Towards Pitch-Insensitive Speaker Verification via Soundfield

Xinfeng Li, Zhicong Zheng, Chen Yan, Chaohao Li, Xiaoyu Ji, and Wenyuan Xu

***Abstract*—Automatic speaker verification systems (ASVs) verify a person's identity by his/her voice and have been widely deployed for user authentication. However, existing ASVs are based on traditional audio spectral features and hence perform poorly in verifying pitch-changed utterances from speakers with cold or sore throat. In this paper, we propose SOFTER (SOundField TrackER), a soundfield-based speaker verification system that can verify speakers regardless of the pitch changes. SOFTER is based on the observation that soundfield features reflect the speaker's vocal tract, mouth, head, torso, etc., which are less affected by the pitch changes in speech signals. SOFTER can be integrated into off-the-shelf smartphones without any hardware modifications. One major challenge is that the soundfield is sensitive to the distance between the speaker and the phone. To solve this problem, we propose a two-stage mechanism combining distance sensing and soundfield reconstruction, which enables to reconstruct the soundfield to a setting similar to the one in the enrollment phase, thus the speaker can be verified from any distance to the phone. We compare SOFTER with 6 state-of-the-art academic and commercial ASVs on two datasets of 134 speakers and 31,000 speech samples. Results show that SOFTER has an equal error rate (EER) of 2.18% and 1.61% on the two datasets, respectively. Moreover, SOFTER outperforms other ASVs by at least 24.67% on average in verifying pitch-varying or pathological speech samples, denoting an evidence of SOFTER's effectiveness in both normal and unhealthy user conditions.**

***Index Terms*—Speaker Verification, Soundfield, Biometrics, Pitch Variation, Pathological Speech.**

I. Introduction

Automatic speaker verification systems (ASVs) authenticate speakers based on their vocal characteristics [1] (i.e., voiceprint) such as pitch and timbre [2]. Compared to the traditional passphrase-based authentication methods, ASV is more convenient as a biometrics mechanism and has been widely used due to its low cost and considerable efficiency for IoT applications. ASVs facilitate people to access various smart applications securely (e.g., APPs login, device unlock) and keep them from malicious intruders while not requiring physical contact. Moreover, taking into account the security and privacy of the users and smart devices, almost off-the-shelf smartphones and smart speakers are integrated with ASV [3]–[5].

However, existing ASVs based on traditional audio spectral features usually perform poorly in verifying pitch-changed utterances, whose extracted voiceprints are different from the registered speaker's in the ASV, thus leading to its misjudgment. Unfortunately, the pitch-variable challenge is inevitable, especially since the nose and throat play essential roles in speech production, which can be significantly affected by the symptoms of a nasal sound or sore throat. Previous works [6], [7] studied the difference between the "cold-affected" and healthy speech in speaker identification. They demonstrated a significant effect of pathology (e.g., nasal congestion, sore throat) on the pitch and found some noisy portions existing when a speaker suffers from hoarseness and coughing. In addition, pathological speech has been demonstrated to introduce a mismatch between registered healthy speaker patterns, resulting in performance degradation [8]. For the sake of robustness in real-world scenarios, ASVs are usually trained with data augmentation techniques. Nevertheless, these strategies cannot tackle the significant performance degradation for speaker identification caused by the pitch changes [9]. Moreover, the pitch modification can even be applied to voice disguise against ASVs [10], [11].

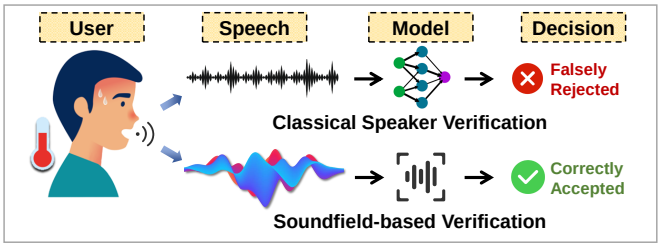


Fig. 1. When a user is in unhealthy conditions, such as suffering from nasal congestion, sore throat, etc., existing automatic speaker verification systems (ASVs) do not authenticate the user as well as they normally do, falsely rejecting the valid user. Nevertheless, soundfield-based verification can perform well regardless of the user's physical condition.

