disorder, which refers to psychiatric disorders or mental disorders, are behavioural or psychological symptoms or patterns, which impact multiple areas of life. Humans feel a vast range of comfortable and uncomfortable emotions [74] and it has been argued that emotional discomfort is a universal human experience [75]. Emotions are transient and continually fluctuating, which may cause both positive and negative effects on human life [76]. On the other end, psychological disorders like stress, depression, anxiety, suicidal behaviour, and distress, which can lead to disabling conditions and impairment in people. Distress is major emotional suffering and highly prevalent in patients with a chronic disease like cancer [77]. Despite the fact that psychological disorders can cause serious consequences, routine screening has not been widely adopted in healthcare due to heavy cost and time requirements [78].

Recent studies have shown the promise of using speech as an effective biomarker for the diagnosis of psychological disorders. Spoken speech can provide a wide range of acoustic features that can be effectively utilised for human emotion detection [53], [79] and diagnosis and monitoring of depression, anxiety, stress, distress, and suicidal behaviour [76]. State-of-the-art deep learning models have improved the performance of emotion recognition, depression, anxiety, stress, distress, and suicidal behaviour detection using speech [80], [81]. Particularly, the combination of CNNs and LSTM has shown great performance in modelling affective behaviours and related disorders from speech [82], [83]. Here CNN is mainly used to extract temporal features and contextual modelling is performed using LSTM. CNNs are also being employed in an end-to-end fashion to extract features from raw speech related to the problem at hand, e.g., depression detection [84]. Convolutional layer in CNNs acts a data-driven filterbank that can produce more generalised features compared to the standard artificial neural networks (ANNs) and other feature-based approaches [33]. In this way, deep learning models are playing an important role in modelling as well as diagnosis of different psychological disorder using speech signal. This demonstrates that the speech technology has great potentials to automate the screening and monitoring of mental illnesses and related disorders, hence alleviating many healthcare challenges.

D. Home and Elder Care

Life expectancy is greatly increasing worldwide, which is causing a higher number of older people in our society [5]. The increased share of the elderly population is shifting the cause of death from parasitic and infectious diseases to chronic

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

TABLE II
COMPARISON OF RECENT STUDIES ON ASR

Studies	Details	Dataset	WER (%)
Chiu et al. [91]	LSTM based sequence-to-sequence LAS model with multi-head attention.	12,500 hour data	5.6
Kriman et al. [99]	Convolutional residual network based end-to-end network is used with Connectionist Temporal Classification (CTC) loss.	LibriSpeech [100]	2.69
Kahn et al. [101]	They explored self-training in the context of end-to-end ASR using wav2letter++ framework [102]	LibriSpeech	5.79
Wang et al. [103]	Transformer-based acoustic models is evaluated for hybrid speech recognition.	LibriSpeech	2.60
Park et al. [104]	They used SpecAugment data augmentation in LSTM based sequence-to-sequence LAS model.	LibriSpeech	2.5
Luscher et al. [105]	Encoder-decoder-attention model in combination with a Transformer language model is evaluated.	LibriSpeech	1.9
Hsu et al. [106]	Meta learning approach for low-resource ASR is evaluated for different languages. Results showed that proposed model significantly outperforms the state-of-the-art models on all target languages	IARPA BABEL project corpus [107]	–

illnesses [85], [86]. Ageing also causes physical limitations that need to be compensated by the assistance of someone or with the services of aged care centres. Elderly people feel isolation, fear, and a sense of helplessness both in-home and at old-age care centres, which cause severe consequences on both physical and mental health. Speech-based assistants are a valuable tool for seniors staying at home or age care centres, especially for those who are not able to use other technology-based services that may require the dexterity of the hands, mobility, and/or good vision. Such systems can also provide them with independence and a better quality of life with physical or cognitive disease [76].

E. Access to Healthcare Services (Particularly in Developing Countries)

Millions of people in developing countries unnecessarily suffer and eventually die from such illnesses that have effective cure and prevention [85]. Generally, people in developing countries have poor literacy skills (i.e., reading, writing). It has been shown that people with low health literacy have a one-in-three chance of misunderstanding the prescribed medication [87]. There is a direct correlation between mortality rates and poor health literacy, approximately 50% to 80% increased mortality risk for individuals with poor health literacy [88]. Language diversity is another challenge in developing countries that reduces the potential benefits of healthcare services such as text messaging and e-health portals [89]. Such healthcare services are also not much useful for low-literates, the blind, the visually impaired, and those that are not computer literate. Speech is a plausible interaction modality for illiterate users and speech-based healthcare services can be ideal for inhabitants of the developing countries.

IV. SPEECH FOR HEALTHCARE: STATE-OF-THE-ART

A. Automatic Speech-Recognition (ASR)

Automatic speech recognition (ASR) is the analogue of machine ear, which enables a computer to recognise uttered speech and transform it into the corresponding sequence of words or sub-words. ASR has witnessed a steady improvement in performance due to the development of cutting-edge ML algorithms. Traditionally, HMM and GMM based models were the main stock of research for ASR. DNN-based ASR systems have become the state-of-the-art by showing huge improvements compared to previous conventional system [90]. Developing and training ASR systems, however, is complicated and requires a lot of pre-processing. Various attempts have been made to reduce the complexity of ASR, paving the way for end-to-end speech recognition [91]. Nowadays end-to-end ASR systems are extensively used and studied for ASR in different languages such as English, Mandarin, Japanese, or French [92]. Similarly, sequence-to-sequence models are also gaining popularity in the automatic speech recognition (ASR) community [91]. Various sequence-to-sequence models including Recurrent Neural Network Transducer (RNNT) [93], Neural Transducer [94], Listen, Attend and Spell (LAS) [95], Recurrent Neural Aligner (RNA) [96], and Monotonic Alignments [97] have been explored in the literature. In addition, transformers [98] based models are also gaining success in ASR field due to their better performance. We presented the performance comparison of different state-of-the-art models in terms of word-error-rate (WER) in Table II. These systems demonstrate promising results on different datasets which shows the feasibility of their integration into healthcare applications.

A major application of ASR in healthcare is to facilitate the generation of clinical documentations [9]. The medical errors caused by bad handwriting can be avoided using speech recognition for medical documentations [108]. Such self-typing systems are also believed to enhance documentation quality and efficiency, as well as improve the satisfaction level of health professionals in clinics or hospitals [71].

Different research studies explored the feasibility of ASR systems for clinical documentation. For instance, in [109] authors evaluated a web-based ASR system in a university hospital for clinical documentation in the German language. They found that medical documentation with ASR increases in documentation speed and amount, and it also has a positive impact on participant mood in contrast to self-typing. Hodgson *et al.* [110] explored the use of ASR for medical transcription of the doctor-patient conversation. They used 14 000 hours of speech and demonstrated that the proposed models achieved promising results on important medical utterances and therefore can be used practically in a clinical setting for transcribing medical conversations. In [111], authors performed a case study in a specialised outpatient department and found that ASR software supports medical doctors by quickly producing patient discharge letters without impairing user satisfaction.

The efficiency of speech recognition is evaluated by Hoyt *et al.* [112] for documenting outpatient encounters in the EHR system at a military hospital and its 12 outlying clinics. Seventy-five clinicians participated to evaluate speech recognition for clinical documentation. Among these participants, 69% of the clinicians continued to use speech recognition in their routine practices and reported that speech recognition for clinical

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

documentation is more convenient, accurate, and expeditious (for instance, speech recognition helped in improving note quality and allowed for closing a patient encounter on the same day.) Similarly, authors in [113] showed that medical speech recognition can perform on a par with humans. Authors in [114] developed a web-based prototype to generate medical reports in the Brazilian Portuguese language using Google Web Speech API and Microsoft Bing Speech API. They found that a system based on Google API achieved an error rate of 12.30%, which was significantly better than those achieved by the Microsoft API 17.68%. Few other studies [115] also highlighted the potential of using ASR for clinical documentation.

