The Catcher in the Field: A Fieldprint based Spoofing Detection for Text-Independent Speaker Verification

Chen Yan*
Zhejiang University
yanchen@zju.edu.cn

Yan Long*
Zhejiang University
ylong@zju.edu.cn

Xiaoyu Ji†
Zhejiang University
xji@zju.edu.cn

Wenyuan Xu†
Zhejiang University
wyxu@zju.edu.cn

ABSTRACT

Verifying the identity of voice inputs is important as voices are increasingly used for sensitive operations. Traditional methods focus on differentiating individuals via the spectrographic features of voices (e.g., voiceprint), yet cannot cope with spoofing attacks, whereby a malicious attacker synthesizes the voice with almost the same voiceprint of a victim or simply replays it. This paper proposes CaField, a text-independent speaker verification method to detect loudspeaker-based voice spoofing attacks with the goal of achieving two seemingly conflicting requirements: usability and security. The key insight of CaField is to construct "fieldprint" with the acoustic biometrics embedded in sound fields, i.e., a physical field of acoustic energy created as the sound propagates over the air, as analogous to "voiceprint". We find that fieldprints can be distinctive between speakers (either humans or loudspeakers), and thus we may detect the speakers being used for spoofing attacks from the authentic users. Our evaluation on a dataset of 20 people and 8 loudspeakers shows that by relying on two on-board microphones to sample sound fields while users talk to the smartphones, CaField achieves a detection accuracy of 99.16% and an equal error rate (EER) of 0.85% across multiple sessions and various voice inputs. CaField supports low audio sample rates at 8 kHz and is robust to various factors including phone displacement, user posture, recording environment, etc.

CCS CONCEPTS

• **Security and privacy → Mobile and wireless security**.

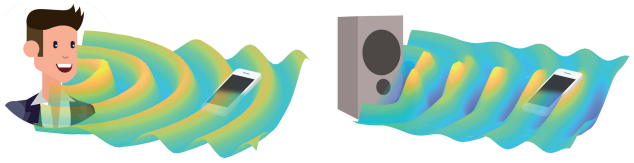
KEYWORDS

fieldprint; speaker verification; spoofing attack; sound field

*Co-first authors.
†Corresponding faculty authors.

CCS '19, November 11–15, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6747-9/19/11…$15.00
https://doi.org/10.1145/3319535.3354248



(a) Authentic user. (b) Spoofing attack.

Figure 1: An illustration of utilizing sound fields to detect loudspeaker-based spoofing attacks with a smartphone. An attacker may use a loudspeaker to generate voices that have almost the same voiceprint as the authentic user's, but it is difficult to replicate the user's sound field in the space.

ACM Reference Format:
Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. 2019. The Catcher in the Field: A Fieldprint based Spoofing Detection for Text-Independent Speaker Verification. In *2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19), November 11–15, 2019, London, United Kingdom.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3319535.3354248

1 INTRODUCTION

Voice inputs have enabled human to vocally instruct smart devices to perform various operations, ranging from making phone calls, posting locations [29], to logging in [59] or even telephone banking [33]. As these operations become increasingly privacy sensitive or security critical, it is important to continuously authenticate who speaks the voice inputs regardless of its content, i.e., text-independent speaker verification. The existing speaker verification mechanisms primarily focus on distinguishing individuals, and typically rely on voice biometrics (e.g., a voiceprint [36]) that is derived from the spectrographic features of utterances. For instance, Siri [6] extracts voiceprints from the wake-up words for owner authentication. Nevertheless, with the help of voice manipulation tools, an attacker can bypass such speaker verification by generating the voice of a victim user via record-and-replay [4, 38, 67], speech synthesis [5, 10, 23], or voice conversion [37, 57, 68], and playing it via a loudspeaker. To cope with such loudspeaker-based spoofing attacks [66], in this paper we propose a user-friendly spoofing detection mechanism (hereafter CaField) that can *continuously authenticate* a voice input. We design CaField to achieve the seemingly conflicting requirements on both usability and security.

Table 1: Similar spoofing detection systems.

System	Source of Distinctiveness	Text Indep.	No Extra Device	Little Pos. Constraint
VoiceLive [71]	Phoneme location	✗	✓	✗
VoiceGesture [70]	Mouth motion	✗	✓	✗
WiVo [43]	Mouth motion	✓	✗	✓
Chetty et al. [21]	Mouth motion	✓	✓	✗
Shang et al. [52]	Throat vibration	✗	✓	✗
VAuth [26]	Body vibration	✓	✗	✓
Chen et al. [20]	Magnetic field	✓	✓	✗
Shiota et al. [55, 56]	Pop noise	✗	✓	✗
VoicePop [63]	Pop noise	✗	✓	✗
2MA [12]	Device ownership	✓	✗	✓
CaField (this work)	Sound field	✓	✓	✓

- **Usability.** `CaField` should impose almost no extra user intervention, and thus it requires *text-independence*, *no extra device*, and *little position constraints*. That is, `CaField` continuously authenticates users as they speak voice inputs; users do not have to carry extra devices [12, 26, 43] other than a smartphone, or being constrained to specific device positions, such as close to the mouth [55, 56, 63, 70, 71], in touch with the throat [52], in front of the face [21], or moving in particular ways [20].
- **Security.** `CaField` shall *distinguish* between an authentic user and spoofing attackers, even if they generate a voice input that has almost the same voiceprint; and the results should be *consistent* across multiple sessions of speaker verification, regardless of the content of the voices, the time of the day, etc.

Achieving security while satisfying all aforementioned requirements on usability is promising yet challenging. The usability requires to record sounds by a single device with little position restrictions, which in turn limits the amount of information available to distinguish between an authentic user and spoofing attackers, especially when the spoofing voices are played by high-quality loudspeakers. Prior work on detecting loudspeaker-based spoofing attacks (as summarized in Tab. 1) shows that it is difficult to achieve all usability requirements, despite that various types of distinctiveness have been explored for security.

To tackle this problem, we utilize the *unexplored acoustic characteristics* that are distinct between an authentic user and attackers regardless of the speech content or device position. The key is to understand the physical process of generating and recording voices in sequential stages: voice production by the user, voice propagation in a medium, and voice receiving and processing at the device. In the voice production stage, the physiological structures of a person, i.e., the shape and size of the vocal cords and vocal tract, determine his/her voice. During propagation, the human voice can be affected by attenuation, ambient noises, diffraction, reverberation, etc., before it is received by a microphone and processed by a speaker verification system. The propagation stage has long been overlooked for its contribution to voice biometrics except for the "harmful" distortion it introduces. However, in this paper, we discover that the voice propagation stage is also affected by physiological features and produces distinctive biometric information, which can be used to differentiate a user from loudspeakers and even from other people. Since sounds propagate in an open space, by nature this stage can be measured with little position constraints. Thus, we may fulfill the requirements on the usability of `CaField` with only one smartphone. Yet achieving text-independence is challenging.

To overcome these challenges, we derive the biometric information that is produced during the voice propagation stage as the "***fieldprint***", analogous to "voiceprints". A fieldprint is extracted from a ***sound field***, which is a physical field of acoustic energy that is created as the sound propagates over the air. As illustrated in Fig. 1, our **key insight** is that the sound field of a speaker (either a human or a loudspeaker) is affected by its physical structure, e.g., physiological features of the mouth, head, and torso, or dimensions of the mechanical components. Sound fields can be distinctive since speakers may have different physical structures, especially between human and loudspeakers. Thus, we may use a fieldprint to verify the identity of a speaker and detect spoofing attacks without relying on liveness information or distorted spoofing sounds.

Based on fieldprints, `CaField`[1] utilizes the on-board microphones of a smartphone to detect spoofing attacks. In particular, `CaField` allows the user to hold the smartphone in any position that he/she is already comfortable with, including but not limited to: next to the ear as making a phone call (the side position) and in front of the chest when interacting with a voice assistant (the front position). Our experiments show that users do not have to hold the smartphone in exactly the same position across multiple sessions, and `CaField` can tolerate modest variance. `CaField` is also fully functional at low audio sample rates such as 8 kHz, which makes it suitable for telephone-based applications of speaker verification. To evaluate its effectiveness, we recruited 20 participants and requested each of them to speak 100 voice commands of their choices in two positions, i.e., in the side and front positions, and we implemented spoofing attacks with 8 loudspeakers of various sizes and qualities. `CaField` is able to accept authentic users and reject spoofing attacks with an accuracy of 99.16% and an equal error rate (EER) of 0.85%, and even distinguish a human participant from others with an accuracy of 98.42% and an EER of 1.84%.

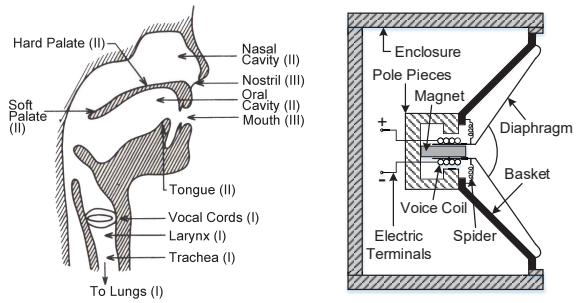
We summarize our major contributions as follows.

- We discover that sound fields can be used to differentiate between authentic users and spoofing attackers, and we design the fieldprint to quantify such distinctions.
- We design `CaField`, a fieldprint-based spoofing detection system that is applicable to text-independent speaker verification without using extra devices and with little position constraints.
- We evaluate the performance of `CaField` based on 20 human participants and 8 loudspeakers that generate 2000 authentic commands and 16000 attempts of spoofing attacks. The results show that `CaField` is highly effective in detecting loudspeaker-based spoofing attacks.

2 BACKGROUND

In this section, we introduce the voice production mechanisms of human and loudspeakers along with the basic concepts of sound fields. To avoid ambiguity, in this paper we refer "speaker" to any type of sound source and use "human (speakers)" and "loudspeakers" respectively to specify biological and electrical sound sources.

[1] `CaField` is short for "the Catcher in the (sound) Field", which is inspired by J. D. Salinger's famous novel "The Catcher in the Rye".



(a) Human speaker. (b) Loudspeaker.

Figure 2: A sectional view of a human vocal tract and a loudspeaker, showing that they rely on distinct sound production mechanisms.

