Digital Biomarkers

Tools and Devices - Review Article

Digit Biomark 2021;5:78–88
DOI: 10.1159/000515346

Received: January 5, 2021
Accepted: February 18, 2021
Published online: April 16, 2021

Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice

Guy Fagherazzi[a] Aurélie Fischer[a] Muhannad Ismael[b] Vladimir Despotovic[c]

[a]Deep Digital Phenotyping Research Unit, Department of Population Health, Luxembourg Institute of Health, Strassen, Luxembourg; [b]IT for Innovation in Services Department (ITIS), Luxembourg Institute of Science and Technology (LIST), Esch-sur-Alzette, Luxembourg; [c]Department of Computer Science, Faculty of Science, Technology and Medicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg

Keywords
Voice · Signal decomposition · Artificial intelligence · Vocal biomarker · COVID-19 · Smart home

Abstract
Diseases can affect organs such as the heart, lungs, brain, muscles, or vocal folds, which can then alter an individual's voice. Therefore, voice analysis using artificial intelligence opens new opportunities for healthcare. From using vocal biomarkers for diagnosis, risk prediction, and remote monitoring of various clinical outcomes and symptoms, we offer in this review an overview of the various applications of voice for health-related purposes. We discuss the potential of this rapidly evolving environment from a research, patient, and clinical perspective. We also discuss the key challenges to overcome in the near future for a substantial and efficient use of voice in healthcare.

Introduction

The human voice is a rich medium which serves as a primary source for communication between individuals. It is one of the most natural, energy-efficient ways of interacting with each other. The voice, as complex arrays of sound coming from our vocal cords, contains various information and plays a fundamental role for social interaction [1] by allowing us to share insights about our emotions, fears, feelings, and excitation by modulating its tone or pitch.

With the purpose of reaching a human-like level, the development of artificial intelligence (AI), technologies, and computer sciences has led the way to new opportunities for the field of digital health, the ultimate purpose of which is to ease the lives of people and healthcare professionals through the leverage of technologies. This is no difference regarding voice. Today, voice technology is even considered as one of the most promising sectors, with healthcare being predicted to be a dominant vertical in voice applications. By 2024, the global voice market is expected to represent up to USD 5,843.8 million [2].

Virtual/vocal assistants on smartphones or in smart home devices such as connected speakers are now mainstream and have opened the way for a considerable use of voice-controlled search. In 2019, 31% of smartphone users worldwide used voice tech at least once a week [3], and 20% of queries on Google's mobile app and Android devices were voice searches. If current voice searches are mostly restricted to basic questions, perspectives for rapid expansion in the healthcare sector are numerous. The evolution of voice technology, audio signal analysis, and natural language processing/understanding methods

karger@karger.com
www.karger.com/dib

Correspondence to:
Guy Fagherazzi, guy.fagherazzi@lih.lu

Table 1. Definitions of key concepts

Keyword	Definition	Example
Audio signal decomposition	Extraction and separation of features from raw audio signals	Decomposition using MFCC for audio feature extraction
Voice feature	One component of the voice audio signal (such as linguistic or acoustic features)	Voice pitch
Vocal biomarker	A feature (or a combination of features) in the voice that has been identified and validated as associated with a clinical outcome	Differentiate people with Parkinson's disease from healthy controls
Vocal assistant	A software agent that performs tasks based on vocal commands or questions	Use voice to manage medication, set up reminders, ask what medication to take at a given moment, and request a prescription refill

have opened the way to numerous potential applications of voice, such as the identification of vocal biomarkers for diagnosis, classification, or patient remote monitoring, or to enhance clinical practice [4].

In this review, we offer a comprehensive overview of all the present and future applications of voice for health-related purposes, whether it be from a research, patient, or clinical perspective. We also discuss the key challenges to overcome in the near future for a large, efficient, and ethical use of voice in healthcare (Table 1).

Search Strategy

References for this review were identified through searches of PubMed/Medline and Web of Science with search terms related to voice, vocal biomarker, voice signature, conversational agents, chatbot, and famous brands or vocal assistants (see the full list of keywords in online suppl. material 1; for all online suppl. material, see www.karger.com/doi/10.1159/000515346). The search was performed on December 26, 2020. Only articles, reviews, and editorials referring to studies in humans and published in English were finally considered. Articles were also identified through searches of the authors' own files and in the grey literature. The final reference list was generated on the basis of originality and relevance to the broad scope of this review.

Vocal Biomarkers

A biomarker is a factor objectively measured and evaluated which represents a biological or pathogenic process, or a pharmacological response to a therapeutic intervention [5], which can be used as a surrogate marker of a clinical endpoint [5]. In the context of voice, a vocal biomarker is a signature, a feature, or a combination of features from the audio signal of the voice that is associated with a clinical outcome and can be used to monitor patients, diagnose a condition, or grade the severity or the stages of a disease or for drug development [6]. It must have all the properties of a traditional biomarker, which are validated analytically, qualified using an evidentiary assessment, and utilized [7].

Parkinson's Disease

Work on vocal biomarkers have mainly been performed in the field of neurodegenerative disorders so far, on Parkinson's disease in particular, where voice disorders are very frequent (as high as 89% [8]) and where voice changes are expected to be utilized as an early diagnostic biomarker [9, 10] or marker of disease progression [11, 12], and could one day supplement the state-of-the art manual exam to assess symptoms to guide treatment initiation [9] or to monitor its efficacy [13]. These voice disorders are mostly related to phonation and articulation, including pitch variations, decreased energy in the higher parts of the harmonic spectrum, and imprecise articulation of vowels and consonants, leading to decreased intelligibility. Even though changes in voice are often overlooked by both patients and physicians in early stages of the disease, the objective measures show changes in voice features [14] in up to 78% of patients with early stage Parkinson's disease [15].