All of these studies highlight several benefits of using ASR for clinical documentation. However, most of these studies are pilot projects. Although the use of ASR can drive interactive clinical documentation, however, careful evaluation is required for EHR [118]. There is still a need to improve the efficiency of ASR for medical documentation to avoid errors that have the potential to cause clinical harms. Furthermore, improved system interoperability and workflow are needed for their successful integration in the clinical setting [119].

B. Speech Biomarkers

Human voice incorporates features that can plausibly be used to discriminate on the basis of gender, age, intelligence, socio-economic status, regional/ethnic origin, education, and occupation [139]. Most importantly for health outcomes, it provides information about various voice disorders, which can be diagnosed by detecting anomalous voice quality, pitch, and loudness that is inappropriate for an individual's age, gender, cultural background or geographic location [120]. Other speech-related disorders include cognitive-communication disorders, communication disorders, swallowing disorders, and an autism spectrum disorder. Speech technology-based solutions have been playing an important role in the diagnostic and monitoring of these disorders. Nowadays, DL models have become the state-of-the-art technique in this domain.

Automatic detection of vocal fold pathologies is of great interest to the voice community as well as the medical community due to its low cost and non-invasive nature. These systems can be used by clinicians to detect the existence of any voice pathologies even in the early stages. Fang *et al.* [121] retrospectively collected normal pathological voice samples of 8 common clinical voice disorders and evaluated both ML models and DNNs. They found that DNNs outperformed other ML models including SVM and GMMs. Authors in [122] used convolutional DBN for voice pathology and showed that CNN can effectively extract features from spectrograms of voice recordings suitable for diagnosing of voice disorders. Harar *et al.* [123] conducted a preliminary study on voice pathology using DNNs and showed that the use of combined CNN-LSTM provides promising results. In [124], the authors investigated a voice pathology detection system using DL on a mobile multimedia healthcare system and voices samples captured using mobile devices. More specifically, the authors used a CNN architecture and reported significantly improved results. In other work, some researchers [125], [126] designed voice pathology detection systems for smart cities but their work used classical ML models in their architectures.

Acoustic analysis of speech is also used for the diagnosis of Alzheimer's disease, which is cognitive impairment and the most common cause of dementia. It has a high prevalence that is increasing rapidly towards an epidemic level. The research community is trying to utilise speech technology to solve this issue. In [127], Lopez-de-Ipina *et al.* proposed a non-linear multi-task approach using a multilayer perceptron (MLP) and CNNs for Alzheimer's detection. The authors evaluated the proposed models using different speech features and reported promising results. In [128], Fraser *et al.* explored the use of linguistic features for the identification of Alzheimer's. They achieved state-of-the-art results and found that linguistic analysis using modern ML is increasingly useful in assessment and clustering of Alzheimer's.

Speech analysis is also being utilised for Parkinson's disease (PD) detection. For instance, authors in [129] designed to diagnosis PD using speech signal. They used DBNs for classification and achieved significantly improved results attesting the power of DBN for speech analysis. Frid *et al.* [130] used raw speech for PD detection using CNNs. They found that relatively small (20 ms) of raw speech contains much information regarding the PD that can be used for classification. Speech technology is also being used for the diagnosis of autism spectrum disorder (ASD) in children. Different studies used DL and speech signals for ASD detection [131], [132] also to assist children with ASD [133].

Human speech provides a wide range of prosodic and spectral features that can be effectively utilised for emotion recognition, depression, distress anxiety, and stress detection. Acoustic features including spectral, prosodic, cepstral, glottal, and Teager energy operators (TEO) were evaluated in for clinical depression detection in adolescents [134]. Authors found that TEO based features produced more promising results compared to all other features. In [135], the authors showed that the prosody and voice quality-related speech features can be used for the identification of suicidal and non-suicidal adolescents. In [136], behaviour prediction of cancer-afflicted patients is performed using different speech features. They showed that speech can effectively be utilised for behavioural prediction in cancer patients. Various other studies [137]–[139] utilised speech feature for identification stress, anxiety and depression in English and also in other languages.

Most of the above-mentioned studies have used publicly available datasets and showed that speech technology can be effectively utilised as biomarkers for the detection of various diseases. However, these systems have not been evaluated in real-life settings. It is important for researchers to focus on the design of systems that can be utilised in clinics and medical hospitals and on reporting real-life performance evaluations.

C. Remote Monitoring

The rising burden on the global healthcare system along with the limited availability of trained healthcare professionals is increasing the demand for infrastructure and technology that can facilitate the remote monitoring of patients. Speech

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

technology-based remote monitoring services have been explored in this regard by the research community. In [140], Hossain used speech along with and facial expressions, which are captured in a multi-sensory environment. The author tested the proposed framework on 100 people and was able to detect the patients' state with an average recognition of 98.2%. In another work, Vatanparvar *et al.* [141] designed a speech privacy preservation method for speech-based remote health monitoring using GAN and reported promising results. A remote system for monitoring of speech-language Intervention is proposed in [142] for the parents of children with ASD. For remote monitoring and assessment of cognitive function in senior people, a system was proposed by Rapcan *et al.* in [143]. The authors used telephone speech recordings and showed that the system can achieve similar results compared to speech recorded in a controlled environment. The SweetHome project was proposed in [144], which used noise-robust multisource ASR to detect distressed sentences in a realistic environment of a smart home to monitor distress situations.

Various studies also have proposed speech remote monitoring systems by exploiting the advanced communication technologies. A 5-G enabled emotion-aware healthcare framework was proposed in [145]. The authors evaluated the proposed healthcare framework on 50 university-level students, who were asked to express the emotion of pain. They used both speech and video as the input of the system and achieved 99.95% of accuracy. A privacy-enhanced emotion recognition system for remote advisory is presented in [146], the authors showed that the proposed system can solve privacy issues while achieving promising results. Similarly, an edge-cloud based privacy-preserving automatic emotion recognition system is proposed which use both speech and visual features [147]. They used CNNs for emotion classification and achieved improved results compared to the state-of-the-art systems. Some other studies [148]–[150] have also used speech signal for remote patients monitoring, however, most of these studies evaluated the proposed models in a controlled setup. Therefore, to ensure a safe and robust operation, it is very important to evaluate these systems in real-life settings.

D. Speech Synthesis

Speech synthesis, also known as text-to-speech (TTS), is an important technology that aims to convert text into speech. Most of the TTS systems use acoustic or linguistic features as an intermediate representation to generate the waveform. Traditionally, the speech waveform was vocoded from these intermediate representation using heuristic methods [151] or using hand-crafted vocoders [152], [153]. Recently, Tacotron 2 [154] used WaveNet [57] as a vocoder to generate waveform from Mel-spectrograms. WaveNet is an autoregressive generative model that can generate relatively realistic human-like speech by using linguistic features. It has the downside of a long inference time due to its autoregressive architecture. To address this issue, various models such as FFT-Net [155], WaveRNN [156], and WaveGlow [157] have been proposed. Nowadays, these neural vocoders have replaced the use of traditional heuristic methods and can dramatically enhance the quality of generated speech. Researchers are also focusing on synthesising more natural speech by transferring prosody [158], style [159], and expressions [160]. All these studies have highlighted the great progress made by TTS systems that shows their suitability in healthcare. We also compared the performance in terms of mean opinion score (MOS) of different state-of-the-art TTS systems in Table III, which depicts that these systems are achieving MOS almost similar to the ground truth speech.