Given the close link between pitch changes and pathological speech, our goal is to implement a pitch-insensitive speaker verification system. As shown in Fig. 1, we envision that such a system can maintain relatively effective in verifying users regardless of their pitch variation and generalize to mitigate the ASVs' performance degradation issue when users are in unhealthy conditions. To achieve the goal, we propose a "**soundfield-based**" approach, inspired by the fact that humans and loudspeakers can be distinguished from the soundfield created via acoustic propagation aspect [12]. Besides, we find that soundfield features of the same speaker are consistent across utterances and are at the same time distinctive between different speakers, which we attribute to the fact that the acoustic propagation forms a soundfield is mainly affected by physiological features such as the speaker's mouth, head, torso, etc., which change little with pitch variation. Based on the same principle, SFF features include multidimensional

X. Li, Z. Zheng, C.Yan (the corresponding author), C. Li, X. Ji, and W. Xu are with the College of Electrical Engineering, Zhejiang University, Hangzhou 310058, China. (Email: xinfengli@zju.edu.cn, zheng_zhicong@zju.edu.cn, yanchen@zju.edu.cn, lchao@zju.edu.cn, xji@zju.edu.cn, wyxu@zju.edu.cn.)

soundfield information and thus can weaken the impact of pitch changes. Our investigation demonstrates two critical properties of soundfield features: *1) Pitch-insensitive: soundfield is more insensitive to pitch variation than monophonic spectral features and can better distinguish speakers in the case of pathological speech compared to the advanced deep learning-based (DL-based) features. 2) Distance-sensitive: the measurement of soundfield is sensitive to user-microphone distance, i.e., the soundfield can differ greatly when the microphone is placed close to the mouth or slightly far away.*

Based on the above properties, we design a system termed SOundField TrackER (hereafter **SOFTER**) that can mitigate the speaker verification performance degradation regardless of pitch variation and overcome the distance-sensitive challenge. Without any dedicated hardware, it leverages the onboard microphones, typically located at the bottom and top of smartphones, to capture the speaker's speech signals propagating in space. We overcome the distance sensitivity by a two-stage soundfield reconstruction mechanism so that users do not have to fix their smartphones in positions the same as the enrollment. Firstly, we design a chirp signal that the onboard loudspeaker of the smartphone can emit to measure the device's distance to the user's mouth. Secondly, we adopt the impulse response to perform the soundfield reconstruction to avoid the performance degradation caused by distance variation, which is regarded as a transfer function between the sound source and microphone. By establishing a distance-oriented impulse response database, we can gain the specific transfer function depending on the measured distance, and the soundfield will be rebuilt to a setting similar to the enrollment stage. Thus users can be verified at different distances.

Unlike classical ASVs, SOFTER does not require training with considerable data. It models *SFF* by the computation-efficient GMM and only requires a few utterances for enrollment to achieve reliable authentication. To comprehensively evaluate its performance, we built two speech datasets named Voice-1 & Voice-2 for pitch-variable and pathological speech experiments, respectively. We also compare SOFTER with 6 state-of-the-art classical ASVs under the same experiment settings. The Voice-1 simulates the pitch changes at five levels of 110 speakers from the VCTK corpus and records their speech samples in a two-channel microphones manner for further soundfield features extraction. The experiments on Voice-1 validate that our system can perform well even in pitch-variable cases, with the EER of 2.18% and maintaining relative EER performance improvements of at least 24.67% than the average of other systems. In addition, we collected Voice-2 containing both the healthy and pathological speech (i.e., consisting of 3 kinds of symptoms: nasal sound, sore throat, or both symptoms simultaneously) of 24 participants. Results positively show that SOFTER presents pronounced advantages over other models in real scenarios, with the EER of 1.61% while the average EER of other ASVs up to 9.49%, and keeps speaker discrimination capability with an EER of 5.04% even when the speakers suffer from severe symptoms.