Alzheimer's Disease and Mild Cognitive Impairment

Subtle changes in voice and language can be observed years before the appearance of prodromal symptoms of Alzheimer's disease [16] and are also detected in early stages of mild cognitive impairment [17]. Both mild cognitive impairment and Alzheimer's disease are proven to

affect the verbal fluency, reflected by the patient's hesitation to speak and slow speech rate, or other impairments, such as word finding difficulties, leading to circumlocution and frequent use of filler sounds (e.g., uh, um), semantic errors, indefinite terms, revision, repetitions, neologisms, lexical and grammatical simplification, as well as loss of semantic abilities in general [18]. Discourse in Alzheimer's disease patients is characterized by reduced coherence, with implausible and irrelevant details [19]. Alterations have been also perceived in prosodic features (pitch variation and modulation, speech rhythm) and may affect the patient's emotional responsiveness [17, 20]. Voice features have the potential to become simple and noninvasive biomarkers for the early diagnosis of conditions associated with dementia [21].

Multiple Sclerosis and Rheumatoid Arthritis

Voice impairment and dysarthria are frequent comorbidities in people with multiple sclerosis [22]. It has also been suggested that voice characteristics and phonatory behaviors should be monitored in the long term to indicate the best window of time to initiate a treatment such as deep brain stimulation in people with multiple sclerosis [23]. Some voice features have already been identified as top candidates to monitor multiple sclerosis: articulation, respiration, and prosody [24]. In people with rheumatoid arthritis, pathological changes in the larynx occur with disease progression; therefore, tracking voice quality features has already been shown to be useful for patient monitoring [25].

Mental Health and Monitoring Emotions

Stress is an established risk factor of vocal symptoms. It was shown that smartphone-based self-assessed stress was correlated with voice features [26]. A positive correlation between stress levels and duration of verbal interaction [27] has also been reported. Voice symptoms seem more frequent in people with high levels of cortisol [28], which is common in patients with depression; therefore, voice characteristics are used to discover depression symptoms [29] or estimate depression severity. The second dimension of a Mel-Frequency Cepstrum Coefficient (MFCC) audio signal decomposition has been shown to discriminate depressive patients from controls [30]. An automated telephone system has been successfully tested to assess biologically based vocal acoustic measures of depression severity and treatment response [31] or to compute a post-traumatic stress disorder mental health score [32]. Beside acoustic measures, the linguistic aspects of voice are likely to be affected in mental diseases. Discourse tends to be incoherent in schizophrenia, manifested by disjointed flow of ideas, nonsensical associations between words, or digressions from the topic. Circumstantial speech is prominent in patients with bipolar and histrionic personality disorders [33]. Recent methodological developments have also allowed for improved emotion recognition accuracy [34], which enables sufficient maturity to be reached for medical research to monitor patients in between visits or to gather real-life information in clinical or epidemiological studies.

Cardiometabolic and Cardiovascular Diseases

A team from the Mayo Clinic has identified several vocal features associated with a history of coronary artery disease [35]. Regarding diabetes, only one study has studied vocal characteristics in people with and without type 2 diabetes showing differences between the 2 groups for many features (jitter, shimmer, smoothed amplitude perturbation quotient, noise to harmonic ratio, relative average perturbation, amplitude perturbation quotient [36]). It has been demonstrated that people with type 2 diabetes with poor glycemic control or with neuropathy had more straining, voice weakness, and a different voice grade [37], and that the most common type 2 diabetes phonatory symptoms were vocal tiring or fatigue and hoarseness [38].

COVID-19 and Other Conditions with Respiratory Symptoms

More recently, considerable research activity has emerged to use respiratory sounds (e.g., coughs, breathing, and voice) as primary sources of information in the context of the COVID-19 pandemic [39]. COVID-19 is a respiratory condition, affecting breathing and voice, and causing, among other symptoms, dry cough, sore throat, excessively breathy voice, and typical breathing patterns. These are all symptoms that can make patients' voices distinctive, creating recognizable voice signatures and enabling the training of algorithms to predict the presence of a SARS-COV-2 infection or as a tool to grade the severity of the disease. Results on vocal biomarkers to aid the diagnosis of COVID-19 by Cambridge University (Area Under the ROC Curve, AUC = 80%), or more recently by MIT scientists (AUC = 97%, based on cough recordings only) are promising [40]. Other projects based on cough sounds are ongoing [41] with the objective of developing a robot-based COVID-19 infection risk evaluation system. Future work should focus on the impact of the age category or the cultural background on the performances of cough-based algorithms, before launching such pre-screening tools on a large scale.

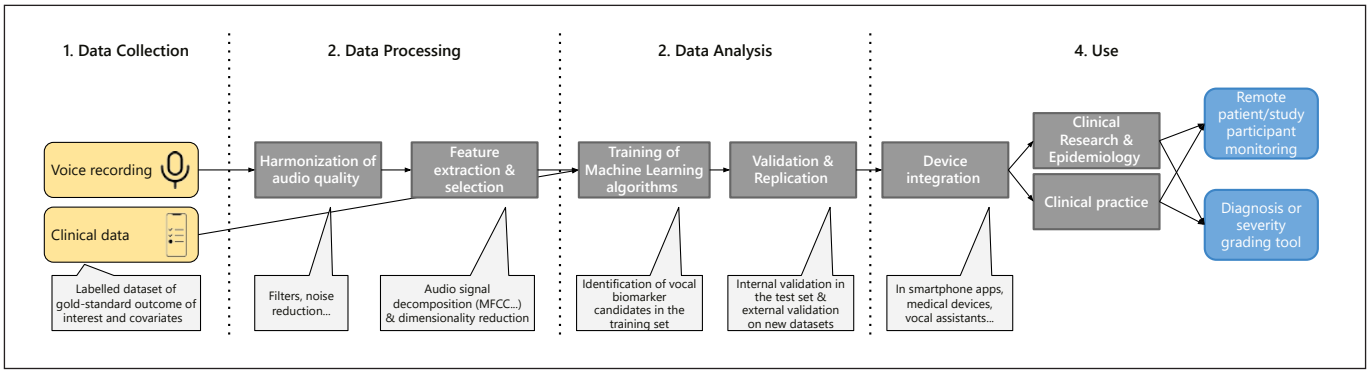

Fig. 1. Pipeline for vocal biomarker identification, from research to practice.