TABLE III
MEAN OPINION SCORE (MOS) EVALUATIONS COMPARISON FOR VARIOUS SYSTEMS

Systems	MOS
Parametric [161]	3.492±0.096
Tacotron [162]	4.001±0.087
Concatenative [163]	4.166±0.091
WaveNet [164]	4.341±0.051
Transformer [164]	4.39± 0.05
Tacotron 2 [154]	4.526 ± 0.066
Ground Truth	4.582± 0.053

Text-to-speech (TTS) solutions can assist healthcare's mission of bettering patient care through the use of assistive and digital tools. It can further enhance digital health technology by speech-based health apps, websites, and emergency call systems, etc. Patients can be verbally reminded about important alerts using TTS, which increases the usability and accessibility of portable health trackers [165]. Different studies evaluated the feasibility of TTS assistive healthcare system. For instance, Liu *et al.* [166] designed a smartphone-based system to help visually impaired people while using Android phones. In [167], the authors presented a design of a TTS-based interactive medication reminder and tracking system for wrist devices. They evaluated the system for both native and non-native English speakers in controlled experiments and achieved very promising results. A voice interactive assistant was designed in [168] to improve adherence to medical treatments. The authors designed this system for stroke patients and tested among several healthy subjects for an initial assessment. However, the final product was reported to be still under development.

TTS health assistive tools can dramatically improve people' health and required costs. For example, a TTS-based system can facilitate patients by offering them an audio version of digital text [169], [170]. This is especially helpful for illiterate individuals, language learners, and the elderly population, and people with learning disabilities or reduced vision. Such systems can also help people by providing an audio version of important medical information such as descriptions of diseases, prescriptions, and drug information leaflets. This avoids drug misuse while making patients cautious about their health [171]. TTS-based systems also allow patients to accurately communicate their needs, which helps to establish a cognitive connection among doctors and patients [172].

E. Some Healthcare Solution Using Speech Technology

As outlined above, speech technology finds its market in healthcare particularly due to potential and impactful use cases. Various voice-enabled healthcare solutions are developed that can help improve the lives of thousands of individuals. We presented the details of some prominent solutions in Table IV.

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
SOME PROMINENT SPEECH TECHNOLOGY BASED HEALTHCARE SOLUTIONS

Application	Product	Details
Physician Notes	Nuance Dragon Medical Practice Edition [173]	It makes easier for practitioners to document care into the EMR using their speech five times faster than typing. It contains more than 60 specialised British English medical vocabularies.
	MDOps [174]	It is voice-enabled interface that allows doctors to dictate clinical data to patients.
	Suki [175]	It is an AI-powered, speech based digital assistant that help the doctors by lifting burden of medical documentation.
	Notable [176]	It uses wearable technology and automatically enrich the patient-physician interaction.
Elderly Care	LiSA [177]	It is a voice enabled learning interface for elderly population that helps them to connect with family, adopt healthy habits, get access to daily routine services, and thrive independence.
	ElliQ [178]	It is is a friendly and intelligent voice interface that offers you various tips and advice, answer your questions, and provide you surprising suggestions.
	LifePod [179]	Its a proactive-voice caregiving service that is designed to monitor daily routines and help improve the quality of life for caregivers and their loved ones.
	Reminder Rosie [180]	It is a is a voice activated hands-free memory aid and daily organiser. It helps individuals to remember their medication, medical appointments, and routine tasks.
Speech Biomarkers	BeyondVerbal [181]	This solution extracts different acoustic features from speech in real time and provides insights on personal wellbeing, emotional condition, and health.
	Cogito [182]	It uses vocal signal and and perform predictive analysis to improve care management using emotional intelligence in real-time.
	Corti [183]	It aims to act as an intelligent partner of medical professionals and help them to make life-saving decisions and diagnose illnesses by listening and analysing the medical interviews or emergency calls.
	WinterLight Labs [184]	It performs analysis on speech and language patterns and to help diagnose and monitor mental illness and cognitive impairment associated with dementia.
Speech and Hearing Difficulty	Ava [185]	It helps deaf and people with hearing impairments by instantly showing them what people say.
	VocaliD [186]	It aims to create unique vocal personalities for individuals having speaking problems.
	Voiceitt [187]	It uses speech recognition to understand non-standard speech and aims to help people with speech motor disabilities.
Patient Engagement	CardioCube [188]	It is a voice interface that aims to help patients with chronic heart disease to manage their health. They can use it to schedule of doctor appointments, give condition updates, request medication refills, etc.
	CareAngel [189]	It is an intelligent virtual nurse assistant that provides a continuous health management at lower cost and improve outcomes.
	Sensely [190]	It is multilingual platform that intelligently guides users about insurance services and healthcare resources.
	Dr. AI [191]	This solution is trained on medical knowledge of thousands of doctors and millions of patients' questions. It provides the individualised support by instantly translating your symptoms into a course of care.

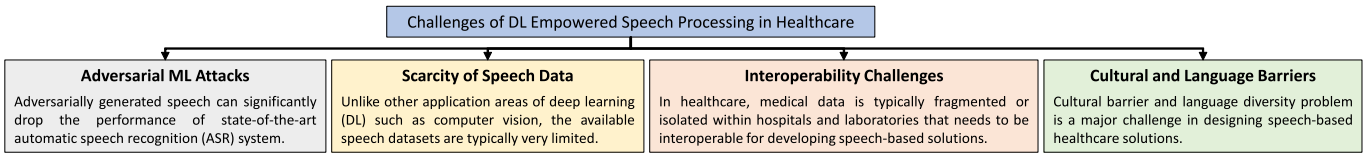

Fig. 5. Challenges of deep learning (DL) empowered speech processing in healthcare.

These solutions are being used by both doctors and patients to change the traditional setting of health provision. The development of speech-based solutions is continuously growing that will significantly impact the current healthcare system in the next several years.

V. SPEECH FOR HEALTHCARE: CHALLENGES

Despite the promising potentials of speech technology in the health domain, there are various hurdles in large-scale deployment of speech-enabled solutions. In this section, we discuss these challenges that need to be addressed to make the rapid adoption of speech technology in healthcare, a taxonomy is depicted in Fig. 5.

A. Speech Based Adversarial Attacks

Even though modern DL-based speech solutions offer great benefits to the current healthcare system, there are still questions about the security of the underlying DL algorithms as recent works have shown that DL models are prone to "adversarial attacks." Adversarial attacks are launched by creating *adversarial examples*, in which non-random imperceptible perturbations are added to input samples through optimization algorithms that aim to fool the classifier and influence it to make incorrect decisions. These attacks are powerful enough to significantly bring down the performance of the state-of-the-art DNN based systems [192]. A popular method for generating such adversarial attacks is to generate the perturbation by utilizing gradient-based methods—many popular attacks such as fast gradient sign method (FGSM) [193], Jacobian-based saliency map attack (JSMA) [194], DeepFool [195], and Carlini and Wagner attacks [196] follow this method. Many more sophisticated attacks also exist including those that rely on non-gradient based methods [197].

Researchers have also proposed various adversarial attacks against speech-based systems. For example, Carlini and Wagner [198] evaluated an iterative optimisation-based adversarial attack against DeepSpeech [199], a state-of-the-art ASR model, with 100% success rate. Some other popular adversarial attacks against speech-based systems include [200]–[204]. The success of these adversarial attacks highlights the vulnerability of speech

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

technology to for healthcare. Therefore, it is necessary to design such systems that preserve patients' privacy in the healthcare settings.

B. Scarcity of Speech Data

To achieve generalisation in DL models, a large amount of data is essential. In the case of speech processing and analysis, the available datasets are typically very limited [79]. Even for a very developed field of ASR, we have transcribed datasets for very few languages compared to the number of spoken languages worldwide. There are more than 5 000 spoken languages globally, however, only 389 languages are spoken by 94% of the world's population.[2] In speech processing, we do not even have speech datasets for 389 languages [79], [205]. Therefore, research in language and speech analysis research is facing the problem of data scarcity [81]. This imbalance, variation, diversity, and dynamics in speech and languages cause hurdles in designing speech-based healthcare systems. For example, the performance of ASR systems degrades when they are evaluated across different languages [81]. Therefore, we need to design more adaptive healthcare solutions using ASR trained on multiple languages data. The datasets related to speech disorders are also very few and have very small sizes [76], [206]. To solve this issue, techniques such as transfer learning, self-taught learning, etc., can be utilised to improve the generalisability of the models.