Our main contributions can be summarized as follows:

- We propose new voice biometrics—soundfield that can improve speaker verification performance in pitch-variable cases, e.g., caused by pathological body conditions. Our investigation validates the feasibility of *SFF* for speaker verification on speakers having three symptoms (nasal sound, sore throat, or both of them).
- We design a speaker verification system called SOFTER, which fully leverages soundfield and overcomes its distance-sensitive challenge. Notably, SOFTER can fit in off-the-shelf smartphones without requiring any hardware modifications. Besides, SOFTER is training-free and can function well with a few utterances to enroll.
- We conduct a comprehensive evaluation on SOFTER and compare it with 6 state-of-the-art classical ASVs based on two datasets we collected. Results demonstrate that our system is effective in verifying speakers regardless of their voice pitch or pathological conditions.

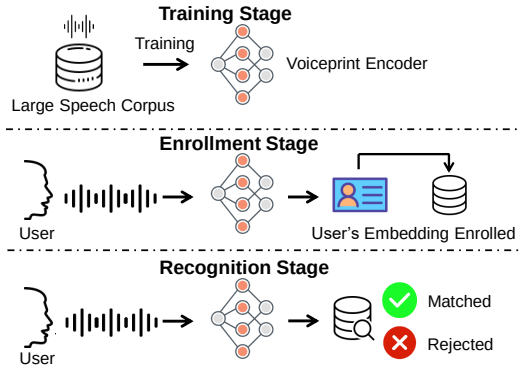


Fig. 2. Classical speaker recognition systems (SRS) consist of training, enrollment, and recognition phase.

II. Background

A. Speaker Recognition Systems (SRSs)

Speaker recognition can be classified into speaker identification/verification. Identification aims to determine from which of the registered speakers a given utterance comes, while verification corresponds to accepting or rejecting the identity claimed by a speaker [13]. SRSs model humans' vocal tract characteristics, generally named "voiceprint", to identify different speakers [1]. As shown in Fig. 2, implementing and utilizing an SRS is usually divided into three phases. Many efforts were devoted to modeling the voiceprint better in the early years. [14] proposed the GMM-UBM, including several representative models named SVM [15] and joint factor analysis [16]. Among the models, the GMM-UBM/I-vectors frontend [17] with probabilistic linear discriminant analysis (PLDA) backend [18], [19] provided state-of-the-art performance for several years. Recently, motivated by the powerful feature extraction capability of deep neural networks (DNNs), many deep learning-based speaker recognition methods were proposed [20]–[22], boosting the better performance of speaker recognition even in complex environments.

B. Speech Production and Sickness Effects

Similar to the fingerprint and face, the voice provides substantial cues that can make listeners distinguish different

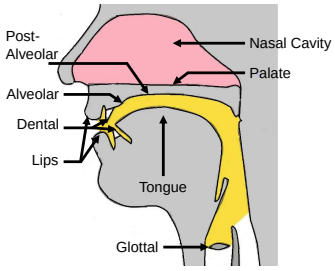


Fig. 3. A sectional view of the human speech production-related apparatus.

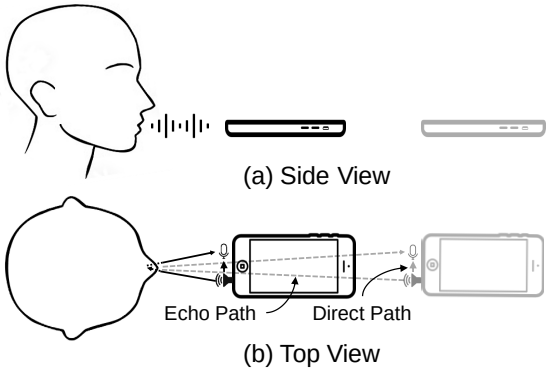


Fig. 4. Schematic of a user using SOFTER. (a): the phone placed horizontally and at the same height as the mouth. (b): the loudspeaker at the bottom of the smartphone returns the range via acoustic signal. The difference between the black and gray color of the phone indicates that the user-phone location can vary.