The Process to Identify a Vocal Biomarker

Below is a description of the typical approach to identify a vocal biomarker (Fig. 1).

Types of Voice Recordings

There is no standard protocol for voice recording to identify vocal biomarkers, but one can classify the sounds emitted from a human's mouth and analyze them for disease diagnostics into 3 main categories: verbal (isolated words, short sentence repetition, reading passage, running speech), vowel/syllable (sustained vowel phonation, diadochokinetic task), and nonverbal vocalizations (coughing, breathing). In a paper from the Mayo Clinic, study participants were asked to perform three 30-s separate voice recordings [35]: read a prespecified text, describe a positive emotional experience, and describe a negative emotional experience. There is an ongoing debate on the efficiency of use of isolated words or text, that are read aloud, and spontaneous conversational speech recordings [15, 42]. In order to have control over the recorded vocal task, but to allow patients to choose their own words to preserve the naturalness, semi-spontaneous voice tasks are designed where the patient is instructed to talk about a particular topic (e.g., picture description or story narration task). Sustained vowel phonations are another common type of recording, where participants are requested to sustain voicing of a vowel for as long and as steadily as they can. Sustained vowel phonations carry information for evaluating dysphonia, and enable estimating a patient's voice without articulatory influences, unaffected by speaking rate, stress, or intonation, and less influenced by the dialect of the speaker [43]. This is particularly helpful for multilingual analyses [44], to avoid confusion caused by different languages or accents. Diadochokinetic tasks are frequently used for the determination of articulatory impairment and include fast repetition of syllables, which combine plosives and vowels (e.g., /pa/-/ta/-/ka/). This task requires rapid movements of the lips, tongue, and soft palate, and reveals the patient's ability to retain their speech rate and/or intelligibility [45].

Sustained vowels and diadochokinetic tasks provide a greater level of control in comparison to conversational speech since they have reduced psychoacoustic complexity with less variability in vocal amplitude, frequency, and quality. However, voice performance is altered to a greater extent in spontaneous speech than in controlled tasks [46]. For example, voice disruptions and voice quality fluctuations are much more evident in conversational speech [43]. It better elicits the dynamic attributes of voice and varying voice patterns that occur in daily voice use, but the feature extraction is more difficult. Thus, the choice of a type of voice recording also depends on the objective: is it primarily diagnostic or developing a more comprehensive understanding of voice disorder.

Data Collection Techniques

Different data collection techniques have been developed over the past decades. They can be grouped into 4 main categories:

1. Studio-based recording includes speech recording into a controlled environment which leads to reduced unwanted acoustics and avoid proximity effects. This often induces an exaggeration of low-frequency sounds due to the proximity of the sound source from a microphone. In general, the recommended distance is between 15 and 30 cm. The collected data via this technique are in general not suitable for a speech application environment.

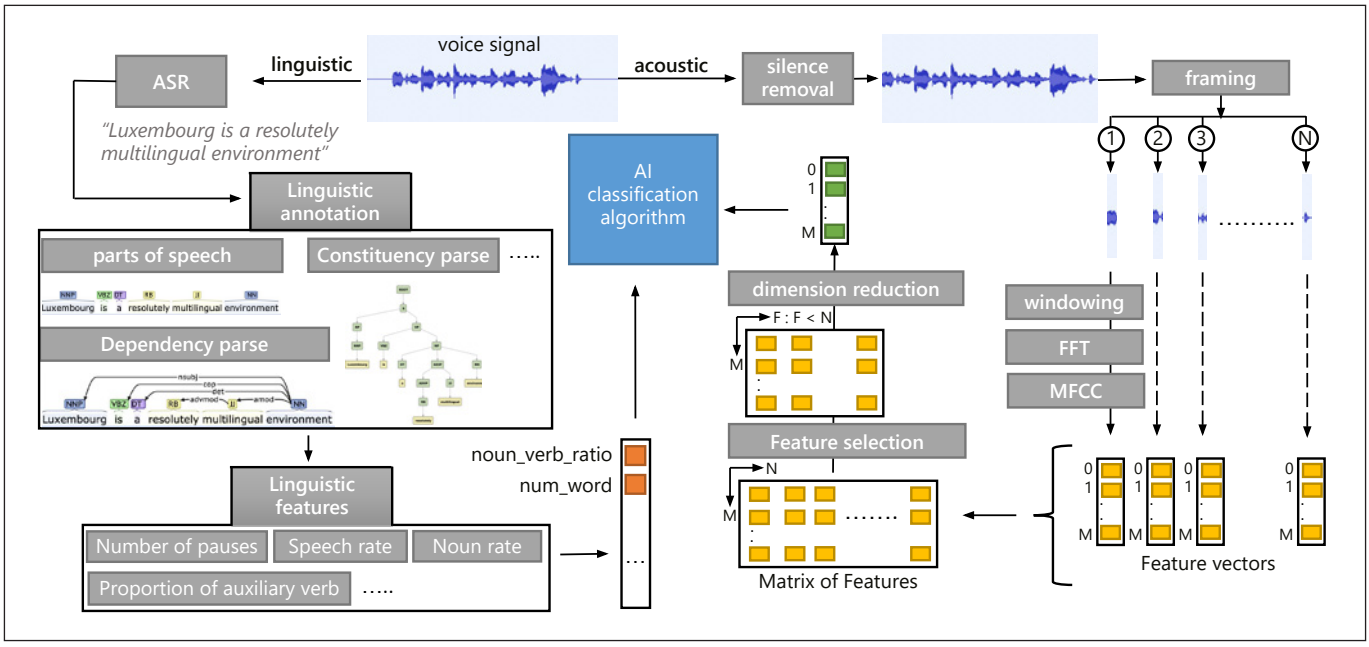

Fig. 2. Representation of a typical voice signal pre-processing and feature extraction using MFCCs. Representation of a typical voice signal pre-processing and linguistic and acoustic feature extraction. Voice signal represents the sound of the following sentence (e.g., "Luxembourg is a resolutely multilingual environment"). ASR refers to automatic speech recognition. Linguistic annotation includes part-of-speech, dependency and constituency parses, and sense tagging. In this diagram, linguistic annotation is applied using tools like CoreNLP. The number of pauses, speech rate, and noun rate are linguistic features and extracted using the BlaBla package, which is a clinical linguistic feature extraction tool. Acoustic features are extracted using MFCCs. The framing step refers to a signal segmentation into *N* samples. Windowing is multiplying of the signal sample by a window function like Hamming to minimize discontinuous signals that can cause noise in the subsequent fast Fourier transform (FFT) step. In this diagram, dimension reduction is represented by the principal component analysis (PCA) method, reducing feature space to a one-dimensional vector.