C. Interoperability Challenges

In current healthcare systems, the data generated from different medical devices, clinical reports, medical correspondence electronic health records (EHRs). These medical data are typically fragmented or isolated within hospitals and laboratories. The interoperability of these medical records is prohibited among different health services providers [86]. On the contrary, different EHRs, medical devices, and other IT systems are also not interoperable if data sharing is allowed among different hospitals [207]. For the effective utilisation of speech in healthcare, we need to enable interoperability in current healthcare systems. This would boost the speed of diagnostic procedures and provide a complete history of patients to medical practitioners. Therefore, the attention of researchers and other stakeholders working in the healthcare sector is required to find solutions for interoperability challenges.

D. Cultural and Language Barriers

Cultural and language diversity reduces the potential use of speech technology in digital healthcare. Digital health is not only a technological but cultural transformation. Cultural barrier is a major challenge for digital transformation, which becomes more prevalent in rural and developing areas. The transformation of people from the classical method to digital health is slowed down by ignoring the importance of cultural changes and the human factors [208]. Therefore, it is important to consider cultural barriers while designing healthcare solutions. In addition to cultural barriers, language diversity problem is another major problem. It is found that the linguistic differences among patients and medical doctors can cause patients to misinterpret medications and suffer unnecessary complications [86]. Therefore, healthcare solutions based speech technology must be trained on multiple languages to work effectively in such situations. However, the development of speech technology-based healthcare systems becomes more challenging for rarely spoken languages. Therefore, it is very important to consider language diversity while designing speech-based healthcare solutions.

VI. Open Issues and Future Directions

Speech technology is expected to drive the change in the healthcare system by changing the conventional ways of medical treatments. However, there are some open issues that require serious attention and consideration of the researchers. Therefore, in this section, we highlighted such open issues and with important pointers and future research directions for the research community.

A. Privacy and Ethical Concerns

When people use speech-based services such as speech recognition or voice authentication, they provide complete possession to their voice recordings to the respective device or software. In these services, speech can be used by an adversary or attacker to extract users' information such as speakers' identity, gender ethnicity information, and emotional state. The adversary can use this information for undesired purposes such as to fool voice-authentication systems. Similarly, users' speech can also be edited or used to create a fake speech that the person never spoke. Several other privacy-related concerns arise while using speech technology-based services [209]. In healthcare systems, information is more personal and very sensitive, and people are more vulnerable to the misuse of their data. Therefore, it is important to utilise speech processing in healthcare by considering both privacy and ethical concerns. In this regard, privacy-preserving DL algorithms can be utilised to protect speaker identity [210], [211], gender identity [212]. Similarly, federated learning [213] is another alternative solution to preserve users' privacy. In federated learning, training data remains decentralised using multiple participating devices.

B. Adoptability and Affordability Issues

Speech-based digital healthcare solutions are intended to be used by all types of users, including people with no literacy or education about smart devices. Usability of healthcare solutions effects adoption of various innovative digital healthcare products or services [85]. Therefore, it is crucial for developers to consider these issues while designing HCI interface for speech based healthcare solutions. It is also important to provide speech technology-based services at an affordable cost as it directly impacts on the adaptability of these solutions [214].

C. IoT Based Solutions

The paradigm of Internet of Things (IoT) offers unprecedented opportunities for digital healthcare solutions by providing an abstraction of infinite, physical, smart, and virtual objects.

[2][Online]. Available: https://www.ethnologue.com/statistics

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

These objects can capture, store, and securely transmit health related information to a public or private cloud and facilitate a new level of automation for the convenience of users. Most importantly, IoT-based solutions are very effective in terms of energy consumption, CPU and memory usage [215]. It is anticipated that IoT will disrupt the current healthcare systems by providing various cutting-edge and highly individualised digital healthcare solutions [5]. These solutions can be utilised for remote monitoring and diagnostics, chronic diseases management, elderly care, and much more. Therefore, it is very important to utilise the opportunities offered by IoT to enhance the effectiveness of speech-based healthcare services.

D. Potentials With 5G

Healthcare expenditure takes a big portion of the national budgets of various countries. For instance, roughly 18% of North America's and 10% of the global economy's gross domestic product (GDP) account for healthcare spending [216]. Technology-enabled healthcare solutions can provide health services outside the hospital setting at remote locations, which can promote adherence to medications, and reduce cost and readmission rates [5]. Telehealth services by utilising 4G LTE have shown sound economic benefits [217], which will be further fuelled by the increasing amount of smartphones and expeditiously improving connectivity with 5G network. It is anticipated that 5G will provide a consistent user experience not only in dense areas but also in remote locations. This will pave the path for telehealth services available everywhere. Therefore, it is important for researchers working on speech processing for health to consider the opportunities offered by communication technologies while designing healthcare solutions.

E. Sustainability Issues

Speech technology-based healthcare solutions have great potentials in healthcare and they are getting great interest and attention among industries and healthcare service providers. However, it is important to understand how any particular speech technology-based healthcare solution can attain a certain level of adoption to achieve scale. The sustainability of any new digital health solutions is always considered uncertain as they involve different various public and private stakeholders [85]. The sustainability of such projects and products become more uncertain in developing countries, where people avoid to use digital technology. Therefore, a proper collaboration is always needed among stakeholders to support the project to make a transition from pilot stage to a self-sustainable long-term project.

VII. Conclusion

Speech technology has unprecedented opportunities for the health domain and these potential opportunities can be reaped to fix the current healthcare system that is continuously facing an increasing burden of the ageing population and chronic diseases. In this paper, we highlighted the potentials of speech technology for healthcare and presented a state-of-the-art work on healthcare from different speech-related including automatic speech recognition (ASR), speech synthesis, and speech processing for different speech-related disorders. The reviewed literature showed that the research on speech processing for healthcare is rapidly evolving with very promising results. However, these results are mainly from some pilot projects or using some publicly available datasets and the available healthcare solutions based on speech technology are being used on a very small scale. There are various factors hindering the growth of speech technology in healthcare that we discussed in detail in this paper. Most importantly, we presented open issues and based on that we outlined future strategies for making speech technologies even more effective in healthcare, which include the utilisation of other emerging technologies like internet of things (IoT) and communication technologies like 5G.

References

[1] World Health Organization *et al.*, "WHO Expert Consultation on Rabies, 2nd Rep.," World Health Organization Tech. Rep. Ser., no. 982, p. 1, 2013.

[2] *The Global Burden of Chronic*. World Health Organization, Geneva, Switzerland, Accessed: Jan. 1, 2020. [Online]. Available: https://www.who.int/nutrition/topics/2_background/en/

[3] *Global Health Workforce Shortage to Reach 12.9 Million in Coming Decades*, World Health Organization: Geneva, Switzerland, 2014. [Online]. Available: http://www.who.int/mediacentre/news/releases/2013/health-workforce-shortage/en

[4] W. Savedoff, "A moving target: Universal access to healthcare services in latin america and the caribbean," Working paper//Inter-American Development Bank, Research Department, Washington, D.C., USA, Tech. Rep. 667, 2009.

[5] S. Latif, J. Qadir, S. Farooq, and M. Imran, "How 5G wireless (and concomitant technologies) will revolutionize healthcare?" *Future Internet*, vol. 9, no. 4, 2017.

[6] S. P. Cunningham *et al.*, "Cloud-based speech technology for assistive technology applications (cloudcast)," in *Proc. AAATE Conf.*, 2017, pp. 322–329.

[7] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: A focused review," *Frontiers Neuroscience*, vol. 10, 2016, Art. no. 429.

[8] M. Johnson *et al.*, "A systematic review of speech recognition technology in health care," *BMC Med. Informat. Decis. Making*, vol. 14, no. 1, 2014, Art. no. 94.

