Human Voice Sensing Through Radio Frequency Technologies: A Comprehensive Review

Yingxiao Wu, *Member, IEEE*, Jianping Han, Zhihua Jian, and Wenyao Xu, *Senior Member, IEEE*

***Abstract*— Voice print is one of the promising solutions for biometric applications by distinctive patterns of certain voice characteristics of an individual. Due to the nonintrusive and high-accuracy sensing nature, many researchers have shown considerable attention to human voice sensing through radio frequency (RF) technologies [e.g., Wi-Fi and millimeter wave (mmWave)] thereby realizing a set of emerging applications. It has become a remarkable research field over the past decade. In this article, we provide a comprehensive technological review of the current state-of-the-art of wireless technologies for human voice sensing. We perform a full analysis of voice print recognition connotations and applications and explore the principles of voice production. Human voice sensing technologies are divided into four modalities. In addition, we summarize the theory of wireless sensing and make an in-depth comparison of different RF technologies for human voice sensing. An overview of human voice sensing is proposed, and this article also highlights voice signal acquisition, signal preprocessing, feature extraction as well as prediction and recognition. Furthermore, this article presents a thorough survey of the current human voice sensing through RF technologies and compares them in detail. We discuss the hardware system architecture and classify various applications of wireless voice sensing. At the end of the article, we conclude the review with future directions and challenges associated with wireless human voice sensing.**

***Index Terms*— Biometrics, millimeter wave (mmWave), surveys and overview, voice sensing, wireless sensing.**

NOMENCLATURE

Acronyms	*Full names*
AW	Acoustic wave.
CFR	Channel frequency response.
cGANs	Conditional generative adversarial nets.
CIR	Channel impulse response.
COVID-19	Coronavirus disease 2019.
CSI	Channel state information.
DNNs	Deep neural networks.
DTW	Dynamic time warping.
DWT	Discrete wavelet transform.
EMW	Electromagnetic wave.
FFT	Fast Fourier transform.
FMCW	Frequency-modulated continuous wave.
GCRN	Gated convolutional recurrent network.
GMM	Gaussian mixed model.
GMM-SVM	GMM-support vector machine.
GMM-UBM	GMM-universal background model.
HCI	Human–computer interaction.
HMMs	Hidden Markov models.
IMCRA	Improved minima controlled recursive averaging.
IoT	Internet of Things.
JFA	Joint factor analysis.
LPC	Linear predictive codes.
LOS	Line of sight.
LDV	Laser Doppler vibrometer.
LSTM	Long–short-term memory.
MCRA	Minima controlled recursive averaging.
MFCCs	Mel frequency cepstral coefficients.
MFC	Mel frequency cepstrum.
mmWave	Millimeter wave.
MOS	Mean opinion score.
NLOS	Non LOS.
PESQ	Perceptual evaluation of speech quality.
PLP	Perceptual linear prediction.
RF	Radio frequency.
RNN	Recurrent neural network.
RSSI	Received signal strength indicator.
RSS	Received signal strength.
SIR	Signal to interference ratio.
SI-SDR	Scale-invariant signal-to-distortion ratio.
STOI	Short-time objective intelligibility.
TOA	Time of arrival.
TOF	Time of flight.
USRP	Universal software radio peripheral.
UWB	Ultra-wide bandwidth.
VCV	Vocal cord vibration.
VUI	Voice user interface.
VQ	Vector quantization.
WARP	Wireless open-access research platform.

Manuscript received 29 October 2023; revised 10 January 2024; accepted 6 February 2024. Date of publication 26 March 2024; date of current version 2 April 2024. This work was supported by the Zhejiang Provincial Natural Science Foundation of China under Grant LY23F010009. The Associate Editor coordinating the review process was Dr. Valentina Bianche. *(Corresponding author: Yingxiao Wu.)*

Yingxiao Wu and Jianping Han are with the College of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China (e-mail: wuyingxiao@hdu.edu.cn; hanjp@hdu.edu.cn).

Zhihua Jian is with the College of Communication Engineering, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China (e-mail: jianzh@hdu.edu.cn).

Wenyao Xu is with the Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14261 USA (e-mail: wenyaoxu@buffalo.edu).

Digital Object Identifier 10.1109/TIM.2024.3381252

I. INTRODUCTION

BIOMETRICS technology is a reliable and fast technology for identifying and authenticating individuals by measuring and analyzing their unique physiological and behavioral

characteristics, such as fingerprint, hand geometry, retinal pattern, and face contour [1]. Biometric traits are difficult to counterfeit and hence contribute to higher accuracy when compared to other authentication methods such as passwords and ID cards. Biometrics technology is widely used in HCI, finance, entertainment, information security, e-government, and other fields, and the global market scale continues to expand. According to the latest Reportlinker.com forecast published in October 2022 [2], amid the COVID-19 crisis, the global market for biometrics is estimated at U.S.$19.5 Billion in the year 2020, and expected to reach a revised size of U.S.$51.2 Billion by 2027. In particular, the use of vocal-based biometrics for user authentication has surged due to the growth of voice-controlled devices and services [3]. Voice print is a strong physiological and behavioral combined biometrics, considered to be just as biologically unique in individuals as a fingerprint [4]. Another natural advantage of vocal-based biometrics in applications is its directive and content variable feature. Compared with fingerprints, facial recognition can only perform simple actions such as shaking the head, blinking, and opening the mouth, vocal-based biometrics is text-independent and can recognize the speaker with random changes of voice content [5]. With the development of wireless sensing technology, RF brings new opportunities for noncontact human voice sensing, To enhance a profound understanding of the history, state-of-the-art, and future of human voice sensing through RF technologies, we present a comprehensive review of human voice sensing via RF in terms of basic principles, techniques, and applications.

A. Overview of Human Voice Sensing and Applications

Many research works have been conducted to develop human voice sensing. The primary difference between voice biometrics and other biometrics is that voice biometrics process acoustic information however most other biometrics are image-based [6]. Campbell [7] presented a comprehensive survey of voice print recognition systems for remote authentication and categorized the modules in speaker recognition, discussing different approaches for each module and challenges confronted with the speaker recognition systems. A summary of the design and development of automatic speaker recognition (ASR) systems is presented in the work. The fundamentals of speaker recognition have been covered with verification and identification. Then, this article introduces the speech processing and basic components of the microphone-based ASR system and discusses the design tradeoffs. Hanifa et al. [8] provided a comprehensive review of the literature on speaker recognition. This article discusses the advances and challenges of speaker recognition made in the last decade and also focuses on the system and structure of speaker recognition, and feature extraction and classifiers methods. The application of speaker recognition in the real-world is also introduced. Li and Zhang [9] made a study on speaker recognition technology, and compares and discusses the existing text-independent speaker recognition technology about advantages and disadvantages, then discuss the directions for the next research. Deep architectures have already had a significant impact on automatic speech recognition, Nassif et al. [10], Papastratis [11], and Zhu et al. [12] provided a thorough statistical analysis on utilizing deep learning for speech-related applications.

B. Human Voice Sensing Technologies

1) Contact and Noncontact Modalities: Human voice sensing, a fundamental component of HCI, communication systems, and healthcare monitoring, can be conducted via both contact and noncontact modalities.

a) Contact voice sensing: Contact-based voice sensing primarily involves the use of wearable devices, such as throat microphones and accelerometers, placed directly on the skin or worn around the neck. These devices capture voice signals through the vibrations of the vocal cords and surrounding tissues during speech. For instance, throat microphones, or laryngophones, have been widely used in high-noise environments like military operations [13]. Accelerometer-based devices [14] also provide a contact-based method for voice sensing and have found applications in laryngeal disorder detection. And bone-conduction sensors [15] at the top of the skull can pick up speech vibrations from the vocal cords to the bones.

b) Noncontact voice sensing: Noncontact voice sensing primarily relies on techniques such as microphone arrays, radar, and laser Doppler vibrometry. Microphone arrays capture sound waves in the air and have been extensively used in smart devices and home automation systems. Radar-based techniques, including mmWave radar, can capture the minute chest and throat movements during speech, providing a noncontact means for voice sensing. Laser Doppler vibrometry, while less common, offers a highly accurate, noncontact method to measure VCVs.

Both contact and noncontact modalities have their advantages and disadvantages. Contact methods can provide a high signal-to-noise ratio (SNR) and are less affected by environmental noise. However, they may cause discomfort and inconvenience to the user due to their intrusive nature. Noncontact methods, while less intrusive, may be more susceptible to environmental noise and require sophisticated algorithms to extract voice signals effectively.

2) Passive and Active Modalities: According to whether the sensors can provide their own detection signals, there are two ways of human voice sensing: passive and active.

a) Passive voice sensing: Passive voice sensing refers to sensors that are activated to capture and measure voice when the acoustic source has audio information. Microphones or speakers are typical passive sensors for voice sensing. Microphones are the most commonly used sensors for acoustic sensing, however, they can only sense the sound from one sound source. In addition, they are vulnerable to fake voice and replay attacks because they collect voice indirectly.

b) Active voice sensing: An active system means that the system itself can generate energy. Active voice sensing automatically sends detection signals and measures the backscatter reflected to the sensor. Electromagnetic (EM) wave-based sensing methods acquire speech or acoustical vibrations by

transmitting EM waves to the objectives, such as light waves, laser, and mmWave.

The acquisition distance and sensitivity of passive voice sensing are limited, the wide frequency bandwidth of EM wave breaks through the sensing details and distance, and EM wave-based active voice sensing may offer continuous and multisource voice measure. Active voice sensing is a better choice for live or continuous works.

C. Objective of the Review

To the best of our knowledge, existing surveys are limited to the scope of speaker recognition and speech recognition based on traditional signal processing methods and AI techniques. In contrast, the purpose of our research is to provide a comprehensive survey on the use of wireless technologies for voice sensing, outline and explore potential areas in the scope of application and solution of human voice sensing, and identify important trends of development. Our contributions can be summarized as follows.

1) Our study presents voice print recognition connotation, explores the principles of voice production and summarizes the fundamentals of wireless sensing-based voice print recognition.
2) Our study focuses on the topic of EMW base sensing techniques for voice print recognition, covering the evolution, principle, and algorithm of traditional wireless signal processing to deep learning approaches.
3) Our study summarizes the application of voice print recognition and elaborates a comprehensive overview of voice print recognition based on wireless sensing techniques.
4) Our study provides a set of research challenges and potential directions along which future studies can refer to enable performance improvement of intelligent signal processing methods for voice print recognition.

The remainder of the article is organized as follows. Section II discusses the related work and our article selection criteria. Section III investigates the voice print recognition connotation, the principles of voice production, and the comparison of different wireless technologies for voice sensing. Then, we study the overview of human voice sensing and the technologies of voice print recognition based on wireless sensing in Section IV. Section V provides the applications of wireless voice sensing. Future directions and discussion are proposed in Section VI, and we finally conclude our survey in Section VII.

The acronyms are given in Nomenclature for easy understanding and simplified expression.

II. Related Work and Paper Selection Criteria

As far as we know, there is currently no comprehensive paper investigating a survey on voice perception based on RF technology. Kabir et al. [16] summarized a systematic review of ASR based on the existing architectures, datasets, techniques, limitations, and challenges, without involving RF sensing technologies for ASR. The reviews [8], [17] mainly discussed speaker recognition with speech recording signal. Anthony and Patil [18] focused on the methodologies of speech emotion recognition using machine learning approaches. Zhang et al. [19] reviewed the technology, platforms, and applications of broad human sensing based on mmWave, including human motion recognition, localization, and biometric measurement. mmWave-based sound recognition as a branch of biometric measurement is investigated. Instead, the purpose of our review is to provide a comprehensive survey on human voice sensing through RF technologies, which focuses on EM-based sensing. The main point of investigation will be the principles of voice production and wireless sensing modality, as well as the voice print sensing technologies and architectures. Furthermore, we are concerned about the advances, challenges, and research trends in this area.

We investigate academic papers concentrating on voice recognition and RF sensing which have been published in the last about 20 years, not just online digital libraries, including IEEE, ACM, Science Direct, Elsevier, and arXiv. The search scope encompasses papers related to voice recognition modalities, RF sensing modalities, and applications of voice recognition. We select representative papers in terms of voice print connotation, applications of wireless voice sensing, and RF technologies for voice sensing based on lasers, Wi-Fi, and mmWave. Moreover, the papers are systematically arranged according to the logical classification of each section for each topic.

III. Fundamentals of Wireless Voice Sensing Technologies

A. Voice Print and Sensing Connotation

1) Voice Print: Voice print, also known as voice biometrics, has a special vocal behavioral characteristic, and the hybrid of phonetic characteristics and pronunciation habits contained in each person's speech process is almost unique. The phonetic characteristics of voice print are mainly the physical shape of the speaker's voice tract and the pronunciation habits which are affected by the physical movement of lips, jaws, tongue, and larynx.

By using an electro-acoustic conversion instrument, the pattern recognition of voice print is to depict the speech pattern with an unknown person's speech sample and understand their speech material. The acoustic features of speech on the map are compared and comprehensively analyzed to obtain the estimate process of whether they are the same.