2. Telephone-based recording which requires data collection from a variety of speakers and handsets where several disadvantages, such as handset noise, a lack of control over the speaker's environment, and bandwidth limitations, are frequent.
3. Web-based recording is a very popular technique for large-scale data collection campaigns and relies on internet access, which is becoming readily available.
4. Smartphone-based recording provides broadband quality using smartphone devices, which are becoming widely available and at a low cost. Smartphone/web-based recording has the same potential drawbacks of telephone-based recording apart from the bandwidth limitation.

A pre-processing step is therefore necessary to overcome most of these limitations.

Audio Pre-Processing

A first step before analyzing the data is the audio pre-processing. This includes steps such as resampling, normalization, noise reduction, framing, and windowing the data [47], as described in Figure 2. The normalization step improves the performance of feature detection by reducing the amount of different information without distorting differences in the ranges of values. Moreover, in traditional non-machine-learning-based approaches for noise detection and reduction, a clean voice estimation is obtained by passing the noisy voice through a linear filter. However, many recent methods work to define mapping functions between clean and noisy voice signals using neural networks. The framing step consists of dividing the voice signal into a number of samples. These are multiplied by a window function to reduce signal leakage effects, which are the discontinuous signals that can cause noise in the subsequent fast Fourier transform. Once these steps have been performed, feature extraction can start.

Audio Feature Extraction

Prior to data analysis, there is a need to convert the audio signal into "features," meaning the most dominating and discriminating characteristics of a signal which

will later contribute to training machine learning algorithms [48]. Various methods are proposed in the literature to identify acoustic features from the temporal, frequency, cepstral, wavelet, and time-frequency domains [48]. The prosodic (pitch, formants, energy, jitter, shimmer) or spectral characteristics (spectral flux, slope, centroid, entropy, roll-off, and flatness), voice quality (zero-crossing rate, harmonic-to-noise ratio, noise-to-harmonic ratio), or phonation (fundamental frequency, pitch period entropy) [49] parameters can be extracted and analyzed. Nonlinear dynamic features, such as correlation dimension, fractal dimension, recurrence period density entropy, or Lempel-Ziv complexity, are able to describe the generation of nonlinear aerodynamic phenomena during voice production. Segmental features, such as MFCCs, may be the most frequently used in speech analysis [35], followed by perceptual linear prediction coefficients, and linear frequency cepstral coefficients [34]. Usually, the first 8–13 MFCC coefficients are sufficient to represent the shape of the spectrum even if some applications need a higher order to capture tone information.

Contrary to acoustic features which are able to capture the motor speech impairments, cognitive impairments may require analyzing linguistic features which reflect the parts of speech, vocabulary diversity, lexical and grammatical complexity, syntactic structures, semantic skills, and sentiment [4]. Before starting linguistics feature extraction and analyzing, linguistic annotation is a necessary step to define the sentence boundaries, parts of speech, named entities, numeric and time values, dependency, and constituency parses. Linguistic analyses often require extended speech production to extract features at all linguistic levels: phonetic and phonological (number of pauses, total pause time, hesitation ratio, speech rate), lexico-semantic (average rate of occurrence for each part of speech, number of repetitions, semantic errors, and closed-class word errors), morphosyntactic and syntactic (number of words per clause, number of dependent and simple clauses, number of clauses per utterance, mean length of utterances), and discourse-pragmatic (cohesion, coherence [19]).

The correct choice of features heavily depends on the voice disorder, disease, and type of voice recording. For example, acoustic features extracted from sustained vowel phonations or diadochokinetic recordings are common in the detection of Parkinson's disease, whereas linguistic features extracted from spontaneous or semi-spontaneous speech may be a more appropriate choice for the estimation of Alzheimer's disease or mental health disorders.

Audio Feature Selection and Dimensionality Reduction

Feature selection methods such as the mRMR (minimum redundancy maximum relevance) [50], Gram-Schmidt orthogonalization [44] allow a subset of the original feature set to be selected without changing them, as illustrated in Figure 2. It removes highly correlated features as well as features with missing values or low variance. This helps to select, for a given outcome of interest, the most relevant set of features to consider for the prediction or classification task. Besides, to avoid a "curse of dimensionality," dimensionality reduction methods such as principal component analysis, linear discriminant analysis, random forests, or stochastic neighbor embedding can be used to transform features and perform data visualization [51].

Training of Algorithms

Following the selection of features, machine or deep learning algorithms, such as support vector machines, hidden Markov models, convolutional or recurrent neural networks, just to name a few, can be trained to automatically predict or classify any clinical, medical, or epidemiological outcome of interest, from vocal features alone or in combination with other health-related data [47]. Algorithms are usually trained on one dataset and then tested on a separate dataset. External validation is still rare in the literature, mainly due to a lack of available data. Although supervised learning algorithms are commonly used as predictive models, extracting the implicit structures and patterns from the voice data using unsupervised learning techniques is also possible. Transfer learning is another promising approach which benefits from pre-training the model on a large voice dataset in a different domain where data are easier to collect, and fine-tuning the model in a target voice dataset, which is typically much smaller.