2.1 Human Speakers

The production of human voice normally includes three stages: I) the generation of an initial sound, II) the reshaping of the sound that creates voice, and III) the emission of the voice. As shown in Fig. 2(a), various organs are involved in the three stages, and physiological differences of them between people lead to distinctive human voices.

(I) *Sound generation.* The energy of voice comes from the air expelled from the lungs. In talking, the air flow vibrates the vocal cords in the larynx, which generate an initial sound at a fundamental frequency and its harmonics that are determined by the length and density of the vocal cords. In whispering, the vocal cords do not vibrate but are held close together, which produces a turbulent flow of air that makes a broadband noise-like sound.

(II) *Sound reshaping.* The initial sound is reshaped on the spectrum after going through the vocal tract (the oral or nasal cavities) and becomes a meaningful voice. The vocal tract acts as a "resonator" that enhances or weakens the spectrum of the initial sound at the resonant frequencies that are determined by the geometry of the cavities. By moving organs including the jaw, tongue, teeth, lips, soft palate, etc., a.k.a. the "articulators", one can change the geometry of the vocal tract and pronounce various phonemes[2].

(III) *Voice emission.* Most of the human voices are radiated in the air through the mouth and nose, and some are emitted by other parts of the human body through bone conduction.

2.2 Loudspeakers

A loudspeaker is a device that converts electrical energies into acoustic energies [9]. Its basic function is to respond to the incoming audio frequency electrical signals by performing physical vibrations (e.g., of a diaphragm), which in turn radiate sound waves corresponding to the electrical signals [58]. In a loudspeaker-based spoofing attack, the electrical signals are provided by the attacker with voice manipulation tools or simply via recording. We show an example of loudspeaker among the numerous types of designs in Fig. 2(b). The key observation here is that almost all loudspeakers produce the desired sounds in only one physical stage by directly vibrating a diaphragm [8]. The sounds, determined by the electrical signals and produced by the diaphragm, are radiated in the air off the diaphragm as well. A part of the sounds is also emitted by the enclosure via mechanical conduction.

[2]Phonemes are the smallest units of speech, which are each perceived to be a single distinctive sound in a language. For example, there are 44 unique phonemes in the English language, and a word is pronounced as a combination of phonemes.

2.3 Sound Fields

In physics, a field is any physical quantity which takes on different values at different points in space [27]. Temperature, for example, is a field. For every point $\boldsymbol{p} = (x, y, z)$ in space, we could specify a temperature value as $T(\boldsymbol{p})$, or $T(\boldsymbol{p}, t)$ if the temperature varies in time. The notion of a field has been a practical utility in describing and analyzing various types of physical phenomena, such as magnetic fields, electric fields, velocity fields, etc.

Likewise, the dispersion of acoustic energy over space can be described by a sound field [11]. A sound field describes the sound pressure for every point in space, i.e., the local pressure deviation from the ambient pressure caused by a sound wave [1]. We follow the nomenclature in [1] and formalize a sound field as $s(\boldsymbol{p}, t)$ in the time domain and $S(\boldsymbol{p}, f)$ in the frequency domain, where f is the sound frequency; a lower case s denotes time domain and upper case S denotes frequency domain. $S(\boldsymbol{p}, f)$ is derived by performing a Fourier transform of $s(\boldsymbol{p}, t)$ over a period of time.

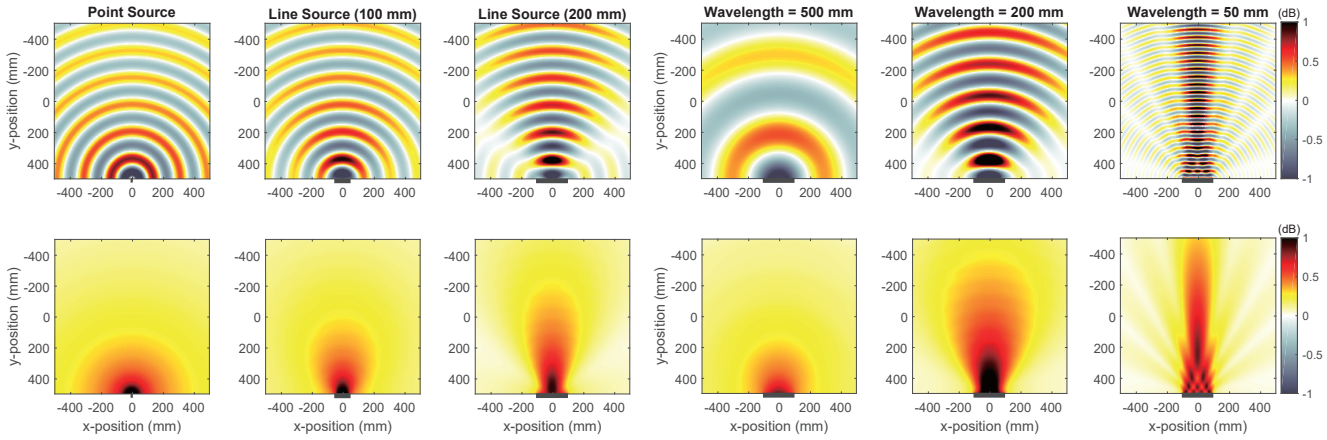
With the concept of a sound field, it becomes easier to understand and model the spatial behaviors of sound propagation. For example, every sound source creates a sound field in the space around it, and the propagation of sounds in the form of sound waves could be modeled by a changing field as if experiencing waves passing through it. The sounds we capture either through our ears or with microphones are the pressure values of a sound field at certain points in space. We investigate hereafter if the difference of voice production mechanisms between human speakers and loudspeakers can reflect distinctiveness in their sound fields.

3 THREAT MODEL

In this paper, we aim to detect the spoofing attacks [66] on text-independent speaker verification systems. An adversary may apply the following three types of attacks:

- *Replay.* An adversary acquires voice samples of the target user through eavesdropping, public speech, etc., and replays the voice samples with loudspeakers [4, 38, 67].
- *Speech synthesis.* An adversary synthesizes utterances in the voice of the target user from text input using speech synthesize technologies and plays with loudspeakers [5, 10, 23, 39].
- *Voice conversion.* An adversary converts a human utterance into the voice of the target user and plays with loudspeakers [37, 57, 68].

As the three major types of spoofing attacks against speaker verification [7], they all rely on loudspeakers to generate the spoofing sounds. Replay attacks are the easiest and most efficient, but an attacker could only generate fake utterances from previously acquired voice samples. In speech synthesis and conversion attacks, an attacker could fabricate any fake utterance in real time, but its similarity to the target user's voice is determined by the performance of synthesis and conversion techniques. Apart from the above loudspeaker-based attacks, there are also human mimicking



(a) Effect of the size of the sound source.

(b) Effect of the sound wavelength.

Figure 3: A 2-D simulation of sound fields in the air (on the 1st row) and their directivity patterns (on the 2nd row). The values of sound fields and their directivity patterns (calculated as the RMS values) are shown in colormaps. The sound sources are modeled as superpositions of point sources centered at coordinate (0, 500) in each plot, and two setups are studied: (a) sources of different size generate a 2 kHz sound; (b) a 200 mm source generates sounds with different wavelengths. It shows that the directivity of a sound field is affected by both the sound source and the generated sound.

attacks, in which a human imitator mimics the voice timbre and prosody of the target user with his/her mouth [32]. Since the threat of human mimicking attacks is not fully understood [66], we do not aim to address such attacks in this paper, even though later we show our mechanism can distinguish a person from other people.

We make the following assumptions on the adversary.

Acoustic attacks. The adversary only attacks speaker verification systems by generating sounds. We do not consider the attack scenario in which the adversary compromises the hardware or OS of the speaker verification device (e.g., a smartphone), or interferes with the device communication. Before the attacks, the adversary may acquire voice samples of the target user through eavesdropping or public resources, but she cannot obtain stereo recordings of the target user in the device positions where he/she uses our system, e.g., by forcing the user. During the attacks, the adversary can be at any location without the user's awareness.

State-of-the-art software and hardware. The adversary may utilize state-of-the-art voice recording, synthesis or conversion software and any types and qualities of microphone and loudspeaker hardware for the attacks.

4 FIELDPRINTS

To exploit sound fields for spoofing detection, we need to answer the following research questions.

RQ1: How can a speaker's sound field reflect its identity?
RQ2: How to utilize a sound field via the extraction of fieldprints and preserve the identity information with desired usability?
RQ3: To what degree do fieldprints show consistency?
RQ4: To what degree do fieldprints show distinctiveness?

We study these questions in the following four subsections. Our answers validate the distinctiveness of sound fields and show fieldprint's advantages on consistency, distinctiveness, and usability, which make it a promising candidate for spoofing detection.

4.1 Sound Field and Speaker Identity

Before investigating the relationship between a speaker's sound field and its identity, we need to characterize a sound field with an abstracted representation that facilitates understanding and comparison. To this end, we follow common practices and describe the shape of a sound field with its *directivity* [1]. Conceptually, directivity refers to how directional the sounds from a sound source are, i.e., the extent to which the sounds are focused into a narrow region in front of the source rather than spread out around it. A sound source that shows no directivity is known as omnidirectional, which only applies to ideal monopole point sources. In practice, all sound sources are directional to some extent. Acoustic directivity is a common experience. For example, when we talk, we subconsciously turn our heads to the listeners, especially when we have to repeat the missed words. This is because the sounds from human are louder in front of the speaker's head than in the back or on the side, i.e., the sound field of a human speaker is front directional.

Our research question then converts to, how does a speaker's identity relate to the directivity of its sound field. For a complete understanding of this issue, we consider all the factors that affect directivity, i.e., the sound source, the sound, and the acoustic environment. Since the reflections and diffractions from the acoustic environment are random interference to sound fields, we evaluate its impact later in Sec. 6. We study the first two factors, the sound source and the sound, separately in the following. To help understand, we visualize the sound fields and their directivity patterns in Fig. 3 with MATLAB simulations using the k-Wave toolbox [60].

Factor 1: how does the sound source affect directivity when the generated sound is the same? An ideal monopole point source radiates sounds from a single point in space. However, in practice, all sound sources exist in measurable sizes. This necessarily suggests that *the same sound could be radiated from different*

points in space at the same time. As a result, the sounds from different parts of a sound source can arrive in phase or out of phase with each other at various locations, and the sound field could be either enhanced or weakened, which in turn creates directivity. Thus, the size of a sound source affects its directivity. For example, in Fig. 3(a), a point source and two line sources that are 100 mm and 200 mm long are simulated to generate the same sound at 2 kHz. It shows that as the size (length) of the source increases, the sound field becomes more directional, which is caused by the increase of out-of-phase cancellation at off-axis angles. In practice, sound sources are much more complex than the line source example. In similar ways, the shape, curve, angle, and material of a source can all affect the directivity pattern beside the size in the 3-D plane.