A voice print is a human body's unique vocal patterns and has long-term stable signal characteristics [20]. For recognition, voice print has the following main advantages. *Naturalness*: voice print is a form of biometric data that comes from the combination of individual physical and behavioral voice patterns, it is not artificially produced. *Uniqueness*: the generation of voice print depends on the vocal organs of the individual which means unique individually distinctive patterns. *Variableness*: voice print is the perfect unity of high variability and uniqueness. The features of a person's voice print will remain invariant and unique even if his speech content or tone is changed randomly. And that voice print remains different even when the speaker deliberately mimics

the voice and tone of others or speakers' environmental and psychological changes. This perfect unity of high variability and uniqueness makes the voice signal itself have a strong anti-attack ability. *Imperceptibleness*: since voice print originates from VCV, it can be collected and authenticated continuously through RF signals in a way that people cannot perceive.

Because of the particularity of voice print generation, voice print recognition technology has physiological characteristics and behavioral characteristics compared with other biometrics. Even if imitation, it is difficult to cover up the most essential pronunciation characteristics and vocal tract characteristics of the speaker. As a method of biometrics identity authentication, voice print recognition is not easy to invalidate, difficult to forge, and better privacy.

2) Voice Print Sensing: Voice signal carries information about both linguistic and indexical cues. Linguistic cues are related to the lexical content of the speaker's intended message, conveying the message by means of phonological, morphological, syntactic, and semantic information within the utterance. Indexical factors, also called paralinguistic or extralinguistic factors, provide information about the speaker's characteristics, including cues to speaker identity, gender, and emotional state information [21].

According to different purposes of identification, voice print sensing can be broadly divided into speech recognition, speaker recognition, language recognition, and health and emotion recognition [7]. The comparison of four voice print sensing is shown in Table I. Speech recognition focuses on converting human speech into text, also known as speech-to-text, its transformation effectiveness is susceptible to the quality of the language and text corpus. It identifies the universal characteristics of each speech unit by excluding the personal characteristics of different speakers. Language recognition focuses on categorizing the languages from audio speeches spoken by speakers according to the characteristics of languages. It identifies the language type to achieve language translation or distinguish multilingual speech. Unlike speech and language recognition, speaker recognition aims to identify the speaker and recognize persons by extracting features from voice pattern, speaking style, and other verbal traits, without concern about the content and meaning of the sound [8]. Health and emotion recognition distinguish different emotional states or medically relevant speech from indexical information of voice which helps detect the mood of the speakers.

The process of recognition encompasses verification and identification. Voice print verification is a one-to-one analysis process that determines whether a speaker is a person he claims to be (makes an identity claim, e.g., by entering an employee license or presenting his ID card). This is different from the voice print identification problem. Identification is a one-to-many analysis process supposing that a speaker is among a group of people or a specific individual. The essence of effective recognition relies on extracting unique features of each prerecorded voice sample and establishing a feature database. For recognition, the sound to be detected is matched against the features in the database, and the speech recognition is realized through meticulous analysis and sophisticated algorithms.

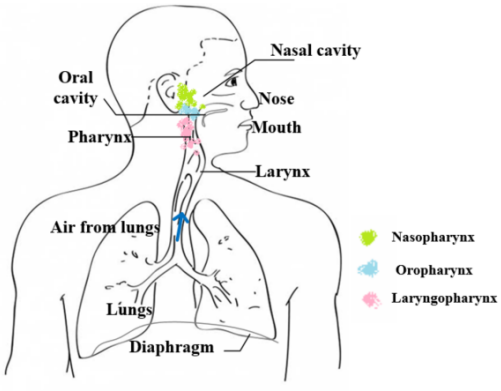

Fig. 1. Anatomy of the human voice production.

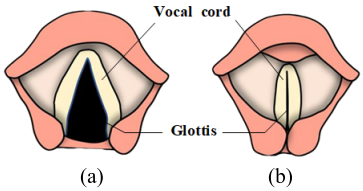

Fig. 2. Diagram of vocal cord structure: (a) opened glottis and (b) closed glottis.

B. Principles of Voice Production

1) Anatomy and Physiology of Voice Production: The sound of your voice is produced when the vocal cords vibrate rapidly in a series of vibration cycles due to aerodynamic forces. There are many structures involved in the human vocal system, including the mouth, nose, oral and nasal cavities, pharynx and larynx, lungs, and diaphragm, as shown in Fig. 1. The generation of sound originates from a complex fluid-structure-acoustic interaction process in the human vocal system, which is also related to the geometry and material properties of these structures [22]. The vocal cord located in the middle of the laryngeal cavity is the main organ for producing sound, at the same time, the lungs provide airflow for pronunciation driven by the diaphragm and chest. All structures above the larynx provide resonance to adjust the sound volume and timbre.

a) Lungs and diaphragm: To exchange oxygen, the lungs' structure contains many small, elastic air sacs called alveoli, which are filled with inhaled air. The movement of air in the lungs is manipulated mainly by the diaphragm which is a dome-shaped muscle located at the bottom of the lungs. When human speech sounds, the main function of the lungs is to provide the air flowing power for humans and store a large amount of air, which will be applied to pass through or may vibrate the vocal cords when you exhale, as shown in Fig. 2. The air passing through the larynx will cause the edges of them to vibrate when the vocal cords are approaching each other. However, no sound will be produced when you breathe and the vocal cords open widely. The vibrations of vocal cords create a series of sounds when we speak and sing.

b) Larynx: The larynx is located in the middle of the neck and is composed of muscle, cartilage, and tendons. The larynx plays a triple role in respiration, pronunciation, and resonance. The most important vocal organ in the larynx is

TABLE I
VOICE PRINT RECOGNITION

Types	Description	Purpose	Application
Speech recognition	Recognizes individual words or phrases and Convert speech into text	Identify the spoken words of the individual speaking	Speech to text or use this text to command the operation of a system
Speaker recognition	Recognizes persons by extracting features from voice pattern, speaking style, and other verbal traits	Identify the speaker	Voice bio-metrics for safety systems and user recognition
Language recognition	Recognizes the type of language in a speech sample	Identify the language type	Spoken language translation or multilingual speech recognition
Health and emotion recognition	Recognizes a person's inner emotional state and auxiliary diagnosis of disease	Identify the speaker's state	Health and social care or human–robot interaction

the vocal cords, also known as vocal folds. Vocal folds are a membranous anatomical structure with two symmetrical left and right folds. Speech, singing, and other vocal actions need to be realized through the movement of the vocal cords. The cartilage of the larynx as a throat fortress supports and holds the muscles of the vocal cords for opening, closing, tightening, and relaxing.

When the vocal cords are closed but not tightly pressed together, the air passing through the laryngeal cavity during exhalation will cause the edges of the vocal cords to vibrate and produce sound. The vibration amplitude of the vocal cords refers to the degree of opening and closing of the vocal cords, which is determined by the impact of the vocalization breath. When the impact of vocalization breath is strong, the vibration force received by the vocal cords is greater, the phase of opening and closing is also greater, and the sound intensity increases. The vibration frequency of the vocal cords refers to the number of times the vocal cords open and close within a second of sound production. The more times the vocal cords open and close, the higher frequency of the vibration, and the higher the sound pitch produced. A typical vocal tract of speech production system is shown in Fig. 1.

c) Pharynx, mouth, and nose: The main structures of the vocal tract include vocal cords deep in the throat, pharyngeal cavity, oral cavity, nasal cavity, soft palate, hard palate, tongue, teeth, and lips. The laryngopharynx, oropharynx, nasopharynx, oral cavity (mouth), and nasal cavity (nose) are five hollow cavities of the vocal tract, three of them make up the pharynx located at the back of the throat and the other two cavities are in the mouth and nose. The anatomy with these five hollow cavities makes the sound of speech resonate and shaped above the vocal cords [23].

Sound is produced when air flows through the vocal cords and causes the vocal cords to vibrate when they are contacted. When the sound wave generated by VCV passes through the throat and oral cavity, it will be amplified and generate resonance in the mouth and back of the throat. And if it goes up through the nasopharynx and head, and down through the chest and abdominal cavity, this leads to a higher cavity resonance. Resonance amplifies sound and also adjusts the color and timbre of the sound by enhancing certain sound qualities. As air flows upward from the lungs, the movements of structures within the vocal tract create each recognizable sound.

2) Voice Print Affecting Factors: According to the anatomy and physiology of voice production, the vocal tract and the shape of the glottal pulse are considered to be the two salient factors that influence vocal vibration [24]. The glottis pulse is generated by the tissue flaps in the vocal cord area and its gaps, which are known as the glottis. The laryngeal muscle that has the effect of tense vocal cords, which is the striated muscle. Its function is to tighten or relax the vocal cords, and open or contract the glottis. So the related glottal pulse parameters include closing instants rate, opening duration, and opening degree of the glottis.

TABLE II
VOICE PRINT AFFECTING FACTORS

Symbol	Influencing factors
Vocal cord	Glottal pulse parameters Musculature and tension of vocal folds Phase information of excitation waveform Energy for speech production
Vocal tract	Length and dimensions of vocal tract Resonance frequency Resonance locations Bandwidth of the speech spectrum Spectral peak location

The vocal tract is located above the glottis and the vocal cavity consists of multiple articulatory organs including the pharyngeal cavity, oral cavity, nasal cavity, and various organs within the vocal tract. The vocal tract is a regulatory cavity that plays an important role to resonate and reshape vocal source signals. The speed and intensity of articulatory organ movement modulate the glottal sound source to produce different sounds and vary from person to person. The main influencing factors on voice print are shown in Table II.

C. RF and mmWave Sensing

Since the COVID-19 virus outbreak in December 2019, the disease has spread to almost every country around the world. The global impacts of COVID-19 are beginning to emerge, and have significantly affected the sensor application market in 2022. Noncontact sensing to improve security certification and healthcare is playing a vital role in future applications. More importantly, most speaker biometrics systems mainly use microphone sensors at present, and this kind of indirect voice sensing method through a media increases the risk of unpredictable attack. For example, fraudsters may eavesdrop on the legitimate user's voice samples or utilize artificial intelligence technologies to clone a synthetic voice against voice-based authentication systems [25], [26]. Therefore, the solution to the major security concerns for voice biometric

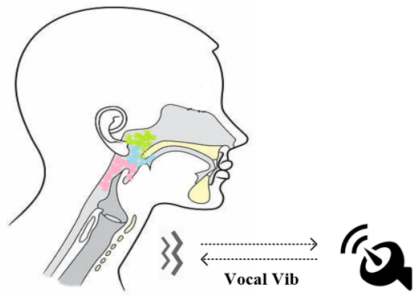

Fig. 3. Wireless sensing for vocal vibration.

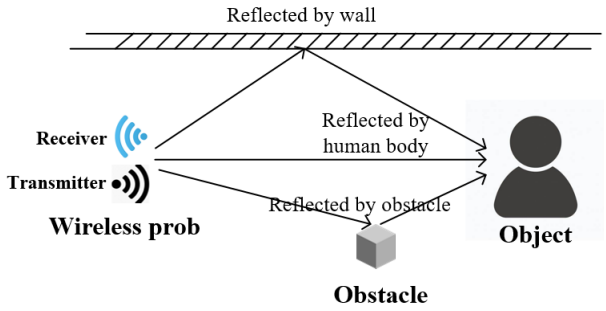

Fig. 4. Multipath effect of wireless signals.

technologies needs a new direct way to acquire and analyze the user's voice according to the analysis of voice print affecting factors.

Sound wave is the transmission form of sound, which is a mechanical wave generated by the vibration of objects, whereas EMWs are produced by accelerating or decelerating charged particles. Therefore, sound waves need a medium for their propagation whereas EMWs can travel without any medium. This is the main difference between sound waves and EMWs. The frequency of sound waves generated by human voice ranges from approximately 60 Hz to 4 kHz [27]. These sound waves cause resonance in nearby objects, these vibrations can be converted into audio signals by using analog electronic devices. To establish the interaction model between speech signal and wireless signal, different wireless measurement techniques have been investigated by many researchers.

Fig. 3 shows vocal vibration is the near-throat skin vibration caused by speech airflow and movement of articulatory organs which will modulate the RF signal, the modulated signal contains speech information.

1) Basic Theory of Wireless Sensing: Wireless sensing can be broadly defined as extracting information from EM signals on possibly everything that they encounter. In the meanwhile, the EM radar sensor has been a promising alternative for various applications associated with phonation. EM radar sensors have become a promising alternative noncontact solution for various applications. Wireless sensing almost can extract information about everything they may encounter from EM signals. Since wireless signals are almost everywhere and can travel through almost any corner of a relatively enclosed space, the propagation of wireless signals would be disturbed by the presence or movement of objects in the surroundings, which leads to tiny changes in multiple reflected signals as shown in Fig. 4. All these multipath signals could contribute to detecting the object's movements [28]. The transmitted signal will be disrupted when an object presents or moves in the sensing area, so it is possible to infer and capture signal feature information by analyzing the differential characteristics between the transmitted signal and received signals reflected off the object. The differential characteristics can be obtained by the change of signal on time, frequency, phase, or magnitude domain.