Testing of Algorithms

Collection of large-scale datasets for people with voice impairments is rarely feasible; therefore, in order to have reliable estimates of the performance, cross-validation and out-of-bootstrap validation techniques can be used. In cross-validation the dataset is randomly partitioned into k approximately equally sized subsets (folds), one being used for testing and the remaining ones for training. The performance is averaged over all folds. Leave-one-out cross-validation is an extreme case of cross-validation when the number of folds is equal to the number of data instances, meaning that the model is trained on all data

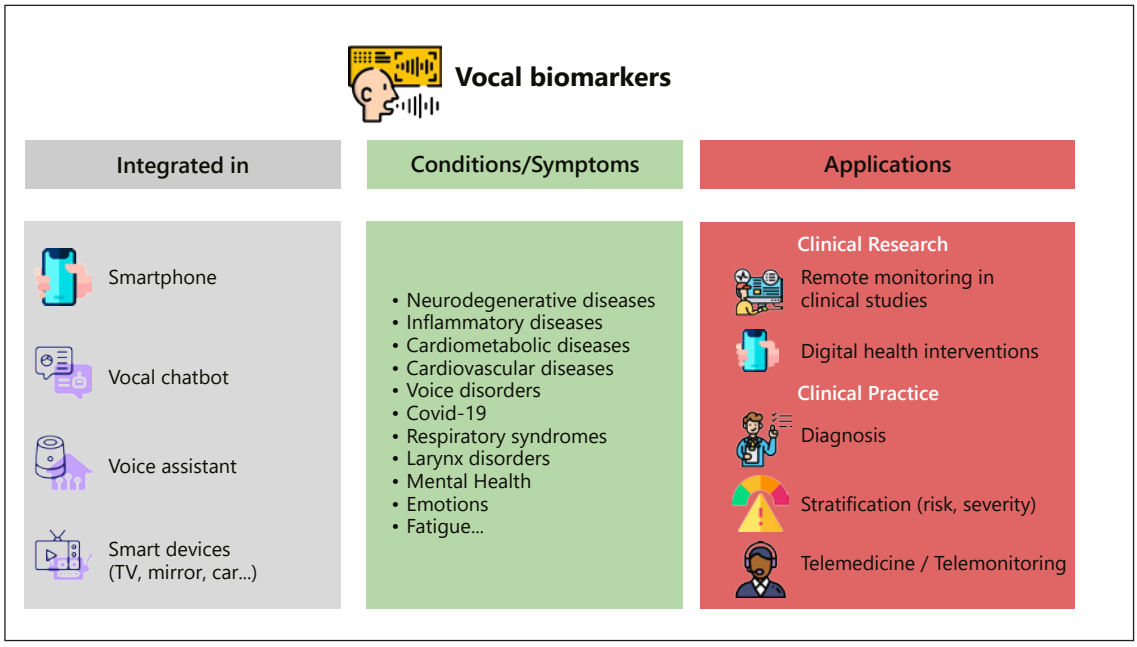

Fig. 3. Overview of present and future use of vocal biomarkers for health.

except one data instance. In bootstrap validation data instances are sampled with replacement from the original dataset, thus producing the surrogate datasets of the same size that may contain repeated data instances or miss data instances from the original dataset. If the unsampled data instances are used for testing, the method is called out-of-bootstrap validation.

Performance Metrics

Various performance metrics are used depending on specific application and the dataset, including accuracy, specificity, sensitivity (recall), precision, F measure, and AUC, just to name a few. The right choice of the metrics is very important since it guides the selection of the prediction model, but also affects interpretation of the results. For example, using accuracy for a heavily imbalanced classification problem could be misleading, since high performance can be reached by a model that always predicts the majority class. Sensitivity-specificity and precision-recall metrics are better choices in that case.

From Research to Clinical Practice

Once a vocal biomarker has been identified, as with any biomarker, the path is still long to a clinical routine use. For vocal biomarkers there are additional challenges, as their validity may be restricted to some languages or accents. The US Food and Drug Administration or European Medicines Agency have not approved any vocal biomarkers yet. Therefore, we can only speculate on the theoretical framework of such a process in the future, taking into account close cases in traditional biomarkers [7] and challenges in digital health. The first step would be to develop standards for vocal biomarker collection and create large-scale voice sample repositories for clinical use. This should be followed by integrating the algorithm into a user-friendly device (smartphone app, smart home device, connected medical device, etc.), co-designed with the end-users if possible. It should then enter sequentially into a feasibility study, one or several clinical trials, as well as real-world studies. It will not be the algorithm alone but its embedding in a connected medical device which will be approved by the agencies, and this major step has not been taken yet. Besides, given the technical constraints, we suspect that the first vocal biomarkers to be validated will be restricted to a specific language or a specific sub-group of the population. A relevant template to help standardizing and evaluating speech-based digital biomarkers has recently been proposed [4]. Health check-ups could one day be performed directly on an everyday device such as a smart mirror to track digital biomarkers, including vocal biomarkers, activity, healthcare status, and body movement [52]. For seniors, voice can also be a preferred medium to communicate inside a smart home to exchange with remote family members, in case of an

Table 2. Technical and ethical challenges for the field of voice technology to move from research to clinical practice

Challenges		Type of studies needed
technical	ethical	
Building and sharing large databanks of highly qualified audio recordings with clinical data and identifying key vocal biomarker candidates	Secure data collection and storage, rely on high-quality, gold-standard clinical data to train algorithms. Transparent definition of the types and frequency of data collected. Privacy preservation and protection of personal data. Article 4.1 of the General Data Protection Regulation of the European Union (GDPR EU) considers the voice as non-anonymous data	Proof of concept studies
Increase audio data harmonization and standardization across studies	Ensure high variability in the profiles to avoid systemic biases	Replication studies
Move from language-, accent-, age-, and culture-specific vocal biomarkers to more universal ones	Maximize open data and open source initiatives to ensure transparency, cross-comparison, and interoperability	
Improve algorithm accuracy	Increase algorithmic explainability	
Embed algorithms into medical devices (apps, vocal assistants, smart mirrors...) and prototyping		Qualitative studies and co-design sessions with end-users
		Usability and pilot studies
Integration within existing IT or telehealth systems	Do not increase existing digital divides and ensure a universal access to innovation	Clinical utility evaluation (randomized controlled trials, marker-based strategy-designed trials) and real-world evaluation studies

emergency or for telemedicine [53, 54]. In pilot studies, it has been shown that it is overall well accepted but highly dependent on the task complexity and the cognitive abilities of the individuals [55].