Remark 1: The directivity of a human is determined by the organs involved in the voice emission stage, i.e., the mouth, nose, and other body parts, while for a loudspeaker it is determined by the vibrating components, e.g., the diaphragm and enclosure. *Thus, it is difficult for loudspeakers to reproduce the human sound fields due to the physical distinctions between electromechanical components and human organs, even if the reproduced sounds are the same.* We will examine this key assumption in the remaining of this paper.

Factor 2: how does the generated sound affect directivity when the sound source is the same? The phase difference of sounds from different parts of a source is neglectable when the wavelength of the generated sound is large with respect to the size of the sound source. However, when the wavelength is comparable to or smaller than the source, the radiation becomes directional despite that the sound source remains unchanged. For example, in Fig. 3(b), a line source that is 200 mm long is simulated to generate sounds of three wavelengths: 500, 200, and 50 mm. Results show that a shorter wavelength (i.e., a higher frequency) leads to higher directivity. When the wavelength is 50 mm (a quarter of the source's length), we observe lobing in the sound field at off-axis angles, which is caused by out-of-phase cancellation.

Remark 2: Considering the influence of wavelength, it is a common practice to specify the sound frequency when quantifying directivity. Since the human voice consists of sounds at various frequencies, we could derive rich directivity patterns if we also measure them at various frequencies.

4.2 Fieldprint Formulation

To utilize the identity information within a sound field, we are motivated to formulate an individually distinctive pattern of voice characteristics from a sound field as a "fieldprint". However, it is impractical, though desirable, to properly sample a sound field over the entire space. Given that nearly all smartphones have at least two microphones (one on the top and one on the bottom), we design a fieldprint based on two sampling locations of a sound field. In particular, we investigate the difference of recordings from the two microphones. Suppose the two microphones are located at $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$, we calculate their difference as the logarithm of the ratio of sound pressure at the two locations:

$$S_R(\boldsymbol{p_1}, \boldsymbol{p_2}, f) = \log \frac{S(\boldsymbol{p_1}, f)}{S(\boldsymbol{p_2}, f)} \tag{1}$$

where $S(\boldsymbol{p}, f)$ is the sound pressure at location $\boldsymbol{p}$ and of frequency f. Though S_R may not fully represent a speaker's directivity, we envision that it can reflect a speaker's distinctive sound field for the following reasons. First, the difference of directivity between speakers may render varying S_R for the same f, $\boldsymbol{p_1}$, and $\boldsymbol{p_2}$. Second, as f changes for a speaker, $S(\boldsymbol{p_1}, f)$ and $S(\boldsymbol{p_2}, f)$ may not change in step due to the change of directivity patterns, which leads to varying S_R at various frequencies. The distinctiveness of a speaker can be increased dramatically when combining the S_R at various frequencies.

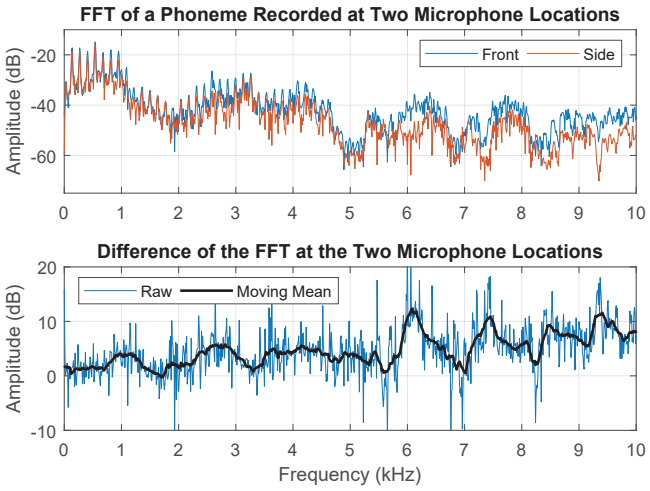


Figure 4: The sound field sampled at two microphone locations (front and side) of a human speaker when pronouncing the phoneme /ō/ (as in "note"), and the derived fieldprint. We show the acoustic energy with amplitude spectra.

Therefore, we formalize a fieldprint as a vector of S_R at various frequencies that are acquired by a smartphone's two microphones:

$$\mathcal{F}(\boldsymbol{p_1}, \boldsymbol{p_2}) = [S_R(\boldsymbol{p_1}, \boldsymbol{p_2}, f_1),\ S_R(\boldsymbol{p_1}, \boldsymbol{p_2}, f_2),\ \ldots\ S_R(\boldsymbol{p_1}, \boldsymbol{p_2}, f_n)] \tag{2}$$

where n is the fieldprint dimension and $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$ are the microphone locations. The choice of frequency f will be discussed in the design of our system. Technically, a fieldprint can be derived by performing Fourier Transform on a stereo recording and calculating the difference of the two channels in the logarithmic scale. To investigate the existence of fieldprints, we simultaneously recorded a person pronouncing a phoneme with two microphones that are 10 cm away in front and on the side of his head. In Fig. 4, we show Fast Fourier Transform (FFT) of the entire pronounced phoneme between 0–10 kHz with a frequency resolution of 10 Hz. The two audio channels are indistinguishable by human ears, which can be confirmed from the resemblance of their FFT plots. However, due to the non-uniform distribution of the sound field and its changing directivity patterns, the difference of their FFT, i.e., the fieldprint, is non-zero and changes with the frequency.

4.3 Consistency of Fieldprint

Since our mechanism relies on a fieldprint to identify a user and detect spoofing attacks, our objective on consistency then depends on to what degree can the fieldprints of a person be consistent between multiple sessions of speaker verification. We study two

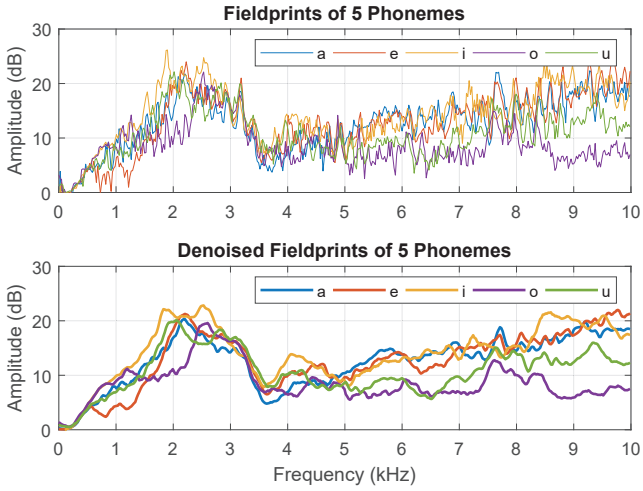


Figure 5: Fieldprints of 5 long vowel phonemes (as in pronouncing the 5 letters "A E I O U"), and the denoised fieldprints for clearer observation. It shows that fieldprints vary for phonemes, which is a challenge for text-independence.

factors that have major impacts on the consistency of a fieldprint: the speech content, and the microphone locations.

4.3.1 Impact of the Speech Content. Humans change the shapes of their articulators to pronounce various phonemes, which may change the sound field and fieldprint. To study the impact of speech content, we measured the fieldprints of 44 English phonemes pronounced by a human speaker. We used a smartphone and kept its location for all phonemes. We show the 5 long vowels as an example in Fig. 5, because their pronunciations take similar amounts of time but require distinct mouth shapes. Although the fieldprints of different phonemes show similar shapes, their difference is obvious especially when the sound frequency is above 500 Hz. Our results agree with a previous study [42] on the directivity of people in pronouncing various phonemes, which also found a sharp deviation above 500 Hz. The impact of speech content poses a challenge for exploiting consistent fieldprints in text-independent scenarios.

Long-Time Average Fieldprint (LTAF). To tackle this challenge, we investigate the average fieldprint of a person when the utterance consists of words and sentences, which contains various types of phonemes. In such a case, the human voice may approach a more *phonetically balanced* state, in which the dynamic influence of individual phonemes may be averaged out and static effects of the sound source become prominent if they occur regularly enough [40, 44]. This also corresponds to the normal usage of speaker verification: a user's identity is verified continuously based on normal voice inputs rather than only a phoneme. Suppose an utterance is divided into m frames that each lasts T in time. A sound field can be stable within a single time frame if T is comparable to the duration of typical phonemes. Assume the fieldprint extracted in the ith frame is $\mathcal{F}_i(\boldsymbol{p}_1, \boldsymbol{p}_2)$, we formalize a *long-time average fieldprint* (LTAF) in the form of

$$\mathcal{F}_{LTA}(\boldsymbol{p}_1, \boldsymbol{p}_2) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{F}_i(\boldsymbol{p}_1, \boldsymbol{p}_2) \tag{3}$$

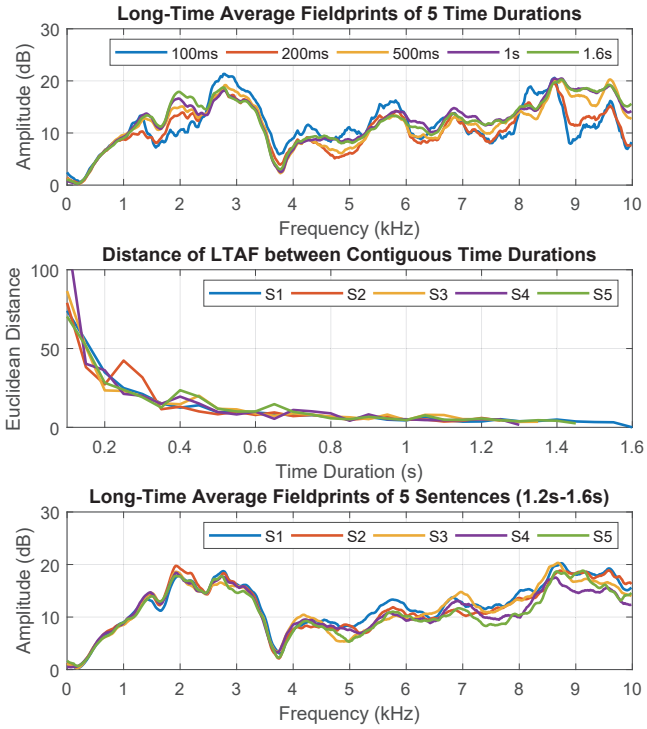


Figure 6: From top to bottom: the long-time average fieldprint (LTAF) of a sentence in its 5 time durations (all starting from the beginning), the Euclidean distance of the LTAF between contiguous time durations for 5 sentences, and the LTAFs of 5 sentences. It shows that an increased time duration renders a more stable LTAF, and different sentences have highly similar LTAFs.