speakers — the anatomy of each individual's vocal apparatus is distinct [23], as shown in Fig. 3. Specifically, differences in the thickness of the vocal folds, the various shapes of a person's palate or nasal cavity, etc., and the habit of the tongue or vocal tract moves can lead to differences in articulation, accent, and other acoustic properties between speakers. When people are in unhealthy physical conditions, the voice they utter can be affected, named pathological speech. We divide these impacts into two classes: semantic impact and voice impact. Semantic impacts widely exist among those with Alzheimer's disease (AD) and Parkinson's disease (PD). However, our work mainly focuses on voice impacts because it is more common in ASV problems. This kind of illness is usually accompanied by symptoms of nasal congestion or sore throat, which affects speech production, termed "pathological-affected speech". In particular, nasal speech is caused by irregular closure of the soft palate during a cold, resulting in the abnormal resonance of air in the nasal cavity due to too much or too little air passing through the nose. Hoarse voice is caused by laryngitis, an inflammation that leads the vocal cords to swell and exceptionally slows their vibrations down [8].

C. Sound Fields

When the sound is produced and uttered from the mouth, similar to physical phenomena such as the electric field and magnetic field, the sound signal propagates over the air, which forms a soundfield, describing the time-variant sound pressure at each location. Although we envision sampling the entire spatial soundfield to describe speech signals perfectly, it requires multiple sophisticated distributed microphones, which is impractical in most application scenarios. CaField [12] proposed "fieldprint", which utilized a simplified method of acquiring soundfield with two microphones and formulated the soundfield by calculating the logarithm of the ratio of sound pressure between these two microphones:

$$S_r(\mathbf{p}_1, \mathbf{p}_2, f) = log\frac{S(\mathbf{p}_1, f)}{S(\mathbf{p}_2, f)} \tag{1}$$

where $S(\mathbf{p}, f)$ is the sound pressure at location $\mathbf{p}$ and of frequency f by performing a Fast Fourier Transform (FFT) on every frame of each-channel signal. By integrating each frequency component, we obtain Eq. 2, which delivers a reliable liveness detection.

$$\mathcal{F}(\mathbf{p}_1, \mathbf{p}_2) = [S_r(\mathbf{p}_1, \mathbf{p}_2, f_1), ..., S_r(\mathbf{p}_1, \mathbf{p}_2, f_n)] \tag{2}$$

Inspired by fieldprint, we proposes the soundfield-based feature (*SFF*), which holds significant properties such as intra-speaker consistency and inter-speaker distinctiveness to perform speaker verification. We augment it with long-time average normalization, calculated as Eq. 3:

$$SFF(p_1, p_2) = \frac{1}{L}\sum_{i=1}^{L}\mathcal{S}_i(p_1, p_2) \tag{3}$$

where L is the number of frames in the time domain, denoting that *SFF* is downscaled from $n \times L$ to an n-dim vector. The critical difference between the *SFF* and other traditional audio spectral features is that the soundfield not only retains the speaker's voice information but also introduces the identity information as the sound propagation is affected by the nature of the person's mouth, face, head, and torso.

III. Soundfield Investigation

In this section, we explore soundfield more in-depth with three research questions:

RQ1: What is the correlation between soundfield and user-phone distance?

RQ2: How robust is the soundfield to changes in pitch?

RQ3: Whether the soundfield can be generalized from pitch-variable tasks to pathological speaker verification?