[9] S. Durling and J. Lumsden, "Speech recognition use in healthcare applications," in *Proc. 6th Int. Conf. Advances Mobile Comput. Multimedia*, 2008, pp. 473–478.

[10] E. Coiera, B. Kocaballi, J. Halamka, and L. Laranjo, "The digital scribe," *NPJ Digit. Medicine*, vol. 1, no. 1, pp. 1–5, 2018.

[11] T. Hodgson and E. Coiera, "Risks and benefits of speech recognition for clinical documentation: A systematic review," *J. Amer. Med. Informat. Assoc.*, vol. 23, no. e1, pp. e169–e179, 2016.

[12] L. Laranjo *et al.*, "Conversational agents in healthcare: A systematic review," *J. Amer. Med. Informat. Assoc.*, vol. 25, no. 9, pp. 1248–1258, 2018.

[13] S. V. Blackley, J. Huynh, L. Wang, Z. Korach, and L. Zhou, "Speech recognition for clinical documentation from 1990 to 2018: A systematic review," *J. Amer. Med. Informat. Assoc.*, vol. 26, no. 4, pp. 324–338, 2019.gg

[14] N. Cummins, A. Baird, and B. Schuller, "Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning," *Methods*, vol. 151, pp. 41–54, 2018.

[15] A. Keerio, B. K. Mitra, P. Birch, R. Young, and C. Chatwin, "On preprocessing of speech signals," *Int. J. Signal Process.*, vol. 5, no. 3, pp. 216–222, 2009.

[16] A. L. Higgins, S. F. Boll, and J. E. Porter, "Noise suppression and channel equalization preprocessor for speech and speaker recognizers: Method and apparatus," U.S. Patent 6 266 633, Jul. 24, 2001.

[17] H. Zen and H. Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4470–4474.

[18] U. D. Reichel and H. R. Pfitzinger, "Text preprocessing for speech synthesis," 2006.

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

[19] S. Arunachalam, D. Gould, E. Andersen, D. Byrd, and S. Narayanan, "Politeness and frustration language in child-machine interactions," in *Proc. 17th Eur. Conf. Speech Commun. Technol.*, 2001, pp. 2675–2678.

[20] C. Elliot, "The affective reasoner: A process model of emotions in a multi-agent system," M.S. thesis, Northwestern Univ. Inst. Learning Sciences, Northwestern, IL, USA, 1992.

[21] R. Cowie *et al.*, "What a neural net needs to know about emotion words," *Comput. Intell. Appl.*, vol. 404, pp. 5311–5316, 1999.

[22] A. Zolnay and R. Haeb-Umbach, "Acoustic feature combination for speech recognition," 2006.

[23] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech—A review," in *Toward Robotic Socially Believable Behaving Systems-Volume I*. Berlin, Germany: Springer, 2016, pp. 205–238.

[24] X. Zhang, Y. Sun, and S. Duan, "Progress in speech emotion recognition," in *Proc. IEEE Region 10 Conf.*, 2015, pp. 1–6.

[25] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2081–2090, Sep. 2011.

[26] A. König *et al.*, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's Dementia: Diagnosis, Assessment Disease Monit.*, vol. 1, no. 1, pp. 112–124, 2015.

[27] Y. Tahir *et al.*, "Non-verbal speech analysis of interviews with schizophrenic patients," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2016, pp. 5810–5814.

[28] P. Neammalai, S. Phimoltares, and C. Lursinsap, "Speech and music classification using hybrid form of spectrogram and fourier transformation," in *Proc. Asia-Pacific Signal Inform. Process. Assoc. Annu. Summit Conf.*, 2014, pp. 1–6.

[29] M. Srinivas, D. Roy, and C. K. Mohan, "Learning sparse dictionaries for music and speech classification," in *Proc. 19th Int. Conf. Digital. Signal. Process.*, 2014, pp. 673–675.

[30] G. Sell and P. Clark, "Music tonality features for speech/music discrimination," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 2489–2493.

[31] E. Mezghani, M. Charfeddine, C. B. Amar, and H. Nicolas, "Multifeature speech/music discrimination based on mid-term level statistics and supervised classifiers," in *Proc. IEEE/ACS 13th Int. Conf. Comput. Syst. Appl.*, 2016, pp. 1–8.

[32] D. Palaz, M. M. Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4295–4299.

[33] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Proc. Interspeech*, 2019, pp. 3920–3924.

[34] H. Muckenhirn, M. M. Doss, and S. Marcell, "Towards directly modeling raw speech signal for speaker verification using CNNs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4884–4888.

[35] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2013.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[37] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, Cambridge, MA, USA: MIT Press, 1998, pp. 255–258.

[38] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014.

[39] X. Li and X. Wu, "Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, 2015, pp. 4520–4524.

[40] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[41] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2014.

[42] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[43] K. Kumar *et al.*, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inform. Process. Syst.*, 2019, pp. 14 881–14 892.

[44] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[45] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[46] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[47] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 1096–1103.

[48] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *Proc. Interspeech*, 2018, pp. 257–261.

[49] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[50] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[51] S. Latif, M. Usman, R. Rana, and J. Qadir, "Phonocardiographic sensing using deep learning for abnormal heartbeat detection," *IEEE Sensors J.*, vol. 18, no. 22, pp. 9393–9400, Nov. 2018.

[52] A. Qayyum, S. Latif, and J. Qadir, "Quran reciter identification: A deep learning approach," in *Proc. 7th Int. Conf. Comput. Commun. Eng.*, 2018, pp. 492–497.

[53] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," in *Proc. Interspeech*, 2018, pp. 3107–3111.

[54] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, vol. 32, Paper II–1278.

[55] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4743–4751.

[56] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, vol. 48, pp. 1747–1756.

[57] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016, p. 165.

[58] A. van den Oord *et al.*, "Conditional image generation with pixelCNN decoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4790–4798.

[59] M. Schatz *et al.*, "Reliability and predictive validity of the asthma control test administered by telephone calls using speech recognition technology," *J. Allergy Clin. Immunol.*, vol. 119, no. 2, pp. 336–343, 2007.

[60] "Statistics on voice, speech, and language," The National Institute on Deafness and Other Communication Disorders (NIDCD), Jul. 2016. [Online]. Available: https://www.nidcd.nih.gov/health/statistics/statistics-voice-speech-and-language, Accessed: Dec. 14, 2019.

[61] "Deafness and hearing loss," World Health Organization, Geneva, Switzerland, 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss, Accessed: Dec. 14, 2019.

[62] O. Wendt, *Assistive Technology: Principles and Applications for Communication Disorders and Special Education*. Leiden, The Netherlands: Brill, 2011.

[63] O. Saz, S.-C. Yin, E. Lleida, R. Rose, C. Vaquero, and W. R. Rodríguez, "Tools and technologies for computer-aided speech and language therapy," *Speech Commun.*, vol. 51, no. 10, pp. 948–967, 2009.

[64] S.-A. Selouani, M. S. Yakoub, and D. O'Shaughnessy, "Alternative speech communication system for persons with severe speech disorders," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, 2009, Art. no. 540409.

[65] G. Potamianos and C. Neti, "Automatic speechreading of impaired speech," in *Proc. Int. Conf. Auditory-Visual Speech Process.*, 2001, pp. 177–182.

[66] R. Nayar, "Towards designing speech technology based assistive interfaces for children's speech therapy," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 609–613.

[67] I. Laaridh, W. Kheder, C. Fredouille, and C. Meunier, "Automatic prediction of speech evaluation metrics for dysarthric speech," in *Proc. Interspeech*, 2017, pp. 1834–1838.

[68] D. S. Wallace, "The role of speech recognition in clinical documentation," *Nuance Commun.*, 2018, Accessed: Dec. 14, 2019. [Online]. Available: https://www.hisa.org.au/slides/hic18/wed/SimonWallace.pdf

[69] M. Heinzer, "Essential elements of nursing notes and the transition to electronic health records," *J. Healthcare Inform. Manag.*, vol. 24, no. 4, pp. 53–59, 2010.