We envision that the LTAF may reduce the influence of individual phonemes, as similar spectrum averaging techniques [40] have been adopted by the acoustics community to observe and compare the acoustical properties of sound sources from highly random speeches. We recorded the human utterances of 10 Harvard sentences [51] listed in Appendix A, which contain specific phonemes at the same frequency they appear in English. These sentences represent the common distribution of English phonemes in normal voice inputs, and they are used for standardized testing of VoIP, cellular, and other telephone systems. The upper plot of Fig. 6 shows the LTAF of the first sentence as the time duration (i.e., N in Eq. 3) increases. Results show that the LTAF converges to a stable pattern as the duration increases, e.g., the LTAFs of 1s and 1.6s duration nearly coincide. To more accurately quantify this effect, we calculate the Euclidean distance of LTAFs between contiguous time durations (increased by 50 ms), i.e., the distance between the LTAFs of duration T and $T - 0.05$ s, as T increases. The results on 5 sentences shown in the middle plot of Fig. 6 validate a stable LTAF when the time duration is approximately above 1 second. The lower plot of Fig. 6 shows that the LTAFs of various sentences are nearly identical (with an average distance of 41), especially below 4 kHz (with an average distance of 17). Thus, we are able to confirm

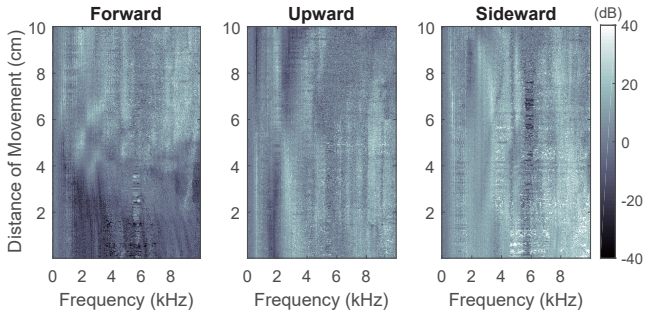


Figure 7: The fieldprints (in colormaps) of a person pronouncing the phoneme /ā/ (as in "late") while moving the recording smartphone in three directions near the speaker's ear: forward, upward, and sideward. It shows that the displacement of microphones can change the fieldprint.

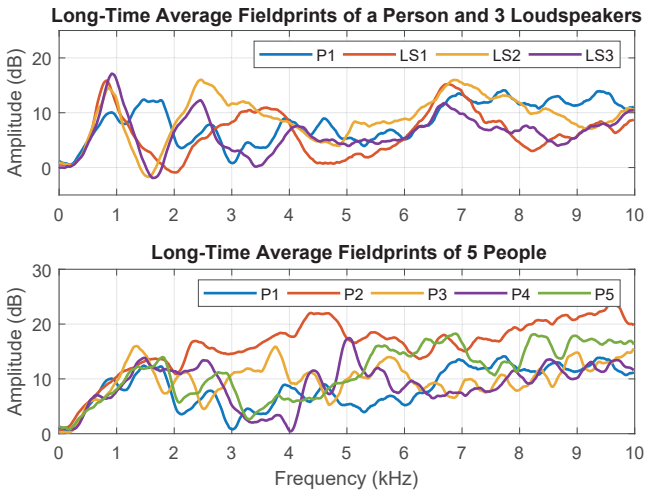


Figure 8: The LTAFs of a person (P1) and 3 replaying loudspeakers (LS1–LS3), and the LTAFs of 5 people (P1–P5). The utterance and microphone locations are the same. It shows that the LTAFs can be distinguished between a person and loudspeakers, and even between different people.

that the long-time average fieldprint meets our requirement on consistency (and usability) for text-independent scenarios.

4.3.2 Impact of the Microphone Locations. The variation of microphone locations between multiple sessions of speaker verification may change the fieldprint. To investigate, we recorded a person continuously pronouncing a phoneme while the smartphone was moved in three directions beside his ear. Fig. 7 shows that a fieldprint is robust to microphone displacements within a few centimeters. However, displacements of more than 4 cm shall be avoided in general, especially in the forward direction. Therefore, users are required to hold the smartphone in the same position for a consistent fieldprint. It is easier for a user to restore the same position by using non-moving parts of his/her body, e.g., an ear or the chest, as the anchor point when holding the device. We will more thoroughly evaluate the robustness and usability of `CaField` in Section 6.4.

4.4 Distinctiveness of Fieldprint

4.4.1 Distinctiveness between a Person and Loudspeakers. Though our simulation suggests that it is difficult for loudspeakers to reproduce human sound fields, the detection of spoofing attacks nonetheless depends on the distinctiveness of fieldprints between an authentic person and loudspeakers. We investigate the feasibility with a proof-of-concept experiment; large-scale experiments will be reported in Sec. 6. We compare the LTAFs of a person (same in Sec. 4.3) saying "show me my messages" and 3 loudspeakers (the first 3 in Tab. 2) replaying it, which all vary in size and shape. We recorded the person and loudspeakers with the same smartphone in similar positions. The upper plot of Fig. 8 shows that the LTAFs of the 3 loudspeakers are very different from the person's. The average Euclidean distance of the replaying loudspeakers' LTAFs to the person's is 106. While in comparison to the results in Sec. 4.3, the average distance is only 41 between the person's own LTAFs even when the speech content varies.

4.4.2 Distinctiveness between People. We also investigate the feasibility of distinguishing a person from other people. Similarly, large-scale experiments will be reported in Sec. 6. We recorded 5 people saying "show me my messages" with the same smartphone in similar positions beside their ears. Their LTAFs in the lower plot of Fig. 8 clearly shows the distinctiveness between people, with an average Euclidean distance of 121. We assume the difference between people's mouth, head, and torso can affect directivity in different frequency ranges and create distinctive fieldprints, as a study has found that directivity varies substantially among singers, and increased directivity could occur through a larger mouth opening, a larger or flatter face, and a larger torso [18].

Effect of the mouth. Since most of the human voice is radiated by the mouth, the mouth opening is equivalent to the major sound source. The maximum mouth opening of an adult is around 5 cm on average, with a range from 3.2 cm to 7.5 cm [69]. Therefore, the mouth opening will mostly affect directivity above 2.3 kHz.

Effect of the head. The human head blocks backward sound radiation and reinforces forward radiation with reflections from the face, which is known as the head shadow effect [46, 48]. For a head width of 17 cm, directivity above 1 kHz will mostly be affected.

Effect of the torso. Similar to the head shadow effect, the human torso provides further reflecting and shadowing effects, with a larger size extending its impact to a lower frequency range.

4.5 Fieldprint Observations

Fieldprint shows the following advantages in spoofing detection:

(1) **Consistency.** The fieldprint of a user is consistent in the form of long-time average fieldprint (LTAF) for varying speech content and with modest microphone displacement.
(2) **Distinctiveness.** The LTAF of a user is distinctive from that of loudspeakers and other people.
(3) **Usability.** The formulation of a fieldprint is only based on two microphones on a smartphone and poses little constraints on the device positions. An LTAF is text-independent and can be used to continuously authenticate the voice input.

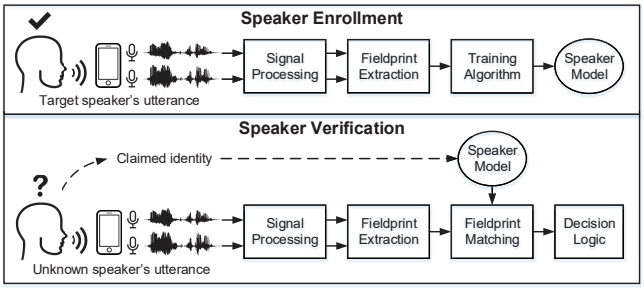


Figure 9: A modular representation of CaField.

5 DESIGN

Powered by the advantages of fieldprints, we design a spoofing detection system named CaField.

5.1 Overview

The core idea of CaField is to detect spoofing attacks by matching the long-time average fieldprint of an utterance with a previously enrolled profile of the claimed identity. CaField consists of four major modules: *Signal Processing*, *Fieldprint Extraction*, *Fieldprint Matching*, and *Decision Logic*. Fig. 9 shows a modular representation of CaField. The upper panel is the enrollment process, and the lower one is the verification process. A user's voice input is recorded simultaneously by two microphones on a smartphone. The two audio channels are first processed in the time and frequency domain to accurately quantify the sound pressure at the two microphone locations. A number of fieldprints are extracted from the processed signals, and an LTAF is then calculated to emphasize a phonetically balanced directivity pattern of the user. A feature vector is further extracted from the LTAF with a filterbank in order to lower its dimension. In the enrollment mode, a speaker model is trained using the feature vectors of a target speaker. In the verification mode, the feature vectors extracted from an unknown speaker are compared with the model of the claimed identity to give a similarity score. The speaker's identity is verified if the similarity score exceeds a predefined threshold, otherwise, it is rejected.

5.2 Signal Processing

Pre-Processing. We first calculate the root mean square of the two audio channels to ensure that the microphone that receives higher acoustic energy is always the $\boldsymbol{p_1}$ in Eq. 1. Then we perform voice activity detection to remove the non-speech portions of the signals, including the silent pauses and breaks within the utterance and transient background noises that appear as pulses. These parts of signals will impair the accuracy of fieldprints because they do not reflect the sound field of the speaker. We edit the two channels at the same time to keep them always synchronous.