Future of Voice for Health

In this review, we have summarized the main fields of use today and in the coming years. Soon, the field will likely move from audio only to video; adding images to the voice will help to better characterize patients, including their emotions or other health characteristics from facial recognition, which, in combination with vocal biomarkers, will ease the remote monitoring of health [56–61]. The increase in data transfer capabilities, using the 5G networks and future updates, combined with an increasing proportion of the population with a smartphone equipped with a vocal assistant or at-home devices, will ease the collection and processing of large vocal samples in raw format or high definition [62]. From a research point of view, we can expect further inclusion of voice-related secondary endpoints in trials and real-world studies. From a healthcare point of view, the inclusion of voice analysis in health call centers will enable augmented consultations, a more accurate authentication of the caller, and real-time analysis of health-related features. Voice technologies will soon be further integrated into the development of virtual doctors and virtual/digital clinics [63] (Fig. 3).

Ethical and Technological Challenges to Tackle

Voice technologies and vocal biomarkers have to take the language and accent into account before being used on a large scale, otherwise they may increase systemic biases towards people from specific regions, backgrounds, or with a specific accent, and could increase a pre-existing digital and socioeconomic divide already present in some minorities (Table 2). To that extent, the voice technology field can learn from other fields, such as radiology for which the use of AI is much more advanced and where systemic biases have already been documented [64]. On top of that, some voice-specific issues will have to be dealt with, as for many applications of vocal biomarkers it is

likely that language-, accent-, age-, or culture-specific features are identified first, before moving to more universal, language- and accent-independent features. The right balance will have to be found between hyper-personalization for a given user and universal assessment of the clinical benefit of a vocal biomarker. There is also a need to improve natural language processing and understanding capabilities, relevance, and the accuracy of answers of vocal assistants, increase the fluidity in human-vocal chatbots interaction, and include emotions and empathy in the dialogue, if we ever want to reach massive and long-term adoption.

The validation of vocal biomarkers against gold-standards is mandatory for a safe use of voice to monitor health-related outcomes. Too few studies are available yet to enable a switch from novelty in small feasibility studies to large-scale clinical development [65].

One now needs proper evaluation of usability, adaptability, efficacy, and safety, but also sociological and ethical implications of using vocal biomarkers and voice technologies. The question of interoperability with existing technologies, integration within the various health systems, and long-term business models remains to be solved. Gathering more data is required to make reliable estimates; therefore, we strongly recommend the establishment of large data banks of labelled audio datasets with associated clinical outcomes. The next step will be to embed the algorithms in a digital device (should it be a vocal assistant, a smartphone, or a smart mirror [52]) and run prospective randomized controlled trials, real-world evaluation, and qualitative studies before envisaging a scale-up. The field needs to move towards a standardization of vocal biomarker collection in terms of data and formats to work with, to ensure cross-comparisons, compatibility, and transferability. Sharing data is also needed, as it will ensure the development of more accurate vocal biomarkers and voice technologies. As any field impacted by AI, voice technologies or vocal biomarkers need to rely on algorithms trained on diverse datasets to limit biases towards under-represented groups of the population.

Voice data is considered sensitive as it can be used to reveal the person's identity, demographic or ethnic origin, or in cases of vocal biomarkers also the health status. Measures, such as encrypting voice data, splitting data into random components, each of them independently processed to securely process voice data without privacy leakage, or learning data representations from which sensitive identifiable information is removed, just to name a few, should be used to address ethical concerns related to voice data collection and processing.

Conclusion

We have discussed numerous applications in healthcare, both for patients and for healthcare professionals. It becomes clear that voice will be increasingly used in future health systems: vocal biomarkers will track key health parameters remotely and will be used for deep phenotyping patients or designing innovative trials, opening the way to precision medicine [9], while voice technologies will be integrated into clinical practice to ease the lives of both patients and healthcare professionals. For the field to reach maturity, we need to move from a technology-oriented approach to a more health-oriented one, by creating studies and high-value datasets for providing evidence of the benefits of such an approach.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

G.F. and A.F. are supported by the Luxembourg Institute of Health, Luxembourg National Research Fund (FNR; Predi-COVID, grant No. 14716273), and the André Losch Foundation. M.I. and V.D. are supported by Luxembourg Institute of Science and Technology (LIST) and University of Luxembourg (UL), respectively, as well as by FNR (CDCVA, grant No. 14856868).

Author Contributions

All authors designed the study and drafted the first version, critically revised, and approved the final version of the manuscript.

References

1 Grossmann T, Vaish A, Franz J, Schroeder R, Stoneking M, Friederici AD. Emotional voice processing: investigating the role of genetic variation in the serotonin transporter across development. PLoS One. 2013; 8:e68377.

2 VynZ Research. Voice assistant market. [cited 15 Feb 2021]. Available from: https://www.vynzresearch.com/.

3 VynZ Research. Global voice assistant market is set to reach USD 5,843.8 million by 2024. Globenewswire.com [Internet]. Cited 17 Mar 2020. Available from: https://www.globenewswire.com/news-release/2020/01/28/1976318/0/en/Global-Voice-Assistant-Market-is-Set-to-Reach-USD-5-843-8-million-by-2024-Observing-a-CAGR-of-27-7-during-2019-2024-VynZ-Research.html.