[70] J. Shagoury, "Dr. 'Multi-Task': Using speech to build up electronic medical records while caring for patients," in *Advances in Speech Recognition*. Berlin Germany: Springer, 2010, pp. 247–273.

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

[71] S. V. Blackley, V. D. Schubert, F. R. Goss, W. Al Assad, P. M. Garabedian, and L. Zhou, "Physician use of speech recognition versus typing in clinical documentation: A controlled observational study," *Int. J. Med. Informat.*, vol. 141, 2020, Art. no. 104178.

[72] Y. Wang, C. S. Jordan, K. P. Laby, and J. Southard, "Medical tele-robotic system with a head worn device," U.S. Patent 7 262 573, Aug. 28, 2007.

[73] M. Amiribesheli, A. Benmansour, and A. Bouchachia, "A review of smart homes in healthcare," *J. Ambient Intell. Humanized Comput.*, vol. 6, no. 4, pp. 495–517, 2015.

[74] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[75] P. Ekman *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion," *J. Personality Social Psychol.*, vol. 53, no. 4, pp. 4–712, 1987.

[76] R. Rana *et al.*, "Automated screening for distress: A perspective for the future," *Eur. J. Cancer Care*, vol. 28, no. 4, 2019, Art. no. e13033.

[77] L. E. Carlson and B. D. Bultz, "Cancer distress screening: Needs, models, and methods," *J. Psychosomatic Res.*, vol. 55, no. 5, pp. 403–409, 2003.

[78] J. A. Chiles, M. J. Lambert, and A. L. Hatch, "The impact of psychological interventions on medical cost offset: A meta-analytic review," *Clin. Psychology: Sci. Practice.*, vol. 6, no. 2, pp. 204–220, 1999.

[79] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Cross corpus speech emotion classification—An effective transfer learning technique," 2018, *arXiv:1801.06353*.

[80] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, 2015.

[81] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Deep representation learning in speech processing: Challenges, recent advances, and future trends," 2020, *arXiv:2001.00378*.

[82] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, 2016, pp. 35–42.

[83] M. A. Shahin, J. Epps, and B. Ahmed, "Automatic classification of lexical stress in english and arabic languages using deep learning," in *Proc. Interspeech*, 2016, pp. 175–179.

[84] S. P. Dubagunta, B. Vlasenko, and M. M. Doss, "Learning voice source related information for depression detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6525–6529.

[85] S. Latif *et al.*, "Mobile technologies for managing non-communicable diseases in developing countries," in *Proc. Mobile Appl. Solutions Social Inclusion*, 2018, pp. 261–287.

[86] S. Latif, R. Rana, J. Qadir, A. Ali, M. A. Imran, and M. S. Younis, "Mobile health in the developing world: Review of literature and lessons from a case study," *IEEE Access*, vol. 5, pp. 11 540–11 556, 2017.

[87] M. S. Wolf *et al.*, "To err is human: Patient misinterpretations of prescription drug label instructions," *Patient Educ. Counseling*, vol. 67, no. 3, pp. 293–300, 2007.

[88] D. W. Baker, M. S. Wolf, J. Feinglass, J. A. Thompson, J. A. Gazmararian, and J. Huang, "Health literacy and mortality among elderly persons," *Archives Internal Medicine*, vol. 167, no. 14, pp. 1503–1509, 2007.

[89] O. M. Oyelami, C. O. Uwadia, and N. A. Omoregbe, "Prospects of voice-enabled healthcare system in the developing nations," in *Proc. 1st Int. Conf. Mobile Comput., Wireless Commun., E-Health, M-Health, Telemed.*, LAUTECH, Ogbomoso, Nigeria, Nov. 18–20, 2008.

[90] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5934–5938.

[91] C.-C. Chiu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4774–4778.

[92] F. Boyer and J.-L. Rouas, "End-to-end speech recognition: A review for the french language," 2019, *arXiv:1910.08502*.

[93] A. Graves, "Sequence transduction with recurrent neural networks," in *Proc. Int. Conf. Mach. Learn. Workshop Representation Learn.*, 2012.

[94] N. Jaitly, Q. V. Le, O. Vinyals, I. Sutskever, D. Sussillo, and S. Bengio, "An online sequence-to-sequence model using partial conditioning," in *Proc. Advances Neural Inform. Process. Syst.*, 2016, pp. 5067–5075.

[95] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2016, pp. 4960–4964.

[96] H. Sak, M. Shannon, K. Rao, and F. Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping," in *Proc. Interspeech*, 2017, vol. 8, pp. 1298–1302.

[97] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *Proc. 34th Int. Conf. Mach. Learn.-Vol. 70*, 2017, pp. 2837–2846.

[98] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[99] S. Kriman *et al.*, "QuartzNet: Deep automatic speech recognition with 1D time-channel separable convolutions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6124–6128.

[100] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.

[101] J. Kahn, A. Lee, and A. Hannun, "Self-training for end-to-end speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 7084–7088.

[102] V. Pratap *et al.*, "Wav2Letter++: A fast open-source speech recognition system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6460–6464.

[103] Y. Wang *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6874–6878.

[104] D. S. Park *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[105] C. Lüscher *et al.*, "RWTH ASR systems for LibriSpeech: Hybrid vs attention," *Proc. Interspeech*, 2019, pp. 231–235.

[106] J.-Y. Hsu, Y.-J. Chen, and H.-Y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2020, pp. 7844–7848.

[107] M. J. Gales, K. M. Knill, A. Ragni, and S. P. Rath, "Speech recognition and keyword spotting for low-resource languages: Babel project research at cued," in *Proc. Spoken Lang. Technologies Under-Resourced Lang.*, 2014, pp. 16–23.

[108] W. Cheshire, "Doctors' handwriting gone digital: An ethical assessment of voice recognition technology in medicine," *Ethics Med.*, vol. 29, no. 2, pp. 71–77, Jun. 2013.

[109] M. Vogel, W. Kaisers, R. Wassmuth, and E. Mayatepek, "Analysis of documentation speed using web-based medical speech recognition technology: Randomized controlled trial," *J. Medical. Internet. Research.*, vol. 17, no. 11, 2015, Paper e247.

[110] C.-C. Chiu *et al.*, "Speech recognition for medical conversations," in *Proc. Interspeech*, 2018, pp. 2972–2976.

[111] C. Ahlgrim, O. Maenner, and M. W. Baumstark, "Introduction of digital speech recognition in a specialised outpatient department: A case study," *BMC Med. Informat. Decis. Making*, vol. 16, no. 1, pp. 1–8, 2016.

[112] R. Hoyt and A. Yoshihashi, "Lessons learned from implementation of voice recognition for documentation in the military electronic health record system," *Perspectives Health Inform. Management/AHIMA, Amer. Health Inform. Manag. Assoc.*, vol. 7, 2010.

[113] E. Edwards *et al.*, "Medical speech recognition: Reaching parity with humans," in *Proc. Int. Conf. Speech and Comput.*, 2017, pp. 512–524.

[114] T. F. de Toledo, H. D. Lee, N. Spolaor, C. S. R. Coy, and F. C. Wu, "Web system prototype based on speech recognition to construct medical reports in brazilian portuguese," *Int. J. Med. Informat.*, vol. 121, pp. 39–52, 2019.

[115] J. Du Toit, R. Hattingh, and R. Pitcher, "The accuracy of radiology speech recognition reports in a multilingual South African teaching hospital," *BMC Med. Imag.*, vol. 15, no. 1, p. 8, 2015.

[116] D. Suendermann-Oeft, S. Ghaffarzadegan, E. Edwards, W. Salloum, and M. Miller, "A system for automated extraction of clinical standard codes in spoken medical reports," in *Proc. Workshop SLT*, 2016.

[117] R. H. Strahan and M. E. Schneider-Kolsky, "Voice recognition versus transcriptionist: Error rates and productivity in MRI reporting," *J. Med. Imag. Radiation. Oncol.*, vol. 54, no. 5, pp. 411–414, 2010.