Short-Time Fourier Transform. After generating a signal that has little to no pauses between words and phrases, we perform short-time Fourier transform (STFT) analysis to measure the sound field at various time and frequencies. We segment the signals into time frames of 20 ms, with a 50% overlap between successive ones. The 20 ms frame length is shorter in duration than typical phonemes, which allows the signal within a frame to be stationary. The overlap ensures that even the shortest phonemes will have an overall effect on the resulting fieldprints. Before computing the signals' spectrogram, we multiply the signals in each frame with a Hamming window function to reduce the influence of spectral leakage. Though the window function decreases the amplitude of phonemes that happen to occur early or late in the frame, this issue can be remedied by the overlap between frames. To optimize the performance of Fast Fourier Transform (FFT), we set the FFT input length to be the next power of 2 from the original signal length. For example, when the audio sample rate is 48 kHz, a 1024-point FFT is performed on a 20 ms frame which contains 960 data points, resulting in a spectral frequency resolution of 48 kHz/1024 = 46.875 Hz.

After signal processing, we transform the two channels of audio signals into two sequences of amplitude spectrum, which represent a time-varying sound field at the two microphone locations.

5.3 Fieldprint Extraction

Given the two channels' spectra, we extract a sequence of fieldprints following Eq. 1 and 2. It is worth noting that by calculating the logarithm of the ratio in Eq. 1, we also diminish the influence of the speaker's volume. After that, we average them to derive a long-time average fieldprint (LTAF) of the utterance following Eq. 3. The size of an LTAF is dependent on the number of points for FFT and the selected frequency range. For example, with a 48 kHz sample rate and 1024-point FFT, the LTAF will have 513 elements if the frequency range is 0–24 kHz and 215 elements if it is 0–10 kHz.

Considering that CaField is designed for smartphones, we need to save the computational cost with cost-effective fieldprint matching algorithms. Low-dimensional features are thus desirable, because traditional statistical models cannot handle high-dimensional data [50], and the number of required training samples grows exponentially with the number of features [34]. We decrease the dimension of LTAF by multiplying it with a filterbank, which is a series of bandpass filters that each returns the average value in a particular frequency band. We consider different frequencies to have equal effects on the sound field, thus for the filterbank, we adopt rectangular-shaped bandpass filters that are evenly located along the linear frequency axis. If the number of filters in a filterbank is N, the final result is an N-dimensional feature vector.

5.4 Fieldprint Matching

We model the distribution of a speaker's feature vectors with a Gaussian mixture model (GMM) [49], which is a stochastic model composed of a finite mixture of multivariate Gaussian components. GMM is computationally inexpensive and has been the *de facto* method in traditional speaker verification. For an N-dimensional feature vector $\boldsymbol{x}$, the mixture probability density function is:

$$p(\boldsymbol{x}|\lambda) = \sum_{i=1}^{M} w_i p_i(\boldsymbol{x}) \tag{4}$$

where M is the number of Gaussian components, and w_i is the mixing weight of the ith Gaussian component $p_i(\boldsymbol{x})$, which is parameterized by an $N \times 1$ mean vector $\boldsymbol{\mu_i}$ and an $N \times N$ covariance

matrix Σ_i as:

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}|\Sigma_i|^{1/2}} e^{-(1/2)(\boldsymbol{x}-\boldsymbol{\mu_i})'\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu_i})} \quad (5)$$

The mixture weight satisfies $\sum_{i=1}^{M} w_i = 1$. A GMM is trained for a specific user based on the feature vectors extracted in the enrollment using the iterative expectation maximization (EM) algorithm [25]. In the verification mode, feature vectors are extracted from the utterances of an unknown speaker, and their likelihood values are calculated with the trained model of the claimed identity and compared with a predefined threshold.

6 EVALUATION

In this section, we evaluate the performance of CaField on the detection of spoofing attacks, i.e., how well can CaField distinguish an authentic human speaker from loudspeaker-based imposters. We also briefly report the results on distinguishing an authentic human speaker from other people, as a reference to detecting human mimicking attacks in the future. We report the overall performance from the two perspectives and specifically evaluate the factors that affect the performance on spoofing detection. We also study the robustness of CaField in verifying the same user across multiple sessions, i.e., factors that may reduce the consistency of fieldprints and damage usability. Unless otherwise specified, a filterbank of 9 filters ranging from 0 to 4 kHz is used in fieldprint extraction.

6.1 Experiment Methodology

Human Voice Data Collection. We collected the utterances from 20 human participants including 6 females and 14 males.[3] The participants were either undergraduate or graduate students recruited in our institute, whose ages ranged from 18 to 36. They were all informed of the purpose of our experiments and the basic ideas of CaField. Their voices were recorded by a Huawei P10 Plus smartphone while they held it in two types of positions: 1) on the side of their heads as if they were making phone calls, and 2) in front of their heads or chests as they normally talk to voice assistants. We believe the two types of positions represent the most common scenarios of speaker verification on a smartphone. We did not pose any restriction on the exact ways that they held the smartphone, yet we did request them to keep the smartphone in the same position for enrollment and verification. For each of these two types of positions, each participant was requested to say 10 English commands for enrollment and 40 for verification. The participants chose their commands freely from ok-google.io, which provides a random list of voice commands in 25 categories that are commonly used to interact with the Google voice assistant. The participants were free to choose 2 commands from each category, and a total of 100 different commands were collected from each participant in the two positions, amounting to 2000 commands for the human voice dataset. The recording was performed in an office room with background noises such as people talking, walking, and HVAC noises.

[3]We followed the local regulations to protect the rights and welfare of the human participants despite the absence of IRB in our institute.

Table 2: Loudspeakers used for collecting spoofing attacks.

No.	Type	Manuf.	Model	Size (L*W*H in cm)
1	Loudspeaker	HiVi	M-50W	$14.9 \times 15.2 \times 17.6$
2	Loudspeaker	Sonos	PLAY:5	$12.5 \times 36.3 \times 21.6$
3	Loudspeaker	JBL	GO	$3.2 \times 8.5 \times 7.0$
4	Television	Samsung	UA55MUF30Z	$7.5 \times 123.5 \times 71.3$
5	Laptop	Samsung	900X5M	$23.7 \times 34.6 \times 1.5$
6	Smartphone	LG	Nexus 5X	$14.7 \times 7.3 \times 0.7$
7	Smartphone	Motorola	G4 Plus	$15.4 \times 7.7 \times 0.8$
8	Smart speaker	Amazon	Echo	$8.2 \times 8.2 \times 23.4$

Spoofing Attack Data Collection. We implemented spoofing attacks by replaying the above collected human voice[4] with loudspeakers and recording them with the same Huawei smartphone. Since our method is to distinguish a human speaker from loudspeakers and is text-independent, we do not intentionally differentiate replay, speech synthesis, and voice conversion attacks as they all use loudspeakers. We utilized the 8 loudspeakers listed in Tab. 2, which are of various sizes and qualities. Each human-generated recording was replayed by all the loudspeakers and recorded in similar positions, i.e., on the side and in front of the loudspeaker, same in the human voice collection, because we assume an attacker is most likely to fabricate fieldprints similar to those of human speakers in similar positions. There are in total 16000 spoofing attack attempts. Spoofing attacks in random positions and using human-shaped loudspeakers will be discussed later in Section 7.1.

Metrics. We use the following metrics throughout the evaluation. *False Acceptance Rate* (FAR): it characterizes the rate at which an imposter is wrongly accepted by the system and considered as a legitimate user. *False Rejection Rate* (FRR): it characterizes the rate at which a legitimate user is falsely rejected by the system. *Equal Error Rate* (EER): it shows a balanced view of the FAR and FRR and is defined as the rate at which the FAR equals to the FRR. *Accuracy*: it measures the overall probability that the system could accept legitimate users and reject imposters.

Effectiveness of the Spoofing Attack Dataset. To validate the effectiveness of our attacks, we implemented a traditional speaker verification system based on the classic Gaussian Mixture Models. The system is fully functional with our human voice dataset as it could verify authentic users with an average FRR of 0.5% and reject other users with an FAR of 3.3%. With the spoofing attack dataset, the FAR is 65.4%, which corresponds to the performance of spoofing attacks reported in [4, 67] and confirms the effectiveness.

6.2 Overall Performance

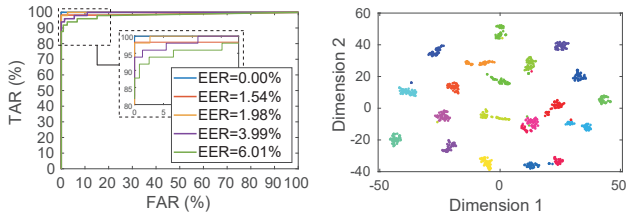
For each participant, we build two models separately for the two smartphone positions. Each model is trained based on the 10 commands in the enrollment, and the rest 40 commands are considered as positive samples in the verification. We summarize the overall performance of CaField in Tab. 3 and discuss in the following.

Distinguishing Authentic Users from Loudspeaker-Based Imposters. For each user model, the negative samples include all the replayed commands of that participant by all loudspeakers. By

[4]We replayed one channel of the stereo recordings that shows better audio quality, as our threat model excludes that the attacker obtains both two channels.

Table 3: The overall performance of CaField.

Distinguish Authentic Users	Accuracy	FAR	FRR	EER
From loudspeaker-based imposters	99.16%	0.82%	0.97%	0.85%
From other human participants	98.42%	1.87%	1.43%	1.84%



(a) ROC curves in spoofing detection. (b) Feature separation of 20 participants.

Figure 10: The ROC curves of five participants in spoofing detection, and the distinctive features of 20 participants (in the side position) after dimension reduction with t-SNE.

averaging the results of the 40 participant models, we derive an overall accuracy of 99.16%, an FAR of 0.82%, an FRR of 0.97%, and an EER of 0.85%. The results show that our system is highly effective in both accepting authentic users and rejecting imposters. We also observe that the performance varies for different participants. For example, Fig. 10(a) shows the receiver operating characteristic (ROC) curves of 5 participants whose EERs are 0.00%, 1.54%, 1.98%, 3.99%, and 6.01% respectively. Though our system outperforms most prior work in Tab. 1, we do not emphasize the comparison since most of them were evaluated with different setups, assumptions, and separate datasets of voices and test commands.

Distinguishing Authentic Users from Other Human Participants. For each user model, the negative samples include all the human voice data from the other 19 participants. By averaging the results of 40 participant models, we get an overall accuracy of 98.42%, an FAR of 1.87%, an FRR of 1.43%, and an EER of 1.84%. In Fig. 10(b), the feature vectors of the 20 participants are visually clustered after dimension reduction with t-Distributed Stochastic Neighbor Embedding (t-SNE) [41]. The results show that with fieldprints we can differentiate different human speakers, and possibly human mimicking attacks as well.