Wireless sensing signals have several different domain features that can be extracted from raw signals and used for identification and certification.

a) Time domain features: The time domain describes the relationship between physical signals or mathematical functions and time. Time domain analysis refers to analyzing the signal directly on a time series of physical signals or mathematical functions in reference to time. Some commonly used time domain features include mean, variance, Kurtosis, Skewness, high-order statistics, and distribution. Another time domain feature is CIR which depicts individual multipath components in the time domain and facilitates to separate LOS paths for accurate ranging [29]. ToA and ToF are classic methods used in ranging the distance of an object or used in positioning.

b) Frequency domain features: Frequency domain analysis reveals the inherent frequency characteristics of the signal and the close relationship between the signal time characteristics and its frequency characteristics. The signals in the time domain can be transformed into the frequency-domain leveraging Fourier transformation. The frequency domain contains a large variety of spectral representations that are either difficult to observe or invisible at all in the time domain. The characteristics of frequency domain signals can be depicted by statistical features such as entropy, peak frequency, CFR, and spectrum distribution. CFR is the Fourier transform of the CIR, it describes the impact of frequency-selective fading channels on signal propagation through amplitude-frequency characteristics and phase-frequency characteristics.

c) Phase domain features: The phase of the signal serves as a measure of the time offset relative to the beginning of the cycle. With FMCW radar, the phase transformed with time can be used to estimate the amplitude and periodicity of vibrating objects. Phase modulation of reflected waves caused by weak physiological movements on the body surface is inversely proportional to the wavelength, so the voice signal is hidden in the phase change information.

d) Particle domain features: Acoustic signal can be received in air medium by mmWave radar because of the movement of the particles and particle clouds caused by acoustic source, and it induces fluctuations in EM field intensity and frequency changes with the changes of sound wave's scalar nature, which depends on factors such as sound potential and particle size, density and concentration [30].

e) Deep features: Compared with traditional learning methods, features are often extracted by experienced experts, while deep learning methods automatically extract features from denoised signals through artificial neural networks. Deep learning methods reduce human intervention for deep feature extraction, making it more comprehensive and effective.

Typical deep learning networks include generative adversarial networks (GANs), transformer networks, and convolutional neural networks (CNNs).

There are many signal models built for feature extraction to facilitate voice print acquisition and recognition. We have identified five frequently used wireless sensing technologies based on different model characteristics, including RSSI, CSI, FMCW, Doppler shift, and scattering in particles.

2) Received Signal Strength Indicator/CSI: To quantify the difference in the received RF signal via multiple paths, researchers can evaluate the physical layer properties over wireless channels, such as RSSI and CSI.

RSSI is an estimated measure of the power level between the transmitter and receiver. According to the attenuation model between signal and distance, the distance between nodes can be calculated based on RSSI. RSSI is also often used in positioning applications through three-point positioning. RSSI is the most commonly used measurement for wireless sensing applications. The limitation of RSSI is that the propagation of wireless signals will be affected by obstacles, multipath effects, and other factors, which indicate they are not reliable or consistent in complex, high-mobility environments. And another indicator can characterize multipath propagation, which is known as CSI. CSI is a fine-grained indicator that expresses channel state and presents the amplitude and phase of multipath propagation at different frequencies (for the orthogonal frequency division modulation (OFDM) based wireless standards (e.g., IEEE 802.11a/g/n/ac/ax), the different frequencies are corresponding to different subcarriers), thus more accurately represents the combined effect of scattering, fading and power attenuation with the distance [31].

To achieve accurate and reliable human activity sensing, CSI reveals the channel response at each subcarrier frequency f and different timing t. It is generally expressed as the superposition of the multipath effect

$$H(f,t) = \sum_{i=1}^{N} a_i(f,t) \cdot e^{j\theta_i(f,t)} \tag{1}$$

where $a_i(f,t)$ is the amplitude attenuation at ith signal path, $\theta_i(f,t)$ is the phase change, and N is the number of reflection multipaths. Each CSI depicts the amplitude and phase of one OFDM subcarrier. By analyzing the state changes in RSSI or CSI, it is possible to infer changes in the surrounding environment and capture the object's movements.

3) Doppler/FMCW: Doppler shift effects are another wireless signal property that can be used to perform noncontact sensing. Specifically, the principle of the Doppler effect is it senses the relative motion by measuring the frequency variation of the received wireless signal reflected from the human body movements. If Doppler radar radiates a special continuous wave with frequency modulation, it is called FMCW radar which can measure the speed of the target and the distance of the target from the radar, and may also distinguish multiple targets.

The Doppler/FMCW radar synthesizer generates periodic frequency chirps transmitted by the TX antenna. The transmitted chirp $X(t)$ is denoted by a start frequency f_c, bandwidth B, and duration T_c as follows:

$$X(t) = e^{j\left(2\pi f_c t + \pi \frac{B}{T_c} t^2\right)}. \tag{2}$$

The TX transmitted chirp is reflected from an object and received by the RX antenna, the received reflected signal is a delayed version of the transmitted chirp, so the receiver signal is

$$R(t) = e^{j\left(2\pi f_c (t-t_d) + \pi \frac{B}{T_c} (t-t_d)^2\right)} \tag{3}$$

where t_d is the time delay, $t_d = 2d/c$, d is the range between the object and the radar and c is the speed of light.

The TX signal $X(t)$ and RX signal $R(t)$ are then mixed and filtered by low-pass filtering as a beat-frequency signal $f(t)$

$$f(t) = X(t) \bigotimes R(t) = e^{j(2\pi f_b t + \phi_b)} \tag{4}$$

where f_b is the beat frequency which is equal to the difference of frequencies of $X(t)$ and $R(t)$ and $\bigotimes$ denotes the mixer. Similarly, ϕ_b, referred to as a phase, is equal to the phase difference of the transmitted and received chirps. In particular, due to the time delay related to the range of the target, the relationship between the phase and the object range can be inferred as follows:

$$\phi_b = \frac{4\pi d}{\lambda} \tag{5}$$

where λ refers to the wavelength of the radar signal. According to the Doppler effect, the phase difference is induced by the movement of the target between the transmission of each chirp. By measuring the phase difference across consecutive chirps, the radar can estimate the velocity of the target. The phase difference $\Delta\phi$ is derived from (5) and we can derive the velocity v

$$\Delta\phi = \frac{4\pi v T_c}{\lambda} \tag{6}$$

$$v = \frac{\lambda \Delta\phi}{4\pi T_c} \tag{7}$$

where vT_c means the object movement distance in duration T_c. By using the Doppler frequency shifting principle, any velocity or movement of an object can be derived in the phase changes of the beat-frequency signal [32].

4) Scattering in Particle: EM radiation is one of many forms of energy. When EMWs propagate through matter, the interaction between EMWs and particles in the matter produces unique scattering patterns, which are related to wavelength and particle size. Li [30] used mmWave Doppler radar to detect and recognize the sound wave signals in air medium space, the detection principle is based on the theory of the interaction between the EMW and AW on the large numbers of particles and particle clouds in air and on the interface. The particles and particle clouds moving caused by acoustic source will derive fluctuation in the EM field strength and frequency change. So that voice signals can be captured with EMW scattering or reflecting which is perturbed by the sound. Scattering analysis uses the form of infinite series expansion of vector spherical harmonics, which allows the

cross sections, efficiency factors and distributions of intensity to be predicted using Mie theory [33]. The scattered energy is derived as

$$W_{\text{sca}} = \frac{1}{2} \oiint_S R_e[\mathbf{E}_{\text{sca}} \times \mathbf{H}^*_{\text{sca}}] \cdot \mathbf{n} dS, \ [W] \tag{8}$$

where R_e refers to the real part and $[W]$ indicates the unit of scattered energy rate. $\mathbf{E}_{\text{sca}}$ represents the electrical intensity vector of the scattered waves, $\mathbf{H}_{\text{sca}}$ represents the magnetic intensity vector of the scattered waves calculated from Faraday's law and $\mathbf{n}$ is a unit vector normal to the imaginary sphere [34]. S represents the surface of an imaginary sphere around the particle. Wireless sensing schemes can be captured with a change in the response of the object's EM field under the EMW.

D. Comparison of Different Wireless Technologies for Voice Sensing

A voice print is related to physical factors which are the size and shape of the individual vocal tract and glottal pulse, as well as contains a combination of an individual's accent, inflection, and rhythm. Therefore, voice print is one of the most unique modalities of identification that a person can produce. Based on this feature, voice biometric recognition for identity verification become one of the most in-demand and promising security technologies. The technology first emerged in the late 1990s and has continued evolving and improving ever since. In 1867, Alexander Melville Bell, the father of telephone inventor Alexander Graham Bell, invented a language called Universal Alphabetics for future voice biometrics research [35], this system replicated the position of a mouth when a particular person is speaking a certain speech pattern, and therefore transcribe what and how a person is saying. According to the voice print acquisition method, voice sensing modality developed from the speech and acoustic transducers stage to the wireless sensing stage.

Up to now, multitudinous research groups have developed a number of different techniques for detecting voice signals, which can be obtained from the vibrational information of particles in an air medium or object surface caused by sound. Conventional speech and acoustic transducers, such as microphones, microphone array [36], a pressure transducers, detect voice signals by perceiving the motion of air particles and reworking them into electric signals when sound propagates through an air medium [37]. These conventional speech and acoustic transducers are based on the theory that voice print can easily be heard and recorded when conducted by air in free space, so this method is almost limited to the air-conducted voice print acquisition and suffers from some serious shortcomings such as limited detection distance, low directional sensitivity, low speaker recognition rates, high susceptibility to acoustic interference noise, and narrow acoustic frequency bandwidth [38]. In the meanwhile, voice sensing based on conventional acoustic transducers or wearable vibrators is a passive or contact method to acquire voice signals. To overcome sensitivity decreases of conventional speech and acoustic transducers due to mechanical resonance and the damping effect, Cho et al. [39] developed a flexible and wearable vibration-responsive sensor, which consists of an ultrathin polymer film and a diaphragm with tiny holes. This sensor is a contact method by attaching to a neck, it can precisely recognize voice by neck skin even in noisy surroundings and at a low voice volume with a mouth mask worn.

To overcome the shortcomings of the traditional acquisition method, the voice print acquisition method which does not depend on conduction by air is required. The promising technique that has been thoroughly explored is RF technology for voice sensing. Previous studies have explored RF (e.g., Laser, LDV, Wi-Fi, and mmWave) based techniques [40] to perform voice print recognition which is generated by a reflected or scattered signal from a person that exists in free space. Reflected signal changes RF phase and attenuation which contains an individual's vocal fold vibration directly related to the voice source can be further utilized for recognition tasks. RF has some special features relative to traditional voice acquisition methods, such as low-range attenuation, a nice sense of direction, wide frequency bandwidth, and high sensitivity. Voice print based on RF is an active and noncontact sensing modality.

To summarize, the major characteristics of different wireless technologies for voice sensing are as follows.

1) The primary frequency of lasers is located between visible, infrared light, and ultraviolet regions of the spectrum, with a frequency range of approximately 405–790 THz. Lidar has extremely high detection accuracy and can measure the motion at the micrometer level. However, it is expensive and cannot detect vibrations under NLOS.
2) Wi-Fi is extensively deployed for IoT. The main used approaches of Wi-Fi-based sensing include RSS and CSI. Wi-Fi can sense NLOS vibrations by RSS since the object vibration modulates the RSS of the reflected signal. However, the vibration-sensing ability of Wi-Fi is much lower than lasers, and it suffers from severe signal reflection and scattering losses.
3) UWB uses a broad spectrum frequency bandwidth of over 1 GHz. Compared to Wi-Fi, UWB measures the distance by estimating the TOF of signals, and because its ultrashort width of pulse signals is over a large bandwidth, UWB owns excellent time resolution. In addition, it consumes much less energy than Wi-Fi.
4) mmWaves operate in the extremely high-frequency range, within the range of 30–300 GHz. The major advantages of mmWave are high frequency and wide bandwidth. similar to Lidar, a mmWave radar system can measure the range, velocity, and angle information of the objects by capturing the reflected signal. The vibration sensing method based on mmWave is based on the Doppler shift, which maps the object vibration by analyzing the phase change of the reflected signals. The vibration sensing accuracy of mmWave wave is higher than Wi-Fi. It can also detect objects in NLOS and it's a robust all-weather radar.