[118] T. Hodgson, F. Magrabi, and E. Coiera, "Evaluating the usability of speech recognition to create clinical documentation using a commercial electronic health record," *Int. J. Med. Informat.*, vol. 113, pp. 38–42, 2018.

[119] T. Hodgson, F. Magrabi, and E. Coiera, "Efficiency and safety of speech recognition for documentation in the electronic health record," *J. Amer. Med. Informat. Assoc.*, vol. 24, no. 6, pp. 1127–1133, 2017.

[120] A. Aronson and D. Bless, *Clinical Voice Disorders*. New York, NY, USA: Thieme, 2009.

[121] S.-H. Fang *et al.*, "Detection of pathological voice using cepstrum vectors: A deep learning approach," *J. Voice*, vol. 33, no. 5, pp. 634–641, 2019.

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

[122] H. Wu, J. J. Soraghan, A. Lowit, and G. Di Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," in *Proc. Interspeech*, 2018, vol. 2018, pp. 446–450.

[123] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: A preliminary study," in *Proc. Int. Conf. Workshop Bioinspired Intell.*, 2017, pp. 1–4.

[124] M. Alhussein and G. Muhammad, "Voice pathology detection using deep learning on mobile healthcare framework," *IEEE Access*, vol. 6, pp. 41 034–41 041, 2018.

[125] M. S. Hossain, G. Muhammad, and A. Alamri, "Smart healthcare monitoring: A voice pathology detection paradigm for smart cities," *Multimedia Syst.*, vol. 25, no. 5, pp. 565–575, 2019.

[126] Z. Ali, G. Muhammad, and M. F. Alhamid, "An automatic health monitoring system for patients suffering from voice complications in smart cities," *IEEE Access*, vol. 5, pp. 3900–3908, 2017.

[127] K. Lopez-de Ipina *et al.*, "Advances on automatic speech analysis for early detection of Alzheimer disease: A non-linear multitask approach," *Current Alzheimer Res.*, vol. 15, no. 2, pp. 139–148, 2018.

[128] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimers disease in narrative speech," *J. Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[129] A. H. Al-Fatlawi, M. H. Jabardi, and S. H. Ling, "Efficient diagnosis system for Parkinson's disease using deep belief network," in *Proc. IEEE Congr. Evol. Comput.*, 2016, pp. 1324–1330.

[130] A. Frid, A. Kantor, D. Svechin, and L. M. Manevitz, "Diagnosis of parkinson's disease from continuous speech using deep convolutional networks without manual selection of features," in *Proc. IEEE Int. Conf. Sci. Elect. Eng.*, 2016, pp. 1–4.

[131] J. Deng, N. Cummins, M. Schmitt, K. Qian, F. Ringeval, and B. Schuller, "Speech-based diagnosis of autism spectrum condition by generative adversarial network representations," in *Proc. Int. Conf. Digital. Health.*, 2017, pp. 53–57.

[132] S. Amiriparian *et al.*, "Recognition of echolalic autistic child vocalisations utilising convolutional recurrent neural networks," in *Proc. Interspeech*, 2018, pp. 2334–2338.

[133] T. She, X. Kang, S. Nishide, and F. Ren, "Improving LEO robot conversational ability via deep learning algorithms for children with autism," in *Proc. 5th IEEE Int. Conf. Cloud Comput. Intell. Syst.*, 2018, pp. 416–420.

[134] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in Adolescents' speech during family interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, Mar. 2010.

[135] S. Scherer, G. Stratou, J. Gratch, and L.-P. Morency, "Investigating voice quality as a speaker-independent indicator of depression and PTSD," in *Proc. Interspeech*, 2013, pp. 847–851.

[136] S. N. Chakravarthula, H. Li, S.-Y. Tseng, M. Reblin, and P. Georgiou, "Predicting behavior in cancer-afflicted patient and spouse interactions using speech and language," in *Proc. Interspeech*, 2019, pp. 3073–3077.

[137] A. R. Avila *et al.*, "Speech-based stress classification based on modulation spectral features and convolutional neural networks," in *Proc. 27th Eur. Signal Process. Conf.*, 2019, pp. 1–5.

[138] V. V. Yerigeri and L. Ragha, "Meta-heuristic approach in neural network for stress detection in marathi speech," *Int. J. Speech Technol.*, vol. 22, no. 4, pp. 937–957, 2019.

[139] Z. Huang, J. Epps, and D. Joachim, "Investigation of speech landmark patterns for depression detection," *IEEE Trans. Affective. Comput.*, to be published.

[140] M. S. Hossain, "Patient state recognition system for healthcare using speech and facial expressions," *J. Med. Syst.*, vol. 40, no. 12, pp. 1–8, 2016.

[141] K. Vatanparvar, V. Nathan, E. Nemati, M. M. Rahman, and J. Kuang, "A generative model for speech segmentation and obfuscation for remote health monitoring," in *Proc. IEEE 16th Int. Conf. Wearable Implantable Body Sensor Netw.*, 2019, pp. 1–4.

[142] M. R. P. Barbosa and F. D. M. Fernandes, "Remote speech-language intervention, with the participation of parents of children with autism," in *Advances Speech-language Pathology*, 2017.

[143] V. Rapcan, S. D'Arcy, N. Penard, I. H. Robertson, and R. B. Reilly, "The use of telephone speech recordings for assessment and monitoring of cognitive function in elderly people," in *Proc. 10th Annu. Conf. Int. Speech Commun. Assoc.*, 2009, pp. 943–947.

[144] M. Vacher, B. Lecouteux, and F. Portet, "Recognition of voice commands by multisource ASR and noise cancellation in a smart home environment," in *Proc. Proc. 20th Eur. Signal Process. Conf.*, 2012, pp. 1663–1667.

[145] M. S. Hossain and G. Muhammad, "Emotion-aware connected healthcare big data towards 5G," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2399–2406, 2017.

[146] M. Thenmozhi and K. Narmadha, "Privacy-enhanced emotion recognition approach for remote health advisory system," in *Artificial Intelligence and Evolutionary Computations in Engineering Systems*. Berlin, Germany: Springer, 2020, pp. 133–142.

[147] M. S. Hossain and G. Muhammad, "Emotion recognition using secure edge and cloud computing," *Inform. Sci.*, vol. 504, pp. 589–601, 2019.

[148] M. S. Hossain and G. Muhammad, "An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework," *IEEE Wireless Commun.*, vol. 26, no. 3, pp. 62–68, Jun. 2019.

[149] Y. Li, Y. Jiang, D. Tian, L. Hu, H. Lu, and Z. Yuan, "AI-enabled emotion communication," *IEEE Netw.*, vol. 33, no. 6, pp. 15–21, Nov./Dec. 2019.

[150] M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," *IEEE Access*, vol. 5, pp. 326–337, 2016.

[151] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, and Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.

[152] M. Morise, F. Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inform. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[153] H. Banno, H. Hata, M. Morise, T. Takahashi, T. Irino, and H. Kawahara, "Implementation of realtime straight speech manipulation system: Report on its first implementation," *Acoust. Sci. Technol.*, vol. 28, no. 3, pp. 140–146, 2007.

[154] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.

[155] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu, "FFTNet: A real-time speaker-dependent neural vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2251–2255.

[156] N. Kalchbrenner *et al.*, "Efficient neural audio synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2410–2419.

[157] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 3617–3621.

[158] R. Skerry-Ryan *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4693–4702.

[159] Y. Wang *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5180–5189.

[160] E. Battenberg *et al.*, "Effective use of variational embedding capacity in expressive end-to-end speech synthesis," 2019, *arXiv:1906.03402*.

[161] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," in *Proc. Interspeech*, 2016, pp. 2273–2277.

[162] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," *Proc. Interspeech*, 2017, pp. 4006–4010.