6.3 Factors Affecting Spoofing Detection

We evaluate the factors that may affect the performance on spoofing detection, including the system parameters, smartphone position and distance, loudspeaker type, and recording smartphone.

Impact of System Parameters. In order to find out how the performance on spoofing detection changes with our system parameters, in particular the filter number and frequency range of the filterbank, we conduct a parameter scanning. The number of filters in a filterbank varies from 1 to 16 at an interval of 1. We use 0 Hz as the lower frequency boundary of the filterbank, and the upper boundary varies from 1 kHz to 16 kHz at an interval of 1 kHz. We use the same training and testing samples for all combinations of these two parameters.

Table 4: The impact of phone position on the performance of spoofing detection, showing superiority on the side.

Position	Accuracy	FRR	FAR	EER
Front	98.74%	2.01%	1.16%	1.28%
Side	99.72%	0.63%	0.34%	0.38%

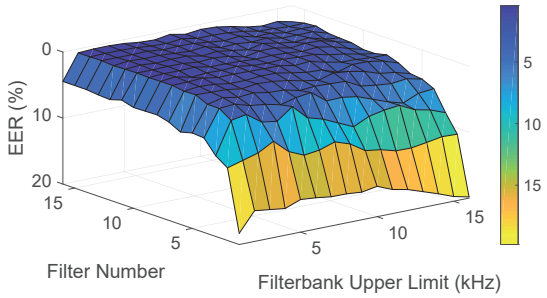


Figure 11: The EERs of spoofing detection under various combinations of system parameters.

As shown in Fig. 11, the performance of CaField generally improves with more filters in the filterbank, which is reasonable as the feature vector represents a more fine-grained fieldprint. In particular, we observe tremendous performance improvement when the number of filters increases from 1 to 4. The performance also increases with the upper frequency boundary of the filterbank, but it slightly worsens when the boundary is above 5 kHz. This in part agrees with our observation in Fig. 6 that the differences of LTAFs between sentences are greater above 4 kHz, whereby the distinctiveness of a person's fieldprints is reduced.

We observe the best performance (i.e., lowest EER) when the filterbank consists of 12 filters under 4 kHz. Although this set of system parameters is not exactly what we use to evaluate the overall performance (9 filters under 4 kHz), their difference in EER is less than 0.2%. This validates the rationality of our model and choice of parameters. The results show another advantage of CaField on supporting *low sample rates*. According to the Nyquist-Shannon sampling theorem [54], a minimum of 8 kHz sample rate is required to correctly sample signals below 4 kHz. Previous microphone-based spoofing detections [70, 71] require sample rates above 48 kHz and recommend 192 kHz.

Impact of Smartphone Position. We compare the performance when the smartphone is held in front or on the side of the participants. The results in Tab. 4 show that holding the smartphone on the side yields slightly better performance: an increase of detection accuracy by nearly 1%, a decrease of FRR by 1.38%, FAR by 0.82%, and EER by 0.9%. We speculate that the reasons are mainly twofold: 1) when held on the side, the smartphone is usually closer to the mouth and thus could extract richer sound field features, and 2) it is easier for users to keep the smartphone in the same position for enrollment and verification when it is held on the side. Nonetheless, both two types of smartphone positions yield EERs less than 2%.

Impact of Smartphone Distance. We study the performance when users hold the smartphone at various distances in front of

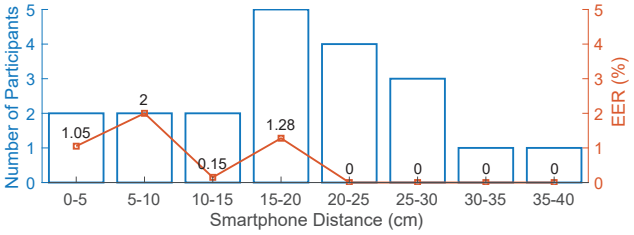


Figure 12: A distribution of the participants' preferred smartphone distances (body-to-microphone) and their impact on the EER of spoofing detection.

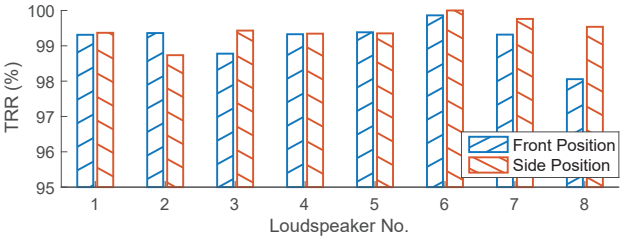


Figure 13: The impact of loudspeaker type on the true rejection rates of spoofing attacks.

them. We show the distribution of their preferred body-to-microphone distances and the corresponding average EERs in Fig. 12. It shows that most participants prefer to hold the smartphone within a distance range of 15–30 cm, yet we do not find an apparent correlation between the distance and performance of spoofing detection.

Impact of Loudspeaker Type. We investigate the performance when the attacker uses different types of loudspeakers. The 8 loudspeakers in Tab. 2 include three traditional loudspeakers, a television, a laptop, two smartphones, and an Amazon Echo. They vary in size and quality and represent the common types of loudspeakers that an attacker may use in various attack scenarios. Among them, the enclosure of the HiVi loudspeaker (No. 1) is closest to a human head by size. We show the true rejection rates (TRR) of these loudspeakers in the front and side positions separately in Fig. 13. It shows that the quality and size of loudspeakers have limited impact on the performance of spoofing detection, which supports our earlier assumption that the sound fields of human speakers and loudspeakers are distinct even if their outer sizes are similar, because they vary in shapes and materials. We also notice that in general, the detection rates in the side position are higher than in the front, which agrees with our earlier results.

Impact of Recording Smartphone. We investigate with two more recording smartphones, a Motorola Moto G4 Plus and an LG Nexus 5X. Both of them yield comparable performance to that of the Huawei smartphone. Notice that `CaField` by design does not allow using one smartphone to enroll and a different one to verify, because the diverse locations, spacing, and models of microphones on a smartphone can lead to distinct fieldprints. This security feature allows `CaField` to authenticate a user and the device at the same time, which reduces the attack surface.

6.4 Robustness and Usability

A system that is not robust to the changing variables between multiple verification sessions will likely have high FRR and bad usability. In the following, we evaluate the robustness of `CaField` to factors that may vary across sessions and applications, including the smartphone displacement, user posture, recording environment, and language of utterance.

Impact of Smartphone Displacement. In view of the location-dependent nature of fieldprints, we evaluate `CaField`'s robustness to smartphone displacement between enrollment and verification. We requested a participant to use our system for 12 consecutive days. In each day, 30 commands were recorded at different times when the user held the smartphone in both the front and side positions. Our system enrolled the participant with only the commands recorded in the first day; it can effectively detect spoofing attacks with an FAR below 1.03% in the side position and 0% in the front position. Though we instructed the participant to restore the smartphone position that was used in enrollment (on the first day), displacement of the smartphone in verification was unavoidable especially as the days last. However, verification results show that the FRRs are all 0% in the side position except for 3 days (5th, 6th, 8th), which are 13.3%, 10.0%, and 3.3% respectively, and the FRRs are all 0% in the front smartphone position. It shows that our system can tolerant modest smartphone displacement with the user's attention. We envision that the impact of smartphone displacement caused by the change of user habits over a longer period of time can be alleviated by updating the user profile regularly.

Impact of User Posture. We evaluate whether `CaField` can still successfully recognize a user if his/her posture changes. We consider the standing and sitting postures since they are most commonly used in speaker verification. We also consider the impact of smartphone positions in these two conditions, because the front position may be more susceptible to the reflections from the lower limbs when seated. The utterances collected from two participants when standing are used for enrollment and those collected when seated are used for verification. The average FRRs in the front and side positions are both 1.02%, which show the robustness to the change of user postures regardless of the smartphone position.

Impact of Recording Environment. We evaluate the robustness of `CaField` when a user enrolls in one place and verifies in another place. We requested two participants to enroll in a meeting room and verify at different places, including an office room, a hallway, an outdoor open space, and a toilet stall. The FRRs are 0.00%, 2.50%, 5.00%, and 18.42% respectively. The results suggest that `CaField` is robust to the change of environment in general, but the reflections and reverberations in a narrow space can affect a speaker's sound field and thus are better avoided.

Impact of Language. In view of the text-independent nature of our system, we evaluate our system's robustness to the language of utterance. We requested a participant to enroll with English commands and verify with Chinese commands, and vice versa. In either condition, we achieve a 100% true acceptance rate. Further observation on the similarity of feature vectors (shown in Appendix B) explains the results and suggests that `CaField` can be used in multilingual applications.

7 DISCUSSION

7.1 Security

Being aware of our mechanism, a motivated attacker may try to elude the detection of CaField with variants of spoofing attacks. We discuss four approaches that are most likely to be exploited and CaField's resistance to them in the following.

Fieldprint Bruteforce Attacks. Although the sound fields of human and loudspeakers are different, it is possible that their fieldprints are similar in different positions. An attacker may try to bruteforce the fieldprint of a user by placing the recording smartphone in random positions around a loudspeaker. To examine the threat of such attacks, we replayed the voice commands of two enrolled users with loudspeaker No. 1, and for each user's voice, we placed the recording smartphone in 100 random positions around the loudspeaker, covering all directions from 1 cm to 1 m away. The recorded voices were used to attack the two user's models (both front and side) trained in Sec. 6. Results showed that the replayed voices were falsely accepted in 3 out of 100 positions (i.e., 3% FAR) for only one user model in the front position. For the other 3 user models, the FARs were all 0%. Though feasible, a successful fieldprint bruteforce attack requires numerous attempts and is easily preventable by setting a limitation for the number of trials.

Sound Field Fabrication Attacks. An attacker may try to fabricate the sound field of a target user with human-shaped loudspeakers, e.g., head and torso simulators (HATS) [17], in hope of a matched fieldprint. HATS are manikins with built-in ear and mouth simulators (microphones and loudspeakers), which are expensive professional acoustic equipment designed for in-situ tests on electroacoustic products. We do not have access to such equipment for an evaluation, but we presume that spoofing attacks with an HATS might fail for two reasons. First, it is difficult to accurately fabricate the sound field of a target user with an HATS. Studies have shown that a standard HATS may not represent the directivity of an average human subject [18, 22, 30]. Even if it does, the personal difference in the sound field makes the attack difficult, as we have shown that a person's fieldprint is distinctive among other people. Second, even if a custom HATS is made in the exact shape of a target user, the material difference between artificial and real heads would cause significant differences among directivity, as found in [14–16].