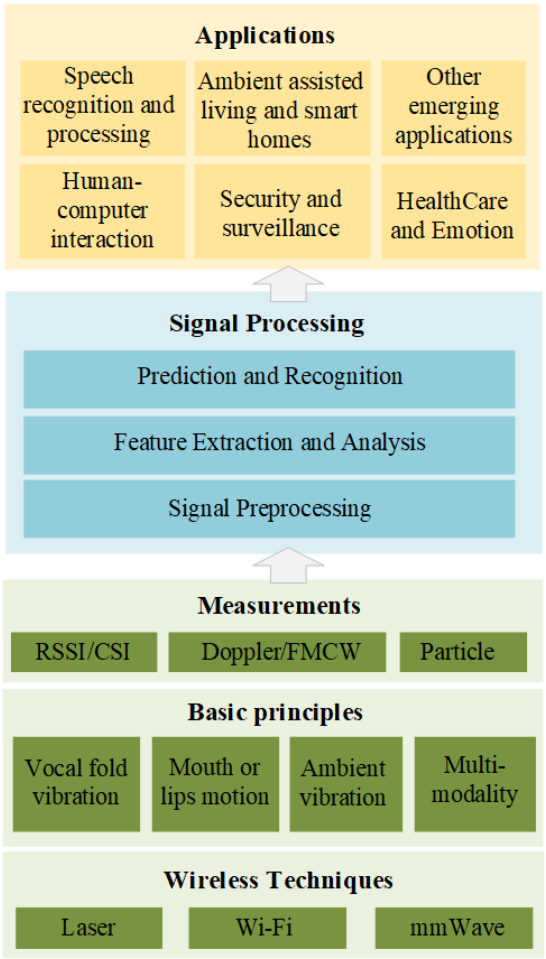

Fig. 5. Typical workflow of vocal print sensing via a wireless signal.

IV. Voice Sensing Technology Based on Wireless Signal

A. Overview of Human Voice Sensing

By taking different types of original signal information as inputs, we can derive diverse outputs from different feature domains across many sensing purposes including detection, recognition, classification, identification, and positioning. Fig. 5 shows a typical workflow of a voice print recognition system using wireless sensing. In order to achieve the noncontact sensing of voice print signs, the voice sensing systems first made an echo signal acquisition which was caused by the micro physiological movements of the body surface based on different wireless sensing methods (e.g., RSSI/CSI, FMCW/Doppler shift, and scattering in particles). To mitigate the impact of interference, ambient noise, and system offset on system performance, a series of signal preprocessing procedures (e.g., filtering, denoising, and calibration) are necessary. Then, the unique features are extracted from the echo signal and perform voice print detection and recognition for wild application.

1) Voice Signal Acquisition: This section presents four types of biometrics for RF-based voice sensing. Sensors collect the reflected signal with hidden voice signals from vocal fold vibration, mouth/lip motion, ambient vibration, and multimodality.

a) Vocal fold vibration: The human voice is determined by the vibration of the vocal cords. Acquiring and analyzing the people's vocal fold vibration directly using RF is the well-optimized solution for voice sensing [41], [42], [43], [44], [45], [46], and it is resistant to security attacks.

b) Mouth or lips motion: The movements and configuration of the mouth or lips are the key to how people produce articulated speech and control them to change the sounds. It is deduced that mouth or lip motion is highly correlated with the voice signal and describes the characteristics of the speech. References [47], [48], and [49] took mouth or lip motion as biometric features for speech recognition by wireless or camera sensing.

c) Ambient vibration: To recover sound signals by reducing the impact of random body motion, the work of [50] proposed a full-field targets vibration measurement via mmWave signals for sound recovery. It is an indirect voice reconstruction approach. References [51], [52], and [53] investigated the sound recovery from speaker or smartphone vibrations. In view of an existing problem in NLOS sound source scenarios, Zhang et al. [54] exploited the principle that the people's voice induces correlated vibrations of the surrounding objects, and presented a mmWave-based voice sensing algorithm under NLOS conditions by sensing the surrounding objects' vibration.

d) Multimodality: The advantage of multimodal fusion over single modal signal is the complementary collaboration between each source signal to get a better sensing performance. To achieve more accurate verification, Dong and Yao [48] and Fan et al. [55] utilized VCV biometrics and lip motion as additional features. Liu et al. [56] fused mmWave signals from subjects' mouths, throats, and audio signals from a microphone as a multimodal feature to improve the accuracy of speech recognition under clutter conditions.

2) Signal Preprocessing: In real-world scenes, the collected raw voice signal may contain various noises and clutter interference, which is crucial to the success of voice print recognition under the condition of inherently low SNR in the received signal. After collecting the raw wireless-detected vocal vibration signal, a whole train of signal preprocessing methods was exploited for making noise and clutter suppression, so as to obtain a relatively clean signal.

a) Noise-reduction: Traditional noise reduction is based on the statistical meaning of minimum tracking noise estimation, MCRA, IMCRA, and other noise estimation methods. According to the characteristics of the signal, some researchers also proposed specific noise reduction methods. Since normal speaking frequency is 150–300 syllables/min, Wang et al. [47] applied a three-order Butterworth IIR bandpass filter to eliminate out-band interference. The received signal may emerge as a dc offset due to an asymmetrical response in the receiver sensitivity, Sami et al. [57] corrected the dc offset by compensating the mean of the signal from the original signal. A noise-driven and short-time phase-spectrum-compensation enhancement algorithm was proposed in [58] to reduce the impacts of various noises under speech detected by both

the radar systems and the microphone, those noises include electrical-circuit noise, harmonic noise, channel noise, and ambient noise. The empirical mode decomposition (EMD) algorithm method is particularly suitable for the analysis and processing of nonlinear and nonstationary signals, so Chen et al. [59] explored EMD to decompose the noisy signal for a radar-sensing vocal vibration signal.

b) Body motion compensation: Random body motion such as head movements, shoulder movements, and arm stretching from speakers is the main dynamic interference in the received signals, which impacts the measurements of feature extraction. Xu et al. [60] developed a fine-grained range profile alignment solution to solve the misalignment problems. To eliminate large body movements while speaking influence, Dong and Yao [48] computed the mean of phase values of every ten samples, these ten samples will be discarded to remove large body movements if the energy of one segment exceeds a predefined threshold. Rodriguez and Li [61] applied a high-pass filter to eliminate the random body movement and its harmonics.

c) Clutter suppression: Unexpected backscatters caused by static or dynamic surrounding objects are the leading obstacle to obtaining precise voice by wireless sensing because they may corrupt the short-term spectral properties. Outdoor (e.g., buildings, snow, and rain) and indoor (e.g., walls, furniture, and appliances) background surrounding objects, even inherent body motion from breathing or heart rate, are all able to reflect the wireless signal and bring relative static clutter. Ozturk et al. [62] extracted the variance at each range-azimuth bin and used a threshold to filtrate static objects, besides, the similar approach also can remove excessive motion of bodies. Xu et al. [60] modeled the background clutter using Weibull distribution and isolated the background clutter with a resilient threshold. Wang et al. [63] applied a Butterworth finite impulse response (FIR) filter on each distance bin to ensure zeros phase distortion. In the meanwhile, dynamic clutter caused by the moving objects (such as passersby and vehicles), its amplitude of spectrum envelop does not obey the Weibull distribution, which is different from static objects, so Xu et al. [60] leveraged the information across multiple RDMs to mitigate the effect of dynamic clutter.

3) Feature Extraction and Analysis: The basic principle of feature extraction is to analyze and unearth a sequence of distinctive features for the purpose of recognition after preprocessing. Feature extraction is the process of selecting useful and relevant key information with regard to the voice signal by eliminating redundant and irrelevant information [64]. Widely used speech features for auditory modeling are cepstral coefficients, which can be parameterized by LPC, PLP, MFCC, DWT, LSF, and DWT. MFCC, LPC, and PLP are the most frequently used feature extraction and parametric representation techniques in the area of speech recognition and speaker verification applications.

a) MFCC: MFCC is often used for feature extraction in speech recognition based on a frequency that reflects human perception of speech. In the field of speech processing, MFC represents the short-term power spectrum of a voice, which is the result of the linear cosine transform of the logarithmic power spectrum of a voice on a nonlinear Mel scale of the frequency. MFCC is an acoustic analysis method based on auditory filters, which extracts cepstrum parameters in the Mel scale frequency domain and collectively make up an MFC [65]. The Mel scale describes the nonlinear characteristics of human ear frequency, which perfectly simulates the processing of human auditory perception [66]. In addition, these coefficients are robust and reliable under variations in speakers and recording conditions [67]. Dong and Yao [48] used MFCC to explore unique biometric features for mmWave-based speaker verification. MFCC is utilized in research [60] to indicate the resonance properties in the vocal tract system. In addition, Sami et al. [57] and Hao et al. [68] introduced the MFCC features as the identification characteristics.

b) LPC: The linear predictive analysis simulates the human phonation principle and is based on the assumption that the vocal tract dominates the attribute of the sound being produced. LPC is the essential low-rate speech coding method and the basic idea is the sampling of a speech can be approximated by a linear combination of several previous speech samples. LPC not only has the prediction function but also is the most powerful method for estimating a number of basic parameters and computational patterns of speech. Unlike MFCC, the linear filters of LPC imitate the fundamental structure of a vocal tract and how to produce a sound [66]. Linear prediction cepstral coefficient (LPCC) is LPC represented in the cepstrum domain. LPCC [69] depicts the frequency characteristics of speech signals and is used to demonstrate the speech signals by a finite number of signal measurements. The LPC coding is used to obtain LPCC coefficients. The main disadvantage of LPCC is its poor description of consonants and its high sensitivity to quantization noise. To reflect more signal features, Xu et al. [60] also selected LPC besides MFCC to characterize the resonance frequency as vocal tract features for speaker recognition. Liu et al. [56] extracted LPC features as remarkable differences to identify wake-up words.

c) PLP: An alternative to the MFCC is the use of PLP coefficients. The PLP model is [70] developed by Hermansky. PLP can be used for speech feature extraction based on the principle of the human ears' auditory system with the physiological and psychological characteristics of speech signals. PLP combines critical frequency bands, intensity loudness compression, and equal loudness preemphasis to eliminate the influence of irrelevant information from speech signals and improve the performance of speech recognition. PLP is used as a set of coefficients of prediction polynomial of all pole models for the vocal tract to obtain a more approximate auditory spectrum based on speech linear prediction analysis [71].

In addition to the traditional feature extraction methods, some researchers also proposed multicluster/class feature selection [47] scheme, CIR [62], to extract representative features for recognition. Furthermore, deep learning-based feature extraction methods are increasingly widely used with the advantages of processing large-scale data, complex tasks, and automatic feature learning. Shen et al. [52] designed an RFMic-PhoneNet for automatic and effective feature

extraction and fusion. Ozturk et al. [62] utilized a dual-path RNN (DPRNN) to preprocess and combine radio and audio modalities for speech enhancement and separation based on mmWave.

4) Prediction and Recognition: Voice prediction and recognition models are dependent on the feature extraction as well as on the task to be dealt with. From the perspective of technology development, voice print prediction and recognition technology has gone through three major stages.

The first stage method starts from the voice print recognition technology based on template matching technology, which is a nonparametric model. In this method, the training characteristic parameter is compared with the test characteristic parameter, and the distortion between them is regarded as the similarity. Based on the difference of signal alignment, it usually requires the same content of speech to be registered and recognized, which is text correlation and therefore has strong limitations [72]. Template matching technology for voice print recognition usually includes VQ [73] and DTW [47], [74] algorithm. VQ is a lossy data compression method. According to the correlation between adjacent voxels, the key operation in the VQ is to select a finite number of vectors to represent a large vector space based on the principle of block coding, the degree of distortion is the criterion for decision. DTW is a nonlinear regularization technique, which combines time regularization with distance measurement to find alignments between two similar time series with different lengths by optimal path matching. DTW is mainly used in the field of speech recognition to detect the similarity between two speech signals.

The second stage is voice print recognition technology based on statistical models or statistical machine learning. A key to understanding human speech signal processing is how to build a model that reflects the dynamic characterization of its sequential pattern, and acoustic modeling based on HMMs, GMMs, GMM-SVM, JFA, and GMM-UBM. has been proposed in many studies [75], [76], [77]. In order to recognize a speaker, we need to extract the features of the target speaker's voice, train them into one or more models, and store them in the model library. When we are in voice print recognition experience, we extract the features in the currently received voice compare them with the models in the model library, and finally confirm who is the speaker of the current voice. Hao et al. [68] proposed a voice security verification method WMHS, which is a combination of weighted MFCCs and hog-based SVM. To highlight the advantages of CNNs used in the system, Sami et al. [57] evaluated the eavesdropping via Lidar sensors with both the traditional SVM and four-layer CNN architecture. The results show SVM does not exhibit adequate discriminatory capabilities for specific applications involving the recognition of digital signals with low SNR conditions. Moreover, the voice print recognition mechanism depends on the biometrics characteristics instead of exploring the acoustic features of the audio signals. Biometrics characteristics are influenced by both the physical structure of an individual's vocal tract and the behavioral characteristics of an individual's articulatory organs (e.g., lips/mouth motions and throat vibrations) [78].