First of all, we define the application scenario, a typical placement of a smartphone held horizontally in front of the mouth [24], as shown in Fig. 4. In this way, we can derive a more remarkable soundfield than other placements because the two-channel signals captured by the top and bottom microphones have the most enormous sound propagation path difference. Moreover, much literature has demonstrated that the higher the wave frequency, the more directional the wave emitted from its source [12], [25], [26]. Thus our design also enables avoiding the problem that cannot capture those weak high-frequency components if their directions are away from the central propagation path, ensuring that we get the effective soundfield.

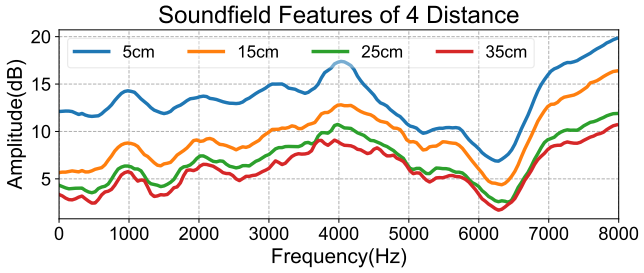


Fig. 5. The *SFF* curves of 4 different distances between the same speaker and microphone, showing the impact of distance.

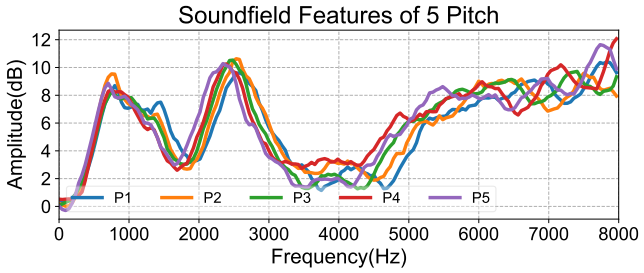


Fig. 6. The *SFF* curves of speech at five pitch levels with the same speaker-phone locations. P1: Descend 1 semitone. P2: Descend 0.5 semitone. P3: origin. P4: Ascend 0.5 semitone. P5: Ascend 1 semitone.

A. Soundfield at Different Distances

We first investigate the relationship between the verification distance and soundfield. As conveyed in black and grey in Fig. 4, when people hold the smartphone, they keep it in the front of their mouths well, yet hard to ensure the fixed user-phone distance. Considering our authentication is specially designed for the near-field application, [27] illustrates that the relationship between pressure and distance is more complex than far-field, implying the distance will significantly affect the soundfield features. We also conduct a preliminary experiment where a speaker utters the same content at 4 distances. Fig.5 shows that soundfield varies with distance, denoting it is sensitive to distance. We analyze the specific reasons as follows: in the near field, the sound energy is divided into *1) energy that directly reaches the microphones after emitting from the mouth*, and *2) energy that circulates back and forth between the mouth and the phone as well as escapes/losses out*, both vary with distance. Given the user-friendly requirements, we need to devise a way to mitigate or even eliminate the non-negligible effect of distance rather than forcing users to hold their devices in a fixed manner, as discussed in Sec. IV.

B. Soundfield of Different Pitch

Since previous works [6]–[8] revealed the significant influence of pitch variation on ASVs, we would like to investigate soundfield's robustness in this case. Specifically, We selected a speaker from the VCTK corpus [28] and simulated his soundfield (described in Sec. V). We refer to the pitch modification methods in [9] by varying the original audio's pitch at 4 levels, within the range of 1 semitone down and up. Then each speech signal propagating over space would be captured by two virtual microphones, saved as stereo

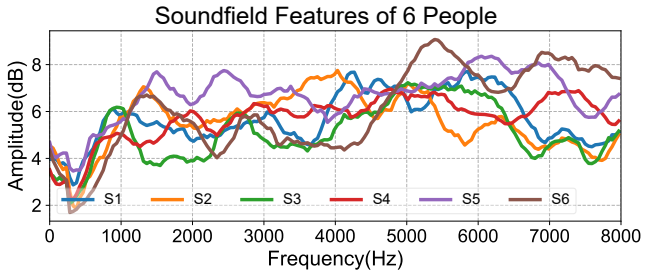