Stereo Replay Attacks. An attacker may utilize two loudspeakers to replay a stereo recording of the target user to a smartphone's two microphones. In this way, the attacker might control what each microphone receives and directly fabricate the fieldprint. Our threat model excludes such attacks because it is difficult to obtain an authentic stereo recording in the first place. Since a fieldprint is extracted from a pair of locations, it is difficult for an adversary to directly measure a user's sound field at the exact microphone locations where the user uses our system, which are normally within a user's intimate space (< 0.45 m) [31]. It is also difficult to acquire it indirectly from public speeches, eavesdropping on phone calls, etc., because currently, nearly no audio application requires stereo recordings of a user's voice.

Twin Mimicking Attacks. Though we have shown fieldprint's distinctiveness between people, it is interesting to know whether the fieldprints of twins will resemble, since their similarity in the physiological features may lead to similar sound fields. We recruited two pairs of twins (all males), including both identical twins that look similar and fraternal twins that look very different. All participants spoke the same 100 phonetically balanced sentences from [28] in both the front and side device positions, and we requested them to mimic their twin brothers under the same setup. CaField can effectively distinguish the fraternal twins with an EER of 0%, yet to our surprise, it can also distinguish the identical twins with an EER of 2%. We examine their feature vectors and show in Appendix B. Compared with the fraternal twins, the feature vectors between the identical twins show higher similarity in shape (especially in the side device position), but we also observe a slight difference in their distributions, which explains the result. We assume that even for identical twins, distinctive fieldprints exist due to the nonidentical habits in talking and using smartphones. With enough practice, twin mimicking attacks might pose a challenge to CaField as to any speaker verification system, though such attacks are extremely rare.

7.2 Limitations and Future Work

Arbitrary Device Positions Across Sessions. Though we do not specify the positions in which users hold their smartphones, they are required to restore the positions used in enrollment. Authenticating users at arbitrary device positions across sessions is a direction for future work. We envision that in the future, we can investigate the relationship between the smartphone position and fieldprint, and thus adapt our model accordingly.

Long-Range Speaker Verification. In this paper, we focus on speaker verification in the near field, e.g., users need to use our system when the devices are within the reach of their hands. Performing speaker verification from a long range has been a recognized challenge even for traditional speaker verification systems due to the reverberation of sounds [35, 45, 47]. Since fieldprints in the far field are less distinctive, long-range speaker verification with fieldprints is a direction for future work. Nevertheless, our system can still detect spoofing attacks from a long range, because fieldprints acquired in the far field will be distinctive from fieldprints of authentic users acquired in the near field.

Liveness Detection. An alternative approach to detecting spoofing attacks is via liveness, i.e., by detecting the live features of human speakers that loudspeakers do not possess. We envision that a speaker's sound field could also reflect its liveness information through the change of directivity. For example, phonemes that are pronounced with larger mouth openings may show higher directivity than those with smaller mouth openings. Thus, we may detect spoofing attacks if we could correlate the dynamic directivity patterns with the speech content in a predictable manner.

Expanding the Dataset. In this paper, we collected our own dataset on 20 people and 8 loudspeakers as CaField requires stereo recording and consistent device position. Evaluating the robustness of our system with a larger dataset of both extra users and spoofing attacks is a direction of future work.

8 RELATED WORK

The research community has spent significant efforts in protecting speaker verification from spoofing attacks, mainly by seeking for

attack traces in two directions: the acoustic artifacts of attacks, and the liveness distinction between human and loudspeakers.

Acoustic artifacts of spoofing attacks can exist on the physical, hardware, and software levels, including the channel noise [64], far-field recordings [61], frequency response of loudspeakers [13], similarity to stored recordings [53], audio editing [62], etc. Subtle attack traces have been uncovered from features in various domains, including the dynamic speech variability [3], spectro-temporal texture [2], higher orders of Mel-cepstral [19], phase spectrum [65], relative phase shift [24], etc. Spoofing detections in this direction may suffer high false acceptance rate (FAR) when a motivated attacker avoids the artifacts with improved attack procedures, e.g., using high-quality microphones and loudspeakers or more advanced voice synthesis and conversion tools. In such a case, the sounds from an authentic user and an imposter may become indistinguishable in both time and frequency domains.

Researches in the other direction have sought for distinctions on liveness between human and loudspeakers that exist even if acoustic artifacts are minimal. For example, the way that a loudspeaker vibrates its diaphragm is distinct from how a person moves his/her mouth in speaking. A live human can be confirmed by measuring the mouth motion and matching it with the speech content. A number of studies have validated this idea by measuring the mouth motion with a camera [21], or through electromagnetic reflections [43] and ultrasonic reflections [70] off the mouth. Another type of approach is based on the observation that a part of the human voice is also conducted through the body besides airborne transmission, and these two channels of voices share commonalities. Feng et al. [26] proposed to measure the body conduction with a wearable device (e.g., glasses), and Shang et al. [52] proposed to measure the low-frequency vibrations of the throat from a smartphone in contact. Some studies explored other characteristics that are exclusive to loudspeakers or human speakers. Chen et al. [20] proposed to detect by the magnetic fields generated by loudspeakers using the magnetometers on smartphones, and Zhang et al. [71] proposed to measure the time difference of arrival (TDoA) to a smartphone's two microphones, which are diverse for various phonemes because they are produced at different locations in an oral cavity. Shiota et al. [55, 56] and Wang et al. [63] proposed to identify live users by detecting the pop noise recorded in human utterances due to the exhalation, but this requires the microphone to be very close to the mouth. `CaField` exploits the difference in sound fields between authentic users and spoofing attackers while achieving a balance of security and usability. It requires only one smartphone with little constraints on the device position, and it supports continuous authentication of the voice input.

The acoustic community has discovered the distinctive directivity patterns of human sound fields. Halkosaari et al. [30] measured the directivity of artificial mouth simulators (loudspeakers) and a group of human subjects and found a difference greater than 10 dB. They suggested that directivity is affected by the aperture size of the mouth, the upper body, and the speech content. [15, 22] also found evident differences in the directivity patterns between human and loudspeaker in an open space and inside cars. These studies were motivated for tests of acoustic effects in the telecommunication industry and not intended for spoofing detection.

9 CONCLUSION

In this paper, we address the challenge of spoofing detection for speaker verification utilizing the difference of sound fields between authentic users and spoofing attackers. We propose a new feature, the fieldprint, to quantify the difference and design a spoofing detection system named `CaField`. It is highly effective in protecting text-independent speaker verification and requires no extra devices other than a smartphone and little constraints on device positions.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for their constructive feedback. This work was funded by the National Key R&D Program of China under grant number 2018YFB0904900 and 2018YFB0904904.

APPENDIX

A HARVARD SENTENCES — LIST 11

(1) Oak is strong and also gives shade.
(2) Cats and dogs each hate the other.
(3) The pipe began to rust while new.
(4) Open the crate but don't break the glass.
(5) Add the sum to the product of these three.
(6) Thieves who rob friends deserve jail.
(7) The ripe taste of cheese improves with age.
(8) Act on these orders with great speed.
(9) The hog crawled under the high fence.
(10) Move the vat over the hot fire.

B FEATURE VECTORS

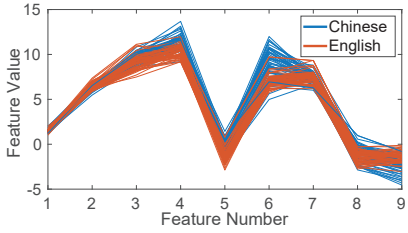
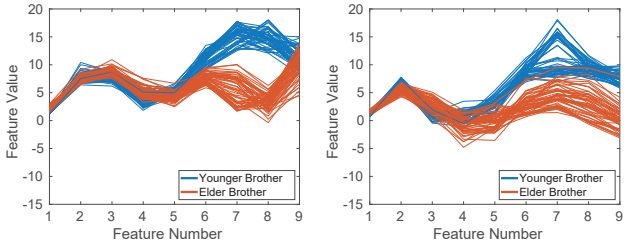
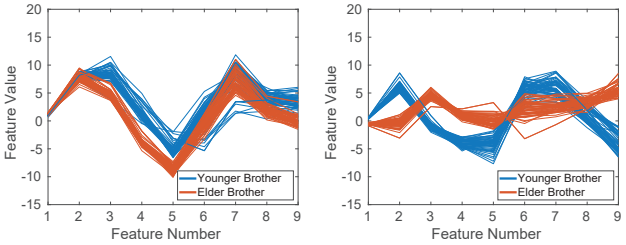


Figure 14: Feature vectors of commands in Chinese and English spoken by the same person. It shows that the language of speech has limited impact on fieldprints.



(a) Side position (b) Front position.

Figure 15: Feature vectors of fraternal twins in the side and front device positions. `CaField` can effectively distinguish the fraternal twins with an EER of 0% in both positions.



(a) Side position (b) Front position.

Figure 16: Feature vectors of identical twins in the side and front device positions. `CaField` can effectively distinguish the identical twins with an EER of 2% in the side position and 0% in the front position.

REFERENCES

[1] Jens Ahrens. 2012. *Analytic methods of sound field synthesis.* Springer Science & Business Media.

[2] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. 2013. A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. In *Proceedings of the 2013 IEEE International Conference on Biometrics: Theory, Applications and Systems*. IEEE, 1–8.

[3] Federico Alegre, Asmaa Amehraye, and Nicholas Evans. 2013. Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3068–3072.

[4] Federico Alegre, Artur Janicki, and Nicholas Evans. 2014. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. In *Proceedings of the 2014 International Conference of the Biometrics Special Interest Group*. IEEE, 1–6.

[5] Federico Alegre, Ravichander Vipperla, Nicholas Evans, and Benoît Fauve. 2012. On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. In *Proceedings of the 20th European Signal Processing Conference*. IEEE, 36–40.

[6] Apple Inc. 2019. Siri - Apple. https://www.apple.com/siri/.

[7] ASVspoof consortium. 2019. ASVspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. http://www.asvspoof.org/.

[8] Glen Ballou. 2012. *Electroacoustic devices: microphones and loudspeakers.* CRC Press.

[9] Glen Ballou. 2015. *Handbook for sound engineers.* Focal Press.

[10] BBC. 2016. Adobe Voco 'Photoshop-for-voice' causes concern. https://www.bbc.com/news/technology-37899902.