The third stage is developed into a deep learning framework for voice print recognition. With the rapid development of deep neural network technology, voice sensing technology has gradually exploited the deep neural network framework, such as DNN [60], [79], RNN [80], and the latest end-to-end system [81], [82], [83]. Voice print signal is a nonstationary random signal whose formatting and sensing process is a complex signal processing, the deep model for voice signal processing is more appropriate for light model [40]. The benefits of deep learning have attracted more attention in the voice print signal processing field. Fan et al. [55] exploited the Cycle-GAN network to recover the high frequency from the fundamental frequency of the speech signal collected by mmWave sensing. Hu et al. [84] proposed cGAN network removes noise and reconstructs the high-quality audio from speaker vibrations signal captured by mmWave radar. RFMicPhoneNet [52] based on GCRN and LSTM is proposed to fuse the multichannel signals from microphone signals and vibration signals and achieve high-quality sound recovery. An end-to-end voice print recognition solution [85] aims to train a machine to convert voice print signal to recognition objective by directly learning the mapping from original signal input to associated labeled objective through a deep learning algorithm. The resulting model is then able to recognize voice print with no further algorithmic components, by reducing manual preprocessing and postprocessing. Traditional voice print recognition systems contain a much more complicated architecture that includes signal acquisition, feature extraction, acoustic modeling, and a variety of other algorithmic techniques in order to be accurate and effective. This makes the training, testing, and code complexity of the system far more difficult than an end-to-end structure-based network. On the whole, an end-to-end solution will reduce the complexity of building a voice print recognition system significantly.

B. Human Voice Sensing Technology

Numerous researches have been published in the world on the technology of voice signal detection using wireless sensing. The laser radars, Wi-Fi, UWB, and mmWave, for this kind of wireless sensing technology of various applications have been investigated in enterprises, universities, and scientific research institutes. The above studies demonstrated the feasibility and effectiveness of the wireless sensor in the exploration of the AW signal or the vibration signal from both the human vocal folds and the glottal structure dynamics. The theory and operating principle of wireless voice sensing are based on the propagation equations of EMWs. The theoretical principle of detecting voice print signals through EMW fields is the same for light radar, lidar, and mmWave radar, however, their difference only lies in the frequency range. Tables III and IV show the research overview on human voice sensing through RF technologies.

1) Based on Lidar: One method to achieve a transformation from sound detection to source motion sensing is a laser. The coherent signal source of laser beams and their very short wavelength enables fine-grained distance measurement, which collects laser signals reflected from vibrating objects

TABLE III
HUMAN VOICE SENSING THROUGH RF TECHNOLOGIES I

Literature	Hardware Type	Freq. Band	Biometrics	Algorithmic Techniques	Range	Applications	Performance
Wang [47]	Wi-Fi CSI	2.4GHz	Mouth movements	Wavelet packet decomposition dynamic time warping (DTW), Context-Based Error Correction	$2m$	Lip reading	Average accuracy of 91% for single individual speaking no more than six words and up to 74% for no more than three people talking simultaneously
Wei [87]	Wi-Fi RSSI	2.485GHz	Loudspeaker	Acoustic-radio transformation (ART) algorithm	1-4m	Acoustic eavesdropping	Almost 100% accuracy when the distance is less than 1m, more than 80% for up to 4m, Peak Signal to Noise Ratio
Wang [63]	UWB	7.29GHz	Sound source	Phase Noise Correction, Static Clutter Suppression, and Vibration Activity Localization algorithm	0.3-9.87m	Extraction and separation of audio vibration	SNR, maximum field of view (FOV) is 50 degrees
Sami [57]	LiDAR	/	Speaker/ Soundbar	Supervised learning techniques	150cm	Audio eavesdropping	Approximately 91% and 90% average accuracy of digit and music classifications
Xu [41]	Millimeter wave	24GHz	Vocal cords	Novel deep neural network(Wave-voice Net)	0.5-2m	Speech sensing	Sensitivity, SSNR (Speaking-Signal-to-Noise Ratio), MCD (Mel-Cepstral Distortion), WER (Word Error Rate)
Hong [42]	Millimeter wave	24GHz	Vocal folds	Variational Mode Decomposition (VMD)	40cm	Vocal folds vibration detection	Relative error is below 10%
Li [38]	Millimeter wave	34.5GHz	Throat	Noise estimation and time-scale adaptation algorithm, wavelet transform denoising methods	2-30m	Speech acquisition	Mean opinion score (MOS) is about 2.3
Jiao [88]	Millimeter wave	34.5GHz	Speakers	Spectral subtraction algorithm, Coherence analysis	4m	Detect speech signals	Mean opinion score (MOS) is 4.4 $\pm$ 0.16
Fan [55]	Millimeter wave	60GHz	Vocal cord vibration and lip motion	Cycle-GAN network	$2m$	Speech Recognition	Average speech recognition accuracy is 92.8%
Dong [48]	Millimeter wave	77GHz	Vocal cord vibration and lip motion	Deep convolutional neural network(CNN)	$2m$	Speaker verification	Accuracy over 95% and replay attacks detection accuracy over 98% and the EER at around 3%
Ozturk [43] Radiomic	Millimeter wave	77GHz	Vocal cord/throat	Radio acoustics neural network	2-4m	Sound detection and sound recovery	4m with 91% mean accuracy and 2m with 70% accuracy, classify the sources with 95% accuracy with 40ms of data, which increases to 99.2% by increasing to 320ms
Li [44]	Millimeter wave	77GHz	Vocal cord and vocal tract	Body Motion Compensation, Clutter removal	20cm	Secure authentication	96% authentication accuracy 98.9% BAC, and 96.8% F-score
Hu [87]	Millimeter wave	77GHz	Speakers	Vibration Extraction methods and cGAN	1-5m	Acoustic eavesdropping	Achieved 8.9% distance average relative error and 9.6% angle average relative error
Liu [56]	Millimeter wave	77GHz	Motion of subjects' mouth and throat, audio signals from a microphone	Voice activity detection method, Noise-resistant Multi-modal attention-based network	1-10m	Speech Recognition	Recognition error rate below 1% in a range of 7 meters
Ozturk [62] Radioses	Millimeter wave	77GHz	Vocal folds	Multimodal deep learning	/	Speech enhancement and separation	SISDR is 15.4, PESQ is 2.83 , STOI is 0.94, SIR is 23.6 with 2-person clean mixtures
Zhang [54]	Millimeter wave	77GHz	Vibrations of the surrounding objects	End-to-end network	/	Voice recognition	A word recognition accuracy 87.21% in NLOS, recognition error 35.1%
Li [50]	Millimeter wave	77GHz	Full-field tiny vibrations targets	Multitarget adaptive fusion enhancement (MAFE) and adaptive chirp mode decomposition (ACMD) method	1.5m	Sound recovery	log-likelihood ratio (LLR), linear predictive coding cepstrum distance (LPC-CD), and the coherence speech intelligibility index (CSII)

to measure subtle motions or vibrations. This instrument, also known as a laser microphone, is mainly considered for the purpose of audio surveillance and eavesdropping through walls and window glass. Shine a laser beam on the objects placed

TABLE IV
HUMAN VOICE SENSING THROUGH RF TECHNOLOGIES II

Literature	Hardware Type	Freq. Band	Biometrics	Algorithmic Techniques	Range	Applications	Performance
Chen [59]	Millimeter wave	94GHz	Human Vocal Folds	Empirical mode decomposition (EMD) and the auto-correlation function (ACF)	1-10m	Voice activity detection	Coherence value
Li [58]	Millimeter wave	94GHz	Throat	Short-time phase-spectrum-compensation algorithm	$2m$	Speech acquisition	Mean opinion score(MOS)
Rodriguez [61]	Millimeter wave	125GHz	Vocal folds	Phase imbalance correction technique	$1m$	Speech signal detection	95nm sinusoidal movement

close to the sound source and capture those induced vibrations to recover the audio information. The laser microphone earliest originated from Léon Theremin of the Soviet Union in later 1940s [86], the basic technique or idea is to leverage a laser light beam for remotely recording sound. Based on the principle of laser-based sensing, Sami et al. [57] designed and implemented LidarPhone for speech- and sound-based attacks. LidarPhone enabled opponents to obtain privacy-sensitive voice information through a laser-based microphone that can recover sounds from subtle vibrating objects located near source objects. Laser microphones required the opponents to aim at the highly reflective surface to achieve high-intensity reflections. Performance shows up to 91% digital classification accuracy and 90% music classification accuracy.

2) Based on Wi-Fi and UWB: With the prevalence of Wi-Fi and IoT technology, almost all electronic devices in home/office environments interconnect wirelessly, many studies have shown the success of exploiting Wi-Fi signals to sense human activities. A series of new research and emerging applications have accelerated the rapid advancement of wireless sensing technologies. By detecting and analyzing RF signal reflection, WiHear [47] enabled Wi-Fi to "hear" people talk within the radio range without setting up any acoustic sensors. They proposed the mouth motion profiles as feature extraction. Measured by PHY layer CSI, WiHear realized lip reading and speech recognition by machine learning after filtering out-band interference and partially eliminating multipath. Wei et al. [87] investigated a new acoustic eavesdropping method through Wi-Fi vibrometry that can penetrate traditional soundproofing isolators. The key idea is the acoustic-radio transformation (ART) algorithm, which recovers and strengthens the loudspeaker's sound by detecting the reflective radio signals disturbed by the loudspeakers. ART algorithm models the process of audio vibration interfering with radio waves based on RSS and signal phase information. UWHear [63] exploited impulse radio UWB (IR-UWB) radar to capture and separate multiple sound sources and vibrations of household appliances. The advantages of IR-UWB sensing lie in it can monitor multiple sound sources of audio vibrations simultaneously and conduct a thorough-wall NLOS sensing in near-field environments, and it can effectively recover and separate sounds from multiple sources merely 25 cm apart.

3) Based on mmWave: With 5G commercial further promoting the development of mmWave hardware and software, mmWave radar sensor is broadly used in the measurement of the vibration signal from both the human vocal folds and the glottal structure dynamics.

a) Single modality: To overcome the microphone speech acquisition shortcoming, Jiao et al. and Li et al. ([38], [88], and [58]) proposed a novel speech sensor based on 34.5 and 94 GHz mmWave radar for speech signal acquisition, they proposed a two-step indirect-conversion transceiver, so as to avoid the severe dc offset problem and the associated $1/f$ noise at the baseband. In particular, Li et al. [58] separated the transmitting and receiving circuits and designed two antennas, the improvement of radar hardware can minimize interference from other directions and increase the detection range. Objective and subjective speech quality evaluation show that 34.5 and 94 GHz radar systems both have a better voice recognition ability than the microphone. The traditional devices of speech signal acquisition are prone to pollution by the high background noise and voice interference, Hong et al. [42] presented a nonacoustic method that designed a highly integrated 24-GHz portable auditory radar system to detect the human vocal folds vibration, then a variational mode decomposition (VMD) based algorithm is introduced to decompose the radar-detected auditory signal and extract the time-varying vocal folds vibration frequency. To overcome the challenge of weak source separation, limited range, and multiple side-channel attacks, RadioMic [43], a mmWave-based sound sensing system, was proposed to capture and reconstruct sound under different scenarios. A radio acoustics model related to radio signals and acoustic signals was introduced to reduce the effect of noise, as well as radio acoustics neural network was designed to solve the extremely ill-posed high-frequency reconstruction problem. Hao et al. [68] proposed a mmWave radar-based text-independent voice security authentication system by exploring VCV, this system is highly robust and can prevent playback attacks. After in-depth research on the basic principles of sound generation and vocal vibrations, Xu et al. [41] proposed a WaveEar, a noise-resistant speech sensing system. WaveEar located the speaker's position among multiple people and targeted the mmWave signal toward the speaker's near throat area to sense VCVs. A new deep neural network was proposed for recovering voice through comprehensive feature extraction. Experimental evaluations in real-world scenarios indicate that WaveEar not only appreciably reduces noise-resistant voice, but also achieves ubiquitous VUI in modern electronic equipment. Li et al. [44] presented a resilient mmWave interrogation

system called vocalprint. Vocalprint directly captured and analyzed the unique disturbance of the skin reflection RF signals around the near-throat area for speaker authentication. An innovative resilience-aware clutter suppression algorithm was proposed to extract the vocal tract and vocal source features, authentication accuracy of vocal print exceeds 96% even under adverse conditions. Acoustic eavesdropping brings significant security and privacy risks, MILLIEAR [84] created a highly effective acoustic eavesdropping attack system. This system can precisely reconstruct the voice of unconstrained vocabulary with varying distances, angles, and different types of soundproofing material by leveraging the high-resolution range of FMCW radar and cGAN models. Wavesdropper [46] presented a solution for word detection of human speech through walls from vocal vibration signals captured by commercial mmWave sensors.

b) Multimodality: Single modal signal for voice print recognition is vulnerable in complex scenarios or under impostor attacks, some researchers consider leveraging the redundancy and complementarity of multimodal fusion to enhance the robustness of system identification. Different from exploring VCV signals using noncontact mmWave sensing, Dong and Yao [48] utilized the radar to capture both VCV and lip motion as multimodal biometrics to enhance the system robustness against imposter attacks for identifying speakers. By continuously sensing the liveness of the speakers using mmWave detecting and deep convolutional networks, the proposed approach can achieve high verification accuracy and stronger robustness. Another multimodal speech recognition system is proposed by Wavoice [56], Wavoice integrated two different voice sensing modalities, namely, mmWave radar signals and audio signals from a microphone. Based on building the inherent correlation model between mmWave and audio signals, Wavoice improved the real-time noise-resistant voice activity detection and user localization with multiple speakers, and a fusion attention mechanism is proposed to refine characteristics and fuse multimodal features for accurate speech recognition. Speech enhancement and speech separation are cocktail party problems in vocalprint recognition, so Ozturk et al. [62] investigated a multimodal deep learning approach using mmWave radio devices, together with a microphone to measure vocal folds vibration. By evaluating the joint audio-radio speech enhancement, separating clean and noisy mixtures using mmWave sensing, the RADIOSES system indicated significant improvements in the performance of existing audio-only methods. Fan et al. [55] also exploited lip motion and vocal-cords vibration as multimodal features for speech recognition.