4 Robin J, Harrison JE, Kaufman LD, Rudzicz F, Simpson W, Yancheva M. Evaluation of Speech-Based Digital Biomarkers: review and Recommendations. Digit Biomark. 2020 Oct; 4(3):99–108.
5 Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001 Mar;69(3):89–95.
6 Kraus VB. Biomarkers as drug development tools: discovery, validation, qualification and use. Nat Rev Rheumatol. 2018 Jun;14(6):354–62.
7 Califf RM. Biomarker definitions and their applications. Exp Biol Med. 2018 Feb;243(3): 213–21.
8 Dashtipour K, Tafreshi A, Lee J, Crawley B. Speech disorders in Parkinson's disease: pathophysiology, medical management and surgical approaches. Neurodegener Dis Manag. 2018 Oct;8(5):337–48.
9 Tracy JM, Özkanca Y, Atkins DC, Hosseini Ghomi R. Investigating voice as a biomarker: deep phenotyping methods for early detection of Parkinson's disease. J Biomed Inform. 2020 Apr;104:103362.
10 Arora S, Visanji NP, Mestre TA, Tsanas A, Al-Dakheel A, Connolly BS, et al. Investigating voice as a biomarker for leucine-rich repeat kinase 2-associated Parkinson's disease. J Parkinsons Dis. 2018;8(4):503–10.
11 Ma A, Lau KK, Thyagarajan D. Voice changes in Parkinson's disease: what are they telling us? J Clin Neurosci. 2020 Feb;72:1–7.
12 Zhan A, Mohan S, Tarolli C, Schneider RB, Adams JL, Sharma S, et al. Using smartphones and machine learning to quantify Parkinson disease severity: the Mobile Parkinson Disease Score. JAMA Neurol. 2018 Jul;75(7): 876–80.
13 Cushnie-Sparrow D, Adams S, Abeyesekera A, Pieterman M, Gilmore G, Jog M. Voice quality severity and responsiveness to levodopa in Parkinson's disease. J Commun Disord. 2018 Nov - Dec;76:1–10.
14 Tsanas A, Little MA, McSharry PE, Ramig LO. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Trans Biomed Eng. 2010 Apr; 57(4):884–93.
15 Rudzicz F. Articulatory Knowledge in the Recognition of Dysarthric Speech. IEEE Trans Audio Speech Lang Process. 2011; 19(4):947–60.
16 Ahmed S, Haigh AM, de Jager CA, Garrard P. Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease. Brain. 2013 Dec;136(Pt 12):3727–37.
17 Toth L, Hoffmann I, Gosztolya G, Vincze V, Szatloczki G, Banreti Z, et al. A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. Curr Alzheimer Res. 2018;15(2):130–8.
18 Reilly J, Peelle JE, Antonucci SM, Grossman M. Anomia as a marker of distinct semantic memory impairments in Alzheimer's disease and semantic dementia. Neuropsychology. 2011 Jul;25(4):413–26.
19 Boschi V, Catricalà E, Consonni M, Chesi C, Moro A, Cappa SF. Connected Speech in Neurodegenerative Language Disorders: A Review. Front Psychol. 2017 Mar;8:269.
20 Martínez-Sánchez F, Meilán JJ, Carro J, Ivanova O. A Prototype for the Voice Analysis Diagnosis of Alzheimer's Disease. J Alzheimers Dis. 2018;64(2):473–81.
21 König A, Satt A, Sorin A, Hoory R, Toledo-Ronen O, Derreumaux A, et al. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimers Dement. 2015 Mar;1(1):112–24.
22 Rusz J, Benova B, Ruzickova H, Novotny M, Tykalova T, Hlavnicka J, et al. Characteristics of motor speech phenotypes in multiple sclerosis. Mult Scler Relat Disord. 2018 Jan;19: 62–9.
23 Pützer M, Wokurek W, Moringlane JR. Evaluation of Phonatory Behavior and Voice Quality in Patients with Multiple Sclerosis Treated with Deep Brain Stimulation. J Voice. 2017 Jul;31(4):483–9.
24 Noffs G, Perera T, Kolbe SC, Shanahan CJ, Boonstra FM, Evans A, et al. What speech can tell us: A systematic review of dysarthria characteristics in Multiple Sclerosis. Autoimmun Rev. 2018 Dec;17(12):1202–9.
25 Kosztyła-Hojna B, Moskal D, Kuryliszyn-Moskal A. Parameters of the assessment of voice quality and clinical manifestation of rheumatoid arthritis. Adv Med Sci. 2015 Sep; 60(2):321–8.
26 Adams P, Rabbi M, Rahman T, Matthews M, Voida A, Gay G, et al. Towards Personal Stress Informatics: Comparing Minimally Invasive Techniques for Measuring Daily Stress in the Wild. Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare. 2014. doi: https://doi.org/10.4108/icst.pervasivehealth.2014.254959.
27 Ferdous R, Osmani V, Beltran Marquez J, Mayora O. Investigating correlation between verbal interactions and perceived stress. Annu Int Conf IEEE Eng Med Biol Soc. 2015 Aug;2015:1612–5.
28 Holmqvist-Jämsén S, Johansson A, Santtila P, Westberg L, von der Pahlen B, Simberg S. Investigating the Role of Salivary Cortisol on Vocal Symptoms. J Speech Lang Hear Res. 2017 Oct;60(10):2781–91.
29 Zhang L, Duvvuri R, Chandra KK, Nguyen T, Ghomi RH. Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative. Depress Anxiety. 2020 Jul;37(7):657–69.
30 Taguchi T, Tachikawa H, Nemoto K, Suzuki M, Nagano T, Tachibana R, et al. Major depressive disorder discrimination using vocal acoustic features. J Affect Disord. 2018 Jan; 225:214–20.
31 Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry. 2012 Oct;72(7):580–7.
32 Xu R, Mei G, Zhang G, Gao P, Judkins T, Cannizzaro M, et al. A voice-based automated system for PTSD screening and monitoring. Stud Health Technol Inform. 2012;173:552–8.
33 Cohen AS, Elvevåg B. Automated computerized analysis of speech in psychiatric disorders. Curr Opin Psychiatry. 2014 May;27(3): 203–9.
34 Daneshfar F, Kabudian SJ. Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. Multimedia Tools Appl. 2020;79(1-2): 1261–89.
35 Maor E, Sara JD, Orbelo DM, Lerman LO, Levanon Y, Lerman A. Voice Signal Characteristics Are Independently Associated With Coronary Artery Disease. Mayo Clin Proc. 2018 Jul;93(7):840–7.
36 Chitkara D, Sharma RK. Voice based detection of type 2 diabetes mellitus. 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). 2016. doi: https://doi.org/10.1109/AEEICB.2016.7538402.
37 Hamdan AL, Jabbour J, Nassar J, Dahouk I, Azar ST. Vocal characteristics in patients with type 2 diabetes mellitus. Eur Arch Otorhinolaryngol. 2012 May;269(5):1489–95.
38 Hamdan AL, Kurban Z, Azar ST. Prevalence of phonatory symptoms in patients with type 2 diabetes mellitus. Acta Diabetol. 2013 Oct; 50(5):731–6.
39 Anthes E. Alexa, do I have COVID-19? Nature. 2020 Oct;586(7827):22–5.
40 Laguarta J, Hueto F, Subirana B. COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open J Eng Med Biol. 2020;1:275–81.
41 Imran A, Posokhova I, Qureshi HN, Masood U, Riaz MS, Ali K, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Inform Med Unlocked. 2020;20:100378.
42 Feenaughty L, Tjaden K, Benedict RH, Weinstock-Guttman B. Speech and pause characteristics in multiple sclerosis: a preliminary study of speakers with high and low neuropsychological test performance. Clin Linguist Phon. 2013 Feb;27(2):134–51.
43 Gerratt BR, Kreiman J, Garellek M. Comparing Measures of Voice Quality From Sustained Phonation and Continuous Speech. J Speech Lang Hear Res. 2016 Oct;59(5):994–1001.
44 Arora S, Baghai-Ravary L, Tsanas A. Developing a large scale population screening tool for the assessment of Parkinson's disease using telephone-quality voice. J Acoust Soc Am. 2019 May;145(5):2871–84.
45 Godino-Llorente JI, Shattuck-Hufnagel S, Choi JY, Moro-Velázquez L, Gómez-García JA. Towards the identification of Idiopathic Parkinson's Disease from the speech. New articulatory kinetic biomarkers. PLoS One. 2017 Dec;12(12):e0189583.