[11] Leo Leroy Beranek and Tim Mellow. 2012. *Acoustics: sound fields and transducers.* Academic Press.

[12] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. 2018. 2MA: Verifying voice commands via two microphone authentication. In *Proceedings of the 2018 Asia Conference on Computer and Communications Security*. ACM, 89–100.

[13] Logan Blue, Luis Vargas, and Patrick Traynor. 2018. Hello, Is It Me You're Looking For?: Differentiating Between Human and Electronic Speakers for Voice Interface Security. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM, 123–133.

[14] Fabio Bozzoli, Paolo Bilzi, and Angelo Farina. 2005. Influence of artificial mouth's directivity in determining speech transmission index. In *Proceedings of the 119th Audio Engineering Society Convention*. Audio Engineering Society.

[15] Fabio Bozzoli and Angelo Farina. 2003. Directivity balloons of real and artificial mouth simulators for measurement of the Speech Transmission Index. In *Proceedings of the Audio Engineering Society Convention 115*. Audio Engineering Society.

[16] Fabio Bozzoli, Angelo Farina, and Michel Viktorovitch. 2005. Balloons of directivity of real and artificial mouth used in determining speech transmission index. In *Proceedings of the Audio Engineering Society Convention 118*. Audio Engineering Society.

[17] Brüel & Kjær. 2019. Head and torso simulators, and ear simulators. https://www.bksv.com/en/products/transducers/ear-simulators.

[18] Densil Cabrera, Pamela J Davis, and Anna Connolly. 2011. Long-term horizontal vocal directivity of opera singers: Effects of singing projection and acoustic environment. *Journal of Voice* 25, 6 (2011), e291–e303.

[19] Lian-Wu Chen, Wu Guo, and Li-Rong Dai. 2010. Speaker verification against synthetic speech. In *Proceedings of the 7th International Symposium on Chinese Spoken Language Processing*. IEEE, 309–312.

[20] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *Proceedings of the 37th IEEE International Conference on Distributed Computing Systems*. IEEE, 183–195.

[21] Girija Chetty and Michael Wagner. 2004. Automated lip feature extraction for liveness verification in audio-video authentication. In *Proceedings of Image and Vision Computing*. University of Canterbury, 17–22.

[22] WT Chu and ACC Warnock. 2002. Detailed directivity of sound fields around human talkers. *Technical Report* RR-104 (2002), 1–47.

[23] Phillip L De Leon, Vijendra Raj Apsingekar, Michael Pucher, and Junichi Yamagishi. 2010. Revisiting the security of speaker verification systems against imposture using synthetic speech. In *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1798–1801.

[24] Phillip L De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga. 2012. Evaluation of speaker verification security and detection of HMM-based synthetic speech. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 8 (2012), 2280–2290.

[25] Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39, 1 (1977), 1–22.

[26] Huan Feng, Kassem Fawaz, and Kang G Shin. 2017. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*. ACM, 343–355.

[27] Richard P Feynman, Robert B Leighton, and Matthew Sands. 1977. The Feynman Lectures on Physics: Mainly Electromagnetism and Matter, Vol. 2.

[28] Qian-Jie Fu, Meimei Zhu, and Xiaosong Wang. 2011. Development and validation of the Mandarin speech perception test. *The Journal of the Acoustical Society of America* 129, 6 (2011), EL267–EL273.

[29] Go Crew. 2018. "Ok Google, Open Go Crew" – Voice Command For Location Sharing. https://go-crew.com/ok-google-open-go-crew-voice-command-for-location-sharing/.

[30] Teemu Halkosaari, Markus Vaalgamaa, and Matti Karjalainen. 2005. Directivity of artificial and human speech. *Journal of the Audio Engineering Society* 53, 7/8 (2005), 620–631.

[31] Edward T. Hall. 1990. *The hidden dimension.* Anchor Books.

[32] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, Timo Leino, and Anne-Maria Laukkanen. 2013. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Proceedings of the Interspeech*. Citeseer, 930–934.

[33] HSBC Group. 2018. Voice ID. https://www.bksv.com/en/products/transducers/ear-simulators.

[34] Anil K Jain, Robert P. W. Duin, and Jianchang Mao. 2000. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1 (2000), 4–37.

[35] Q. Jin, T. Schultz, and A. Waibel. 2007. Far-Field Speaker Recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 7 (Sep 2007), 2023–2032. https://doi.org/10.1109/TASL.2007.902876

[36] Lawrence George Kersta. 1962. Voiceprint identification. *The Journal of the Acoustical Society of America* 34, 5 (1962), 725–725.

[37] Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li. 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4401–4404.

[38] Johan Lindberg and Mats Blomberg. 1999. Vulnerability in speaker verification-a study of technical impostor techniques. In *Proceedings of the 6th European Conference on Speech Communication and Technology*.

[39] Johan Lindberg and Mats Blomberg. 1999. Vulnerability in speaker verification-a study of technical impostor techniques. In *Proceedings of the Sixth European Conference on Speech Communication and Technology*.

[40] Anders Löfqvist and Bengt Mandersson. 1987. Long-time average spectrum of speech and voice analysis. *Folia Phoniatrica et Logopaedica* 39, 5 (1987), 221–229.

[41] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[42] AH Marshall and J Meyer. 1985. The directivity and auditory impressions of singers. *Acta Acustica united with Acustica* 58, 3 (1985), 130–140.

[43] Yan Meng, Zichang Wang, Wei Zhang, Peilin Wu, Haojin Zhu, Xiaohui Liang, and Yao Liu. 2018. WiVo: Enhancing the Security of Voice Control System via Wireless Signal in IoT Environment. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*. ACM, 81–90.

[44] Brian Bruce Monson. 2011. High-frequency energy in singing and speech. (2011).

[45] Ladislav Mošner, Pavel Matějka, Ondřej Novotnỳ, and Jan Honza Černockỳ. 2018. Dereverberation and beamforming in far-field speaker recognition. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5254–5258.

[46] J Müller, F Schon, and J Helms. 2002. Speech understanding in quiet and noise in bilateral users of the MED-EL COMBI 40/40+ cochlear implant system. *Ear and Hearing* 23, 3 (2002), 198–206.

[47] Mahesh Kumar Nandwana, Julien van Hout, Mitchell McLaren, Allen Stauffer, Colleen Richey, Aaron Lawson, and Martin Graciarena. 2018. Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings. In *Proceedings of Interspeech*. ISCA, 1106–1110.

[48] Cherish Oberzut and Laurel Olson. 2003. Directionality and the head-shadow effect. *The Hearing Journal* 56, 4 (2003), 56–58.

[49] Douglas A Reynolds. 1995. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication* 17, 1-2 (1995), 91–108.

[50] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. 2000. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing* 10, 1-3 (2000), 19–41.

[51] EH Rothauser. 1969. IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics* 17 (1969), 225–246.

[52] Jiacheng Shang, Si Chen, and Jie Wu. 2018. Defending Against Voice Spoofing: A Robust Software-based Liveness Detection System. In *Proceedings of the 15th IEEE International Conference on Mobile Ad Hoc and Sensor Systems*. IEEE, 28–36.

[53] Wei Shang and Maryhelen Stevenson. 2010. Score normalization in playback attack detection. In *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1678–1681.

[54] Claude Elwood Shannon. 1998. Communication in the presence of noise. *Proceedings of the Institute of Radio Engineers* 86, 2 (1998), 447–457.

[55] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. 2015. Voice liveness detection algorithms based on pop noise caused by human breath for automatic speaker verification. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association*.

[56] Sayaka Shiota, Fernando Villavicencio, Junichi Yamagishi, Nobutaka Ono, Isao Echizen, and Tomoko Matsui. 2016. Voice liveness detection for speaker verification based on a tandem single/double-channel pop noise detector. In *Proceedings of the Odyssey Speaker and Language Recognition Workshop*. 259–263.

[57] Yannis Stylianou. 2009. Voice transformation: a survey. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3585–3588.

[58] Michael Talbot-Smith. 2012. *Audio engineer's reference book*. Focal Press.

[59] Tencent Inc. 2015. Voiceprint: The New WeChat Password. https://blog.wechat.com/tag/voiceprint/.

[60] Bradley E Treeby and Benjamin T Cox. 2010. k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields. *Journal of Biomedical Optics* 15, 2 (2010), 021314.

[61] Jesús Villalba and Eduardo Lleida. 2011. Detecting replay attacks from far-field recordings on speaker verification systems. In *Proceedings of the European Workshop on Biometrics and Identity Management*. Springer, 274–285.

[62] Jesús Villalba and Eduardo Lleida. 2011. Preventing replay attacks on speaker verification systems. In *Proceedings of the 2011 Carnahan Conference on Security Technology*. IEEE, 1–8.

[63] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. 2019. VoicePop: A Pop Noise based Anti-spoofing System for Voice Authentication on Smartphones. In *Proceedings of the IEEE International Conference on Computer Communications*. IEEE, 1–9.

[64] Zhi-Feng Wang, Gang Wei, and Qian-Hua He. 2011. Channel pattern noise based playback attack detection algorithm for speaker recognition. In *Proceedings of the 2011 International conference on machine learning and cybernetics*, Vol. 4. IEEE, 1708–1713.

[65] Zhizheng Wu, Eng Siong Chng, and Haizhou Li. 2012. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association*.

[66] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and countermeasures for speaker verification: A survey. *Speech Communication* 66 (2015), 130–153.

[67] Zhizheng Wu, Sheng Gao, Eng Siong Cling, and Haizhou Li. 2014. A study on replay attack and anti-spoofing for text-dependent speaker verification. In *Proceedings of the Signal and Information Processing Association Annual Summit and Conference*. IEEE, 1–5.

[68] Zhizheng Wu and Haizhou Li. 2013. Voice conversion and spoofing attack on speaker verification systems. In *Proceedings of the 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 1–9.

[69] Khalid H Zawawi, Emad A Al-Badawi, Silvia Lobo Lobo, Marcello Melis, and Noshir R Mehta. 2003. An index for the measurement of normal maximum mouth opening. *Journal of the Canadian Dental Association* 69, 11 (2003), 737–741.

[70] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 57–71.

[71] Linghan Zhang, Sheng Tan, Jie Yang, and Yingying Chen. 2016. Voicelive: A phoneme localization based liveness detection for voice authentication on smartphones. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1080–1091.