c) Hardware improvements: To achieve higher sensitivity for voice print recognition, some researchers focus on the improvements of radar hardware systems. Zhao et al. [45] presented a portable 24-GHz auditory radar system with a pair of 4×4 antenna arrays which can accurately derive high-resolution time-varying speech information, and the developed radar achieved the advantages of background noise suppression and directional discrimination compared with microphone-recorded experiment. The phase and amplitude imbalance of the I/Q channel due to circuit nonidealities will degrade the receiver performance at last, Rodriguez and Li [61] proposed a new phase imbalance correction method based on cross correlation and ac coupling distortion compensation on a 125-GHz radar for voice information sensing and micro-motion detection. The system sensitivity comes up to measure 95 nm movements. Chen et al. [59] detected the vibration signals of human vocal folds under the use of a 94 GHz mmWave radar sensor. A comprehensive signal processing method that merges EMD with the auto-correlation function (ACF) method is presented for suppressing the noise of the radar-detected signal and extracting the vibration feature.

d) Indirect vibration: However, some researchers found that mmWave-based voice sensing from the near-throat region may not be achieved under NLOS scenarios when human moves or obstacles exist between the radar and the target. Therefore, Zhang et al. [54] proposed AmbiEar which can complete indirect sensing of the human's voice by detecting the vibration of the surrounding objects and tackles the serious problems in NLOS and dynamic scenarios. To utilize the abundant time–frequency information of full field vibrating objects to improve sound recovering, [50] offered a novel concept mmPhone, which is ambient vibration measurement using mmWave signals multitarget adaptive fusion enhancement (MAFE) method to achieve high-quality sound recovery.

C. Hardware Components and System Architecture

To deal with the challenges of low SNR levels, I/Q mismatch, and noise, researchers aimed to optimize the hardware architecture or algorithm design of the wireless sensor system to improve wireless sensing system performance. Depending on the integration degree, the hardware system of wireless sensors for voice sensing can be categorized into two types: off-the-shelf and self-assembly.

1) Off-the-Shelf Sensors: To facilitate customers to develop radar applications, many radar manufacturers introduce radar evaluation modules. The radar evaluation module is programmable and allows developers or researchers to conveniently develop radar applications. Wang et al. [47] implemented the on-off-the-shelf Wi-Fi devices under NLOS and LOS. To monitor multiple sound sources simultaneously, UWHear [63] utilized a commercial IR-UWB radar board for multiple sources recovery and separation at the same time. Sami et al. [57] utilized the lidar sensors equipped on popular commodity robot vacuum cleaners for sensing sounds. For mmWave, The frequency band of the radar evaluation module is concentrated in 24, 60, and 77 GHz, and the number of antennas is generally 1×1 (1 Tx and 1 Rx), 1×2 and 2×4. The 77 GHz radar the most popular frequency range has a bandwidth of 4 GHz (76–86 GHz), and many countries have allocated this band as a dedicated band for automobiles. Compared to 24 GHz, the 77 GHz radar evaluation module has high bandwidth and integrates 2 Tx and 4 Rx for higher measurement accuracy. References [43], [44], [46], [48], [56], and [84] deployed the 77 GHz radar evaluation module for voice sensing. With the application of the evaluation module, developers do not need to focus on the radar hardware and can quickly build an algorithm verification platform by simply configuring the RF front end. A mature commercial radar

evaluation module is the best choice if the focus is on the development of radar signal processing algorithms to solve the difficulties.

2) Self-Assembly Sensors: Due to the limitations of software algorithms for voice sensing, many scholars have developed new radar architectures for research purposes. Xu et al. [41] and Hong et al. [42] designed a 24 GHz mmWave probe with the RF board, the baseband board, and the MCU board. The high-speed difference amplifiers for suppressing noise are settled on the Rx side. The experiments showed that the system can precisely reconstruct speech in noisy environments. Jiao et al. [88] designed a high-sensitivity radar speech sensor operating at 34.5 GHz for speech signal detection. I/Q mismatch will bring signal deterioration which affects the radar sensitivity, Rodriguez and Li [61] proposed a new phase imbalance correction algorithm for a 125-GHz radar to detect speech. The hardware algorithm of cross correlation-based imbalance correction and ac coupling distortion compensation restrain the distortion issue and recover successfully the voice information from throat vibration. Chen et al. [59] utilized an assembled 94-GHz mmWave radar with a 16-channel A/D converter and two separate antennas to improve the detection range and the impact of interference. In addition, Wei et al. [87] presented a basic ART algorithm and integrated it on a WARP software radio testbed with a custom-built FPGA core and 2.485 GHz modulated carrier.

V. Applications of Wireless Voice Sensing

Wireless voice sensing has broad applications in many fields, especially in the area of the IoT, including voice interfaces, health acoustic sensing, sound event monitoring in smart homes and buildings, security authentication, and interactive communication.

A. Speech Recognition and Processing

Speech recognition and processing is the basic application of wireless voice sensing. Speech recognition technology is used to identify spoken words and transform them into readable text based on voice print biometrics. Liu et al. [56] fused mmWave and audio signals to enhance the speech recognition capacity with lower character and word error rates. Ozturk et al. [43], [62] reconstructed the sound from tiny vibrations of object surfaces near the sound source. Zhao et al. [45] implemented speech sensing by collecting the tiny vibration of the human throat.

B. Human–Computer Interaction

With the development of voice recognition and information technology, VUI has vastly improved the modes humans interact with computers through voice or speech commands. Some studies have shown that the current barriers to interaction methods for people with disabilities led to the accelerated introduction of using speech as a means of interaction in HCI [89]. VUI originated from interactive voice response (IVR) in the 1970s. Benefiting from the development of deep learning and natural language processes, the present VUI integrates a variety of artificial intelligence technologies, including speech synthesis and segmentation, automatic speech recognition, and named entity recognition. Humans can interact with computers through natural language, and automatic speech recognition enables VUI to have the capacity for precise understanding of the user's intentions [90]. VUI also plays a crucial role in intelligent scenarios, e.g., smart homes, smart customer services, and robots.

C. Security and Surveillance

Security authentication is one of the most popular biometric applications. Voice print can serve as a biometric feature for speaker representation and identification, and it is a unique voice feature that can distinguish itself from others [91]. The voice print is a distinctive and difficult-to-imitate biological feature. Voice print identification technology has been mainly used in the fields of national defense security, public security technical investigation, judicial correction, and so on, which effectively guarantees national and public security. To enhance home security, Ren et al. [92] proposed a remote authentication system based on voice print for a smart home environment, which can hardly be cheated or cracked according to the characteristics of voice print.

Moreover, banks are using AI-based voice recognition for security function testing which can automatically confirm the identity of a customer for an accurate response such as making payments, transferring money, or credit card payments. Nowadays, online virtual currency payment has become the mainstream way of people's transactions, so the credible identity authentication of online payment is becoming more and more important. Voice print recognition has also been placed more expectations when fingerprint recognition and face recognition have been exposed to a variety of vulnerabilities at the same time.

D. Healthcare and Emotion

Voice recognition in healthcare is an exciting and promising development trend. Researchers found that people with voice disorders may have a relationship with diseases that change an individual's voice with impact on organs (e.g., heart, lungs, brain, muscles, and nervous system) and psychology. References [93], and [60] proposed a DeepVoice system simultaneously combining deep learning and mobile health technology, DeepVoice is a voice print-based Parkinson's Disease identification application system. Similarly, voice print analysis for Parkinson's disease [94] was also studied using MFCC, GMM, and neural networks. Fagherazzi et al. [95] outlined a series of voice applications for health-related purposes. They also summarized present and future applications of vocal biomarkers for healthcare such as cardiometabolic diseases, cardiovascular diseases, voice disorders, and Covid-19.

In the healthcare domain, voice recognition is also being used as a dictation replacement, with physicians leveraging voice recognition technology to streamline their clinical documentation processes. Paperwork and documentation are a vital part of providing quality and proper medical care for patients, in [96], speech-to-text conversion technology is exploited

to improve the documentation process of medical records and reduce the cost and time of recording information. The system helps healthcare providers to improve their productivity regarding paperwork.

Revealing emotions deep inside the speech signals is called speech emotion recognition. Speech emotion recognition can recognize hidden feelings and affective states through tone and pitch, so as to predict pleased, sadness, anger, calm, regret, surprise, fear, neutral, and other emotions. The study [97] analyzed the acoustic characteristics of the audio data and distinguished the underlying motion of the speech and some insights on the human expression of emotion by leveraging machine learning to extract special features.

E. Ambient Assisted Living and Smart Homes

The voice assistant is one of the emerging interactive methods for smart living that developed rapidly in recent years [98]. A voice assistant is a digital assistant that can execute tasks and provide information in response to your questions and commands based on voice recognition, speech synthesis, and natural language processing. Voice assistants can be accessed from a smartphone or built into devices such as smart home appliances, game machines, and even customer assistance, for HCI engaging with intelligent technology using voice. Commodity voice-controlled devices, such as Alibaba's Tmall Genie, Apple's "Hey siri," Google's "OK Google," Xiaomi's Xiaoai classmates, and Microsoft's "Hey Cortana" have integrated speaker identification functions to protect the user information and provide a service through a particular application. XiaoAI classmates support voice print recognition according to the human vocal structure and characteristics based on a microphone array composed of four highly sensitive sensors, after recognizing voice print, Xiao AI will address the user according to the corresponding nickname. The "Hey Siri" [99] detector converts the acoustic pattern of voice at every moment into a probability distribution over speech sounds by employing a deep neural network, then Siri will wake up if the score is high enough after using a temporal integration process to estimate a confidence score of the phrase "Hey Siri." They are deployed as standalone smart devices or integrated within phones, PCs, and specialized equipment (TVs and robot vacuum and mop).

F. Other Emerging Applications

The applications of wireless voice sensing are becoming more and more broadly used in many fields. Speech recognition of the audio files [100] is done to generate the subtitles automatically or automatic translation. Voice hearing aids having improved speech recognition help amplify the sounds in noisy environments for those with hearing loss or help separate the voices of multiple speakers for those with speech disorders. Totakura et al. [101] proposed voice recognition in self-driving autonomous vehicles for the safety of driving, exploiting voice commands to control the car in particular circumstances. Voice sensing can also be applied in the forensic field [102] for crime investigation. Voice recognition solutions in education can also be used to improve pronunciation skills, enhance access for physical disabilities, and assist learning for people with dyslexia and dysgraphia. Based on the idea of integrated sensing and communication, Cui et al. [51] designed a communication solution with mmWave radars which send and receive messages via modulating and decoding the smartphone vibrations.

VI. Discussion and Future Work

With 5G commercial further promoting the construction of the IoT, IoT as one of the development goals, voice print recognition based on wireless sensing as noncontact recognition technology will have a very big application scene. Metrics, privacy, and ethics concerns become the major areas of concern.

A. Metrics for Evaluating Voice Sensing Systems

Voice sensing systems are evaluated based on several key metrics as follows.

1) *Accuracy:* The ability of the system to correctly capture and transcribe voice signals. This is usually measured as a percentage of correctly transcribed words or phrases.
2) *Noise Tolerance:* The system's ability to perform well in various acoustic environments, including those with background noise, is vital.
3) *Robustness:* The system's ability to handle different accents, dialects, and languages, as well as variations in voice due to illness or emotional state.
4) *Latency:* The delay between voice input and the system's response or transcription. Lower latency is desirable for real-time applications.
5) *Energy Efficiency:* Energy efficiency is particularly important for battery-powered devices, this metric refers to how much power the voice sensing system consumes.

B. Privacy Concerns in Wireless Voice Sensing

Wireless voice sensing technologies pose significant privacy concerns in real-world applications as follows.

1) *Eavesdropping:* Unauthorized access to voice data can lead to the exposure of sensitive information.
2) *Data Protection:* Collected wireless voice data must be securely stored and transmitted to prevent unauthorized access.
3) *Anonymity:* In certain applications, the identity of the speaker (e.g., gender and age) must be protected. To protect privacy by removing the sensitive information.

C. Ethical Implications of Using RF Technology

RF technology, while offering many benefits, also raises ethical considerations as follows.

1) *Health Concerns:* While RF technology is generally considered safe, ongoing exposure to RF radiation has raised health concerns that need further investigation (especially in the case of beam-forming, where the user will be exposed to relatively high-intensity RF signals).
2) *Privacy and Surveillance:* RF technology can be used to track individuals or capture data without their knowledge

or consent, raising significant privacy and surveillance concerns.