46 Rusz J, Cmejla R, Tykalova T, Ruzickova H, Klempir J, Majerova V, et al. Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. J Acoust Soc Am. 2013 Sep;134(3):2171–81.
47 Akçay MB, Oğuz K. Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun. 2020; 116:56–76.
48 Sharma G, Umapathy K, Krishnan S. Trends in audio signal feature extraction methods. Appl Acoust. 2020;158:107020.
49 Kanabur V, Harakannanavar SS, Torse D. An extensive review of feature extraction techniques, challenges and trends in automatic speech recognition. IJIGSP. 2019;11(5):1–12.
50 Sajal MS, Ehsan MT, Vaidyanathan R, Wang S, Aziz T, Mamun KA. Telemonitoring Parkinson's disease using machine learning by combining tremor and voice analysis. Brain Inform. 2020 Oct;7(1):12.
51 Ali L, Zhu C, Zhang Z, Liu Y. Automated detection of Parkinson's disease based on multiple types of sustained phonations using linear discriminant analysis and genetically optimized neural network. IEEE J Transl Eng Health Med. 2019 Oct;7:2000410.
52 Miotto R, Danieletto M, Scelza JR, Kidd BA, Dudley JT. Reflecting health: smart mirrors for personalized medicine. npj. Digit Med. 2018;1(1):1–7.
53 Portet F, Vacher M, Golanski C, Roux C, Meillon B. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. Pers Ubiquitous Comput. 2013;17(1):127–44.
54 Perumal S, Javid MS. Voice Enabled Smart Home Assistant for Elderly. Int J Comput Sci Eng. 2019;7(11):30–7.
55 Zhang B, Rau PL, Salvendy G. Design and evaluation of smart home user interface: effects of age, tasks and intelligence level. Behav Inf Technol. 2009;28(3):239–49.
56 Coiera E, Kocaballi B, Halamka J, Laranjo L. The digital scribe. NPJ Digit Med. 2018 Oct; 1(1):58.
57 Kumah-Crystal YA, Pirtle CJ, Whyte HM, Goode ES, Anders SH, Lehmann CU. Electronic health record interactions through voice: a review. Appl Clin Inform. 2018 Jul; 9(3):541–52.
58 Vaida C, Pisla D, Plitea N, Gherman B, Gyurka B, Graur F, et al. Development of a voice controlled surgical robot. In: Pisla D, Ceccarelli M, Husty M, Corves B, editors. New trends in mechanism science. Dordrecht: Springer Netherlands; 2010. pp. 567–74.
59 Nathan CO, Chakradeo V, Malhotra K, D'Agostino H, Patwardhan R. The voice-controlled robotic assist scope holder AESOP for the endoscopic approach to the sella. Skull Base. 2006 Aug;16(3):123–31.
60 Nash DB. "Alexa, refill my omeprazole". Am Health Drug Benefits. 2017 Dec;10(9):439–40.
61 Isyanto H, Arifin AS, Suryanegara M. Design and implementation of IoT-based smart home voice commands for disabled people using Google Assistant. 2020 International Conference on Smart Technology and Applications (ICoSTA). 2020. doi: https://doi.org/10.1109/ICoSTA48221.2020.1570613925.
62 Ahad A, Tahir M, Aman Sheikh M, Ahmed KI, Mughees A, Numani A. Technologies Trend towards 5G Network for Smart Health-Care Using IoT: A Review. Sensors. 2020 Jul; 20(14):E4047.
63 Torous J, Hsin H. Empowering the digital therapeutic relationship: virtual clinics for digital health interventions. NPJ Digit Med. 2018 May;1(1):16.
64 Geis JR, Brady AP, Wu CC, Spencer J, Ranschaert E, Jaremko JL, et al. Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement. Can Assoc Radiol J. 2019 Nov;70(4):329–34.
65 Babrak LM, Menetski J, Rebhan M, Nisato G, Zinggeler M, Brasier N, et al. Traditional and digital biomarkers: two worlds apart? Digit Biomark. 2019 Aug;3(2):92–102.