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

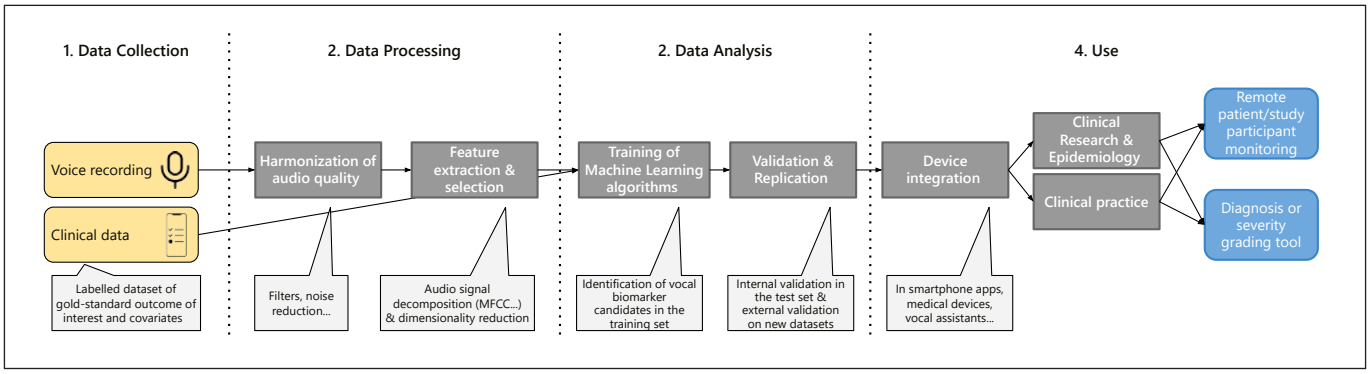

Fig. 1. Pipeline for vocal biomarker identification, from research to practice.

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.

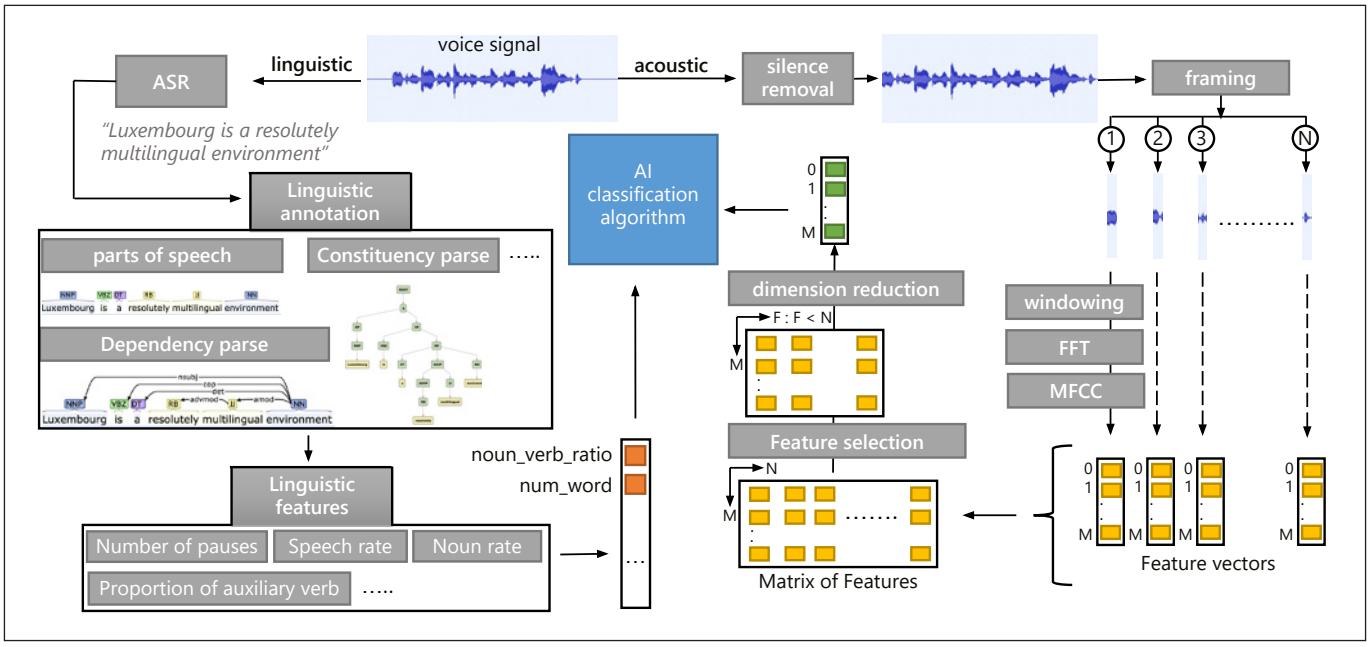

Fig. 2. Representation of a typical voice signal pre-processing and feature extraction using MFCCs. Representation of a typical voice signal pre-processing and linguistic and acoustic feature extraction. Voice signal represents the sound of the following sentence (e.g., "Luxembourg is a resolutely multilingual 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

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