3) *Access and Equity:* As RF technology becomes increasingly integrated into daily life, there's a risk of deepening the digital divide for those without access to these technologies (e.g., people who cannot speak, such as vocal fold disease, might use wireless technology as long as they can move the mouth and facial muscles).

D. Potential Impact on Various Industries and Sectors

The use of voice sensing technologies can significantly impact various industries as follows.

1) *Healthcare:* Voice sensing can be used for patient monitoring, telemedicine, and early detection of certain health conditions, such as vocal cord disorders or neurological conditions affecting speech. In addition, it can also acquire other health-related signals, such as vital signs.
2) *Home Automation:* Voice-activated smart home devices have already transformed home automation, allowing users to control various household systems with voice commands. Especially there are multiple sound sources, such as radio, TV, and so on. Wireless voice sensing is particularly robust in these scenarios.
3) *Security:* Voice recognition can be used as a biometric for user authentication, enhancing security in various applications.

Despite the aforementioned research investigations that have demonstrated the powerful ability of wireless sensing-based voice print recognition, there still exist many challenges and open problems that need further exploration in future research.

1) Voice print acquisition method based on wireless sensing. Different voice print acquisition methods may differ in the sampling rate of the sound source, acquisition distance, directional sensitivity, frequency bandwidth, and sensing sensitivity. Traditional microphones or acoustic sensors depend on conduction by air, and acquisition distance is limited or sensitivity is quite poor. Therefore, acquisition methods that do not depend on air conduction or can prevent the shortcomings of the traditional speech acquisition method are necessary. Voice print acquisition using wireless sensing has been recognized as one of the emerging technologies.
2) Voice print recognition is developing toward multiattribute recognition. Voice print not only contains semantic information but also conveys information on language, voice, age, emotion, pathology, psychophysiology, and other paralinguistic attributes. Multiattribute recognition based on voice print has attracted increasing attention since it could provide important information for comprehensive application analysis.
3) Multimodal biometrics for Voice print recognition. Unimodal biometric systems have limitations, such as lack of distinguishability, nonuniversality, weak anticounterfeiting attacks, and unacceptable error rates. Multimodal biometric systems that integrate the features provided by multiple sources of information to improve detection accuracy and robustness have been becoming the most potential technologies for identity recognition.
4) Deep learning for voice print recognition. Like other recognition technologies, voice print recognition is also developing toward deep learning. Voice print recognition technology experienced the stages of template matching, statistical modeling, statistics-based machine learning, and so on. With the rapid development of artificial intelligence techniques, deep learning is also increasingly used in voice print recognition. Compared with traditional methods, the recognition accuracy of the deep learning method is significantly improved.
5) End-to-end recognition system. In recent years, end-to-end voice recognition technology become one of the most popular research hotspots in the field of voice recognition for its ability to solve complex problems. Compared with the traditional voice print recognition system, the end-to-end system has the advantages of simple modeling, excellent performance, streamlined training and decoding process, and small space occupation, which attracted the attention of industry and academics.
6) Exploring new applications of voice print recognition. Compared with fingerprint, face, and other biometric technologies, the application of voice print recognition seems to be relatively narrow by now due to the immaturity of voice technology. Voice print recognition is limited to specific scenarios in specific areas of expertise, such as finance, smart home, government security, and so on. But after technological breakthroughs in ambient noise, temporal variability of sound, and speech channel diversity, voice print recognition is bound to enter a wider application scene from financial security and other professional fields.

VII. Conclusion

Wireless sensing has been increasingly employed in multiple fields for applications because of its noncontact, high precision, and ubiquitous characteristics. The nature of the interaction of RF with vocal organs provides tremendous opportunities in the field of voice print sensing. In this survey, the purpose is to explore the evolution, depth, and comparison of voice print recognition using wireless signals. We review a comprehensive array of emerging applications correlated with voice print using wireless signals, including HCI, security certification, healthcare, and emotion. To provide an encyclopedic view of wireless sensing in the voice print, we explore the principles of voice production and wireless sensing techniques. We further summarize and compare existing research on voice print recognition based on wireless sensing. In addition, we also point out the potential issue of the current wireless-based sensing approaches and show a few challenges for further research.

References

[1] P. Flynn, A. K. Jain, and A. A. Ross, *Handbook of Biometrics*. Berlin, Germany: Springer, 2007.
[2] Reportlinker. (2022). *Global Industry Analysts*. [Online]. Available: https://www.reportlinker.com/p05957509/Global-Biometrics-Industry.html

[3] J. A. Markowitz, "Voice biometrics," *Commun. ACM*, vol. 43, no. 9, pp. 66–73, 2000.

[4] L. G. Kersta, "Voiceprint identification," *Nature*, vol. 196, no. 4861, pp. 1253–1257, 1962.

[5] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[6] Z. Saquib, N. Salam, R. Nair, and N. Pandey, "Voiceprint recognition systems for remote authentication—A survey," *Int. J. Hybrid Inf. Technol.*, vol. 4, no. 2, pp. 79–97, 2011.

[7] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[8] R. M. Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Comput. Elect. Eng.*, vol. 90, Mar. 2021, Art. no. 107005.

[9] J. Li and J. Zhang, "A study of voice print recognition technology," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2021, pp. 1802–1808.

[10] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[11] I. Papastratis. (2021). *Speech Recognition: A Review of the Different Deep Learning Approaches*. [Online]. Available: https://theaisummer.com/

[12] J. Zhu, L. Chen, D. Xu, and W. Zhao, "Backdoor defence for voice print recognition model based on speech enhancement and weight pruning," *IEEE Access*, vol. 10, pp. 114016–114023, 2022.

[13] A. Vijayan, B. M. Mathai, K. Valsalan, R. R. Johnson, L. R. Mathew, and K. Gopakumar, "Throat microphone speech recognition using mfcc," in *Proc. Int. Conf. Netw. Adv. Comput. Technol. (NetACT)*, Jul. 2017, pp. 392–395.

[14] S. Dubey, A. Mahnan, and J. Konczak, "Real-time voice activity detection using neck-mounted accelerometers for controlling a wearable vibration device to treat speech impairment," in *Proc. Design Med. Devices Conf.*, 2020, pp. 1–6.

[15] Y. Zhou, Y. Chen, Y. Ma, and H. Liu, "A real-time dual-microphone speech enhancement algorithm assisted by bone conduction sensor," *Sensors*, vol. 20, no. 18, p. 5050, Sep. 2020.

[16] M. M. Kabir, M. F. Mridha, J. Shin, I. Jahan, and A. Q. Ohi, "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," *IEEE Access*, vol. 9, pp. 79236–79263, 2021.

[17] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Netw.*, vol. 140, pp. 65–99, Aug. 2021.

[18] A. A. Anthony and C. M. Patil, "Speech emotion recognition systems: A comprehensive review on different methodologies," *Wireless Pers. Commun.*, vol. 130, pp. 515–525, Mar. 2023.

[19] J. Zhang et al., "A survey of mmWave-based human sensing: Technology, platforms and applications," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2052–2087, 4th Quart., 2023.

[20] Y. Han, S. Li, Y. Cao, Q. Ma, and M. Yoshikawa, "Voice-indistinguishability: Protecting voiceprint in privacy-preserving speech data release," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[21] S. A. Borrie, M. J. McAuliffe, J. M. Liss, G. A. O'Beirne, and T. J. Anderson, "The role of linguistic and indexical information in improved recognition of dysarthric speech," *J. Acoust. Soc. Amer.*, vol. 133, no. 1, pp. 474–482, 2013.

[22] Z. Zhang, "Mechanics of human voice production and control," *J. Acoust. Soc. Amer.*, vol. 140, no. 4, pp. 2614–2635, 2016.

[23] (2021). *Anatomy and Physiology of Voice Production*. [Online]. Available: https://startsingingtoday.com/physiology-of-voice-production/

[24] I. R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 1902–1915, 2008.

[25] R. Wang et al., "DeepSonar: Towards effective and robust detection of AI-synthesized fake voices," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1207–1216.

[26] Y. Ren, Z. Fang, D. Liu, and C. Chen, "Replay attack detection based on distortion by loudspeaker for voice authentication," *Multimedia Tools Appl.*, vol. 78, no. 7, pp. 8383–8396, 2019.

[27] *IEEE Standard Methods and Equipment for Measuring the Transmission Characteristics of Analog Voice Frequency Circuits*, IEEE Standard 743-1984, 1984, pp. 1–52.

[28] H. Türkmen, M. S. J. Solaija, A. Tusha, and H. Arslan, "Wireless sensing-enabler of future wireless technologies," *Turkish J. Elect. Eng. Comput. Sci.*, vol. 29, no. 1, pp. 1–17, 2021.

[29] Z. Yang, Z. Zhou, and Y. Liu, "From RSSI to CSI: Indoor localization via channel response," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 1–32, 2013.

[30] Z.-W. Li, "Millimeter wave radar for detecting the speech signal applications," *Int. J. Infr. Millim. Waves*, vol. 17, no. 12, pp. 2175–2183, Dec. 1996.

[31] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1629–1645, 3rd Quart., 2019.

[32] *The Fundamentals of Millimeter Wave Radar Sensors*, Texas Instruments, Dallas, TX, USA, 2020.

[33] S. Yushanov. *Scattering of Electromagnetic Waves by Particles*. Accessed: Mar. 27, 2023. [Online]. Available: https://altasimtechnologies.com/scattering-of-electromagnetic-waves-by-particles/

[34] S. Yushanov, J. S. Crompton, and K. C. Koppenhoefer, "Mie scattering of electromagnetic waves," in *Proc. COMSOL Conf.*, vol. 116, 2013.

[35] S. Akhdar and S. K. Jasra, "Voice biometrics distinction between English, French, Arabic and Spanish using sound cleaner filtering and speechpro SIS II analysis for same individual identification in multilingual societies," *J. Emerg. Forensic Sci. Res.*, vol. 5, no. 1, pp. 65–72, 2020.

[36] K. Kumatani et al., "Microphone array processing for distant speech recognition: Towards real-world deployment," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Dec. 2012, pp. 1–10.

[37] A. G. H. Van der Donk, P. R. Scheeper, W. Olthuis, and P. Bergveld, "Modelling of silicon condenser microphones," *Sens. Actuators A, Phys.*, vol. 40, no. 3, pp. 203–216, 1994.

[38] S. Li et al., "A new kind of non-acoustic speech acquisition method based on millimeter waveradar," *Prog. Electromagn. Res.*, vol. 130, pp. 17–40, 2012.

[39] (2019). *A Wearable Vibration Sensor for Accurate Voice Recognition*. [Online]. Available: https://www.sciencedaily.com/releases/2019/06/190624111517.htm

[40] C. Li, Z. Cao, and Y. Liu, "Deep AI enabled ubiquitous wireless sensing: A survey," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–35, 2021.

[41] C. Xu et al., "WaveEar: Exploring a mmWave-based noise-resistant speech sensing for voice-user interface," in *Proc. 17th Annu. Int. Conf. Mobile Syst. Appl. Services*, 2019, pp. 14–26.

[42] H. Hong et al., "Time-varying vocal folds vibration detection using a 24 GHz portable auditory radar," *Sensors*, vol. 16, no. 8, p. 1181, Jul. 2016.

[43] M. Zahid Ozturk, C. Wu, B. Wang, and K. J. Ray Liu, "RadioMic: Sound sensing via mmWave signals," 2021, *arXiv:2108.03164*.

[44] H. Li et al., "VocalPrint: A mmWave-based unmediated vocal sensing system for secure authentication," *IEEE Trans. Mobile Comput.*, vol. 22, no. 1, pp. 589–606, Jan. 2023, doi: 10.1109/TMC.2021.3084971.

[45] H. Zhao, Z. Peng, H. Hong, X. Zhu, and C. Li, "A portable 24-GHz auditory radar for non-contact speech sensing with background noise rejection and directional discrimination," in *IEEE MTT-S Int. Microw. Symp. Dig.*, May 2016, pp. 1–4.

[46] C. Wang, F. Lin, Z. Ba, F. Zhang, W. Xu, and K. Ren, "Wavesdropper: Through-wall word detection of human speech via commercial mmWave devices," *Proc. ACM Interact. Mobile Wearable Ubiquitous Technol.*, vol. 6, no. 2, pp. 1–26, 2022.

[47] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with Wi-Fi!" *IEEE Trans. Mobile Comput.*, vol. 15, no. 11, pp. 2907–2920, Nov. 2016.

[48] Y. Dong and Y.-D. Yao, "Secure mmWave-radar-based speaker verification for IoT smart home," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3500–3511, Mar. 2021.

[49] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative lip-motion features for biometric speaker identification," in *Proc. Int. Conf. Image Process.*, Oct. 2004, pp. 2023–2026.

[50] S. Li, Y. Xiong, P. Zhou, Z. Ren, and Z. Peng, "MmPhone: Sound recovery using millimeter-wave radios with adaptive fusion enhanced vibration sensing," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 8, pp. 4045–4055, Aug. 2022.