[163] X. Gonzalvo *et al.*, "Recent advances in Google real-time HMM-driven unit selection synthesizer," in *Proc. Interspeech*, 2016, pp. 2238–2242.

[164] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 6706–6713.

[165] Y. Li, A. S. Ng, T. Trinh, and R. McNamara, "Text-to-speech based reminder system," U.S. Patent 6 182 041, Jan. 30, 2001.

[166] K.-C. Liu, C.-H. Wu, S.-Y. Tseng, and Y.-T. Tsai, "Voice helper: A mobile assistive system for visually impaired persons," in *Proc. IEEE Int. Conf. Comput. Inform. Technol.*, 2015, pp. 1400–1405.

[167] A. S. Mondol, I. A. Emi, and J. A. Stankovic, "MedRem: An interactive medication reminder and tracking system on wrist devices," in *Proc. IEEE Wireless Health*, 2016, pp. 1–8.

[168] S. Samyoun, M. A. S. Mondol, I. A. Emi, and J. A. Stankovic, "iAdhere: A voice interactive assistant to improve adherence to medical treatments: Demo abstract," in *Proc. 10th ACM/IEEE Int. Conf. Cyber-Phys. Syst.*, 2019, pp. 334–335.

[169] A. Kumar and S. K. Agarwal, "Spoken web: Using voice as an accessibility tool for disadvantaged people in developing regions," *ACM SIGACCESS Accessibility Comput.*, no. 104, pp. 3–11, 2012.

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.

[170] S. Mhlana, "Development of isiXhosa text-to-speech modules to support e-services in marginalized rural areas," Ph.D. dissertation, Univ. of Fort Hare, Alice, South Africa, 2011.
[171] C. Henton, "Bitter pills to swallow: ASR and TTS have drug problems," *Int. J. Speech Technol.*, vol. 8, no. 3, pp. 247–257, 2005.
[172] S. Marshall and R. R. Hurtig, "Developing a culture of successful communication in acute care settings: Part I. Solving patient-specific issues," *Perspectives ASHA Special Interest Groups*, vol. 4, no. 5, pp. 1028–1036, 2019.
[173] Accessed: Jan. 5, 2020. [Online]. Available: https://www.nuance.com/en-au/healthcare/provider-solutions/speech-recognition/dragon-medical-practice-edition.html
[174] Accessed: Jan. 5, 2020. [Online]. Available: http://mdops.com/
[175] Accessed: Jan. 5, 2020. [Online]. Available: https://www.suki.ai/
[176] Accessed: Jan. 5, 2020. [Online]. Available: https://notablehealth.com/
[177] Accessed: Jan. 5, 2020. [Online]. Available: https://cuidahealth.com/
[178] Accessed: Jan. 5, 2020. [Online]. Available: https://elliq.com/
[179] Accessed: Jan. 5, 2020. [Online]. Available: https://lifepod.com/
[180] Accessed: Jan. 5, 2020. [Online]. Available: https://smpltec.com/reminder-rosie
[181] Accessed: Jan. 5, 2020. [Online]. Available: http://www.beyondverbal.com/
[182] Accessed: Jan. 5, 2020. [Online]. Available: https://www.cogitocorp.com/solutions/care-management/
[183] Accessed: Jan. 5, 2020. [Online]. Available: https://corti.ai/
[184] Accessed: Jan. 5, 2020. [Online]. Available: https://winterlightlabs.com/
[185] Accessed: Jan. 5, 2020. [Online]. Available: https://www.ava.me/
[186] Accessed: Jan. 5, 2020. [Online]. Available: https://vocalid.ai/
[187] Accessed: Jan. 5, 2020. [Online]. Available: http://www.voiceitt.com/
[188] Accessed: Jan. 5, 2020. [Online]. Available: https://cardiocube.com/
[189] Accessed: Jan. 5, 2020. [Online]. Available: https://www.careangel.com/
[190] Accessed: Jan. 5, 2020. [Online]. Available: https://www.sensely.com/
[191] Accessed: Jan. 5, 2020. [Online]. Available: https://medium.com/@HealthTap/dr-a-i-80b4cf06be30
[192] S. Latif, R. Rana, and J. Qadir, "Adversarial machine learning and speech emotion recognition: Utilizing generative adversarial networks for robustness," 2018, *arXiv:1811.11402*.
[193] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2015.
[194] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy*, 2016, pp. 372–387.
[195] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2574–2582.
[196] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.
[197] A. Qayyum, M. Usama, J. Qadir, and A. Al-Fuqaha, "Securing connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward," *IEEE Commun. Surv. Tut.*, vol. 22, no. 2, pp. 998–1026, 2020.
[198] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *Proc. IEEE Secur. Privacy Workshops*, 2018, pp. 1–7.
[199] A. Hannun *et al.*, "Deep speech: Scaling up end-to-end speech recognition," 2014, *arXiv:1412.5567*.
[200] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? Adversarial examples against automatic speech recognition," in *Proc. Conf. Neural Inf. Process. Syst. Mach. Deception Workshop*, 2018.
[201] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," 2018, *arXiv:1808.05665*.
[202] W. Cai, A. Doshi, and R. Valle, "Attacking speaker recognition with deep generative models," 2018, *arXiv:1801.02384*.
[203] S. Hu, X. Shang, Z. Qin, M. Li, Q. Wang, and C. Wang, "Adversarial examples for automatic speech recognition: Attacks and countermeasures," *IEEE Commun. Mag.*, vol. 57, no. 10, pp. 120–126, Oct. 2019.
[204] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.
[205] S. Latif, J. Qadir, and M. Bilal, "Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition," in *Proc. 8th Int. Conf. Affective. Comput. Intell. Interact.*, 2019, pp. 732–737.
[206] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Trans. Affective. Comput.*, to be published.
[207] P. J. Pronovost, *Procuring Interoperability: Achieving High-quality, Connected, and Person-Centered Care*. Washington, DC, USA: NAM. EDU, 2018.
[208] B. Meskó, Z. Drobni, É. Bényei, B. Gergely, and Z. Győrffy, "Digital health is a cultural transformation of traditional healthcare," *Mhealth*, vol. 3, pp. 38–46, 2017.
[209] M. A. Pathak, B. Raj, S. D. Rane, and P. Smaragdis, "Privacy-preserving speech processing: Cryptographic and string-matching frameworks show promise," *IEEE Signal Process. Mag.*, vol. 30, no. 2, pp. 62–74, Mar. 2013.
[210] B. M. L. Srivastava, A. Bellet, M. Tommasi, and E. Vincent, "Privacy-preserving adversarial representation learning in ASR: Reality or illusion?" *Proc. Interpspeech*, 2019, pp. 3700–3704.
[211] S. H. K. Parthasarathi, H. Bourlard, and D. Gatica-Perez, "Wordless sounds: Robust speaker diarization using privacy-preserving audio representations," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 1, pp. 85–98, Jan. 2012.
[212] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," 2019, *arXiv:1910.13212*.
[213] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 1310–1321.
[214] N. Ekekwe and N. Islam, *Disruptive Technologies, Innovation and Global Redesign: Emerging Implications*. Hershey, PA, USA: IGI Global, 2012.
[215] G. Aloi *et al.*, "A mobile multi-technology gateway to enable iot interoperability," in *Proc. IEEE 1st Int. Conf. IoT Des. Implementation*, 2016, pp. 259–264.
[216] "Global health care sector outlook," Deloitte, New York, NY, USA. Accessed: Jan. 1, 2020. [Online]. Available: https://www2.deloitte.com/global/en/pages/life-sciences-and-healthcare/articles/global-health-care-sector-outlook.html
[217] K. Taylor, *Connected Health: How Digital Technology is Transforming Health and Social Care*. London, U.K.: Deloitte Centre for Health Solutions, 2015.

Authorized licensed use limited to: National University Fast. Downloaded on April 30,2024 at 05:56:53 UTC from IEEE Xplore. Restrictions apply.