[51] K. Cui, Q. Yang, Y. Zheng, and J. Han, "mmRipple: Communicating with mmWave radars through smartphone vibration," in *Proc. 22nd Int. Conf. Inf. Process. Sensor Netw.*, 2023, pp. 149–162.

[52] X. Shen, Y. Xiong, S. Li, and Z. Peng, "RFMic-phone: Robust sound acquisition combining millimeter-wave radar and microphone," *IEEE Sensors Lett.*, vol. 6, no. 11, pp. 1–4, Nov. 2022.

[53] P. Hu et al., "Towards unconstrained vocabulary eavesdropping with mmWave radar using GAN," *IEEE Trans. Mobile Comput.*, vol. 23, no. 1, pp. 941–954, Jan. 2024.

[54] J. Zhang, Y. Zhou, R. Xi, S. Li, J. Guo, and Y. He, "AmbiEar: MmWave based voice recognition in NLoS scenarios," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 3, pp. 1–25, 2022.

[55] L. Fan, L. Xie, X. Lu, Y. Li, C. Wang, and S. Lu, "mmMIC: Multi-modal speech recognition based on mmWave radar," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2023, pp. 1–10.

[56] T. Liu et al., "Wavoice: A noise-resistant multi-modal speech recognition system fusing mmWave and audio signals," in *Proc. 19th ACM Conf. Embedded Netw. Sensor Syst.*, 2021, pp. 97–110.

[57] S. Sami, Y. Dai, S. R. X. Tan, N. Roy, and J. Han, "Spying with your robot vacuum cleaner: Eavesdropping via LiDAR sensors," in *Proc. 18th Conf. Embedded Networked Sensor Syst.*, pp. 354–367, 2020.

[58] S. Li et al., "A 94-GHz millimeter-wave sensor for speech signal acquisition," *Sensors*, vol. 13, no. 11, pp. 14248–14260, 2013.

[59] F. Chen, S. Li, Y. Zhang, and J. Wang, "Detection of the vibration signal from human vocal folds using a 94-GHz millimeter-wave radar," *Sensors*, vol. 17, no. 3, p. 543, Mar. 2017.

[60] Z. Xu, J. Wang, Y. Zhang, and X. He, "Voiceprint recognition of Parkinson patients based on deep learning," 2018, *arXiv:1812.06613*.

[61] D. Rodriguez and C. Li, "Sensitivity and distortion analysis of a 125-GHz interferometry radar for submicrometer motion sensing applications," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 12, pp. 5384–5395, Dec. 2019.

[62] M. Z. Ozturk, C. Wu, B. Wang, M. Wu, and K. J. R. Liu, "RadioSES: MmWave-based audioradio speech enhancement and separation system," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1333–1347, 2023.

[63] Z. Wang, Z. Chen, A. D. Singh, L. Garcia, J. Luo, and M. B. Srivastava, "UWHear: Through-wall extraction and separation of audio vibrations using wireless signals," in *Proc. 18th Conf. Embedded Netw. Sens. Syst.*, 2020, pp. 1–14.

[64] S. Sujiya and E. Chandra, "A review on speaker recognition," *Int. J. Eng. Technol.*, vol. 9, pp. 1592–1598, 2017.

[65] N. Dave, "Feature extraction methods LPC, PLP and MFCC in speech recognition," *Int. J. Advance Res. Eng. Technol.*, vol. 1, no. 6, pp. 1–4, 2013.

[66] M. Labied and A. Belangour, "Automatic speech recognition features extraction techniques: A multi-criteria comparison," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, pp. 177–182, 2021.

[67] M. A. Anusuya and S. K. Katti, "Front end analysis of speech recognition: A review," *Int. J. Speech Technol.*, vol. 14, no. 2, pp. 99–145, 2011.

[68] Z. Hao, J. Peng, X. Dang, H. Yan, and R. Wang, "MmSafe: A voice security verification system based on millimeter-wave radar," *Sensors*, vol. 22, no. 23, p. 9309, Nov. 2022.

[69] U. Bhattacharjee, "A comparative study of LPCC and MFCC features for the recognition of Assamese phonemes," *Int. J. Eng. Res. Technol.*, vol. 2, no. 1, pp. 1–7, 2013.

[70] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[71] V. Z. Këpuska and H. A. Elharati, "Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden Markov model classifier in noisy conditions," *J. Comput. Commun.*, vol. 3, no. 6, p. 1, 2015.

[72] L. Gbadamosi, "Voice recognition system using template matching," *Int. J. Res. Comput. Sci.*, vol. 3, no. 5, pp. 13–17, Sep. 2013.

[73] D. Nagajyothi and P. Siddaiah, "Voice recognition based on vector quantization using LBG," in *Computer Communication, Networking and Internet Security*. Berlin, Germany: Springer, 2017, pp. 503–511.

[74] T. B. Amin and I. Mahmood, "Speech recognition using dynamic time warping," in *Proc. 2nd Int. Conf. Adv. Space Technol.*, Nov. 2008, pp. 74–79.

[75] J. Jangir, B. K. Singh, and M. I. Ali, "Voice identification secure system by statistical model of speech signal using normalization technique," *Int. J. Eng.*, vol. 3, no. 1, pp. 2124–2127, 2014.

[76] A. Garg and P. Sharma, "Survey on acoustic modeling and feature extraction for speech recognition," in *Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2016, pp. 2291–2295.

[77] L. R. Rabiner and B. H. Juang, "Speech recognition: Statistical methods," *Encyclopedia Lang. Linguistics*, pp. 1–18, 2006.

[78] R. O'Neil King, "Speech and voice recognition white paper," Biometrics Res. Group, Toronto, ON, Canada, Tech. Rep., 2014.

[79] J. Zhang, "The algorithm of voiceprint recognition model based DNN-RELIANCE," in *Proc. Int. Conf. Comput. Eng. Appl. (ICCEA)*, Mar. 2020, pp. 250–253.

[80] T. S. Vandhana, S. Srivibhushanaa, K. Sidharth, and C. S. Sanoj, "Automatic speech recognition using recurrent neural network," *Int. J. Eng. Res. Technol.*, vol. 9, no. 8, pp. 777–781, 2020.

[81] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4835–4839.

[82] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," 2018, *arXiv:1805.05826*.

[83] N. Kimura, Z. Su, and T. Saeki, "End-to-end deep learning speech recognition model for silent speech challenge," in *Proc. INTERSPEECH*, 2020, pp. 1025–1026.

[84] P. Hu, Y. Ma, P. S. Santhalingam, P. H. Pathak, and X. Cheng, "MILLIEAR: Millimeter-wave acoustic eavesdropping with unconstrained vocabulary," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2022, pp. 11–20.

[85] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 325–351, 2024.

[86] R. P. Muscatell, "Laser microphone," *J. Acoust. Soc. Amer.*, vol. 79, no. 5, p. 1647, 1986.

[87] T. Wei, S. Wang, A. Zhou, and X. Zhang, "Acoustic eavesdropping through wireless vibrometry," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 130–141.

[88] M. Jiao, G. Lu, X. Jing, S. Li, Y. Li, and J. Wang, "A novel radar sensor for the non-contact detection of speech signals," *Sensors*, vol. 10, no. 5, pp. 4622–4633, May 2010.

[89] P. Dabre, R. Gonsalves, R. Chandvaniya, and A. V. Nimkar, "A framework for system interfacing of voice user interface for personal computers," in *Proc. 3rd Int. Conf. Commun. Syst., Comput. IT Appl. (CSCITA)*, Apr. 2020, pp. 1–6.

[90] D. Yu and L. Deng, *Automatic Speech Recognition*, vol. 1. Berlin, Germany: Springer, 2016.

[91] H. N. M. Shah, "Biometric voice recognition in security system," *Indian J. Sci. Technol.*, vol. 7, no. 1, pp. 104–112, Jan. 2013.

[92] H. Ren, Y. Song, S. Yang, and F. Situ, "Secure smart home: A voiceprint and internet based authentication system for remote accessing," in *Proc. 11th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2016, pp. 247–251.

[93] H. Zhang, A. Wang, D. Li, and W. Xu, "DeepVoice: A voiceprint-based mobile health framework for Parkinson's disease identification," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Informat. (BHI)*, Mar. 2018, pp. 214–217.

[94] S. Dasgupta, K. Harisudha, and S. Masunda, "Voiceprint analysis for Parkinson's disease using MFCC, GMM, and instance based learning and multilayer perceptron," in *Proc. IEEE Int. Conf. Power, Control, Signals Instrum. Eng. (ICPCSI)*, Sep. 2017, pp. 1679–1682.

[95] G. Fagherazzi, A. Fischer, M. Ismael, and V. Despotovic, "Voice for health: The use of vocal biomarkers from research to clinical practice," *Digit. Biomarkers*, vol. 5, no. 1, pp. 78–88, Apr. 2021.

[96] S. Ajami, "Use of speech-to-text technology for documentation by healthcare providers," *Nat. Med. J. India*, vol. 29, no. 3, p. 148, 2016.

[97] M. Wadhwa, A. Gupta, and P. K. Pandey, "Speech emotion recognition (SER) through machine learning," Brillio Technol., Indian Inst. Technol., Kharagpur, Tech. Rep., 2020.

[98] P. Cheng and U. Roedig, "Personal voice assistant security and privacy—A survey," *Proc. IEEE*, vol. 110, no. 4, pp. 476–507, Apr. 2022.

[99] (2017). *Hey Siri: An on-Device DNN-Powered Voice Trigger for Apple's Personal Assistant*. [Online]. Available: https://machinelearning.apple.com/research/hey-siri

[100] A. Mathur, T. Saxena, and R. Krishnamurthi, "Generating subtitles automatically using audio extraction and speech recognition," in *Proc. IEEE Int. Conf. Comput. Intell. Commun. Technol.*, Feb. 2015, pp. 621–626.

[101] V. Totakura, B. R. Vuribindi, and E. M. Reddy, "Improved safety of self-driving car using voice recognition through CNN," in *Proc. IOP Conf., Mater. Sci. Eng.*, 2021, vol. 1022, no. 1, Art. no. 012079.

[102] B. V. K. Babu, D. K. Bhargav, R. K. Sah, L. Regalla, and N. Singh, "Forensic speaker recognition system using machine learning," in *Proc. Int. Conf. Sustain. Comput. Data Commun. Syst. (ICSCDS)*, Mar. 2023, pp. 696–701.

Yingxiao Wu (Member, IEEE) received the M.S. degree in computer applications technology from Hangzhou Dianzi University, Hangzhou, China, in 2003, and the Ph.D. degree in signal and information processing from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2010.

She was a Visiting Scholar with the State University of New York at Buffalo, Buffalo, NY, USA, from 2013 to 2014. Since 2021, she has been a Research Fellow with the Department of Computers, Hangzhou Dianzi University. From 2010 to 2019, she was with Huaxin Consulting Company Ltd., Hangzhou, as a Professoriate Senior Engineer, and from 2019 to 2021, she was with Zhejiang Laboratory, Hangzhou, as a Senior Research Specialist. Her research interests include wireless sensing, pervasive computing, Industrial Internet, and deep learning.

Jianping Han received the B.S. and M.S. degrees from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1990 and 1996, respectively, and the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2009.

He is currently a Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou. His current research interests include data intelligence, blockchain, and image information fusion.

Zhihua Jian received the B.Sc. degree in electrical engineering from Nanchang University, Nanchang, China, in 2001, and the M.Sc. and Ph.D. degrees in signal and information processing from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2004 and 2008, respectively.

He is currently an Associate Professor with the School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China. His research interests include speech signal processing, voice conversion, speech spoofing detection, speaker recognition, and machine learning.

Wenyao Xu (Senior Member, IEEE) received the bachelor's and master's degrees from Zhejiang University, Hangzhou, China, in 2006 and 2008, respectively, and the Ph.D. degree from the University of California at Los Angeles, Los Angeles, CA, USA.

He is currently an Associate Professor at the tenure of the Computer Science and Engineering Department, State University of New York at Buffalo, Amherst, NY, USA. The results of his research interest have been published in peer-reviewed top research venues across multiple disciplines, including computer science conferences (e.g., ACM MobiCom, SenSys, MobiSys, UbiComp, ASPLOS, ISCA, HPCA, Oakland, NDSS, and CCS), biomedical engineering journals (e.g., IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS, and IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS), and medicine journals (e.g., *The Lancet*). To date, his group has authored or coauthored more than 180 peer-reviewed papers. His research interests include exploring novel sensing and computing technologies to build up innovative Internet of Things (IoT) systems for high-impact human-technology applications in the fields of smart health and cyber-security. His inventions have been filed within the U.S. and internationally as patents, and have been licensed to industrial players. His research has been reported in high-impact media outlets, including the Discovery Channel, CNN, *NPR*, and *The Wall Street Journal*.

Dr. Xu is the technical program committee member of numerous conferences in the field of smart health and IoT. To date, his group has been the recipient of nine best paper awards, two best paper nominations, and three international best design awards. He has been the TPC Co-Chair of the IEEE Body Sensor Networks in 2018. He is currently an Associate Editor of IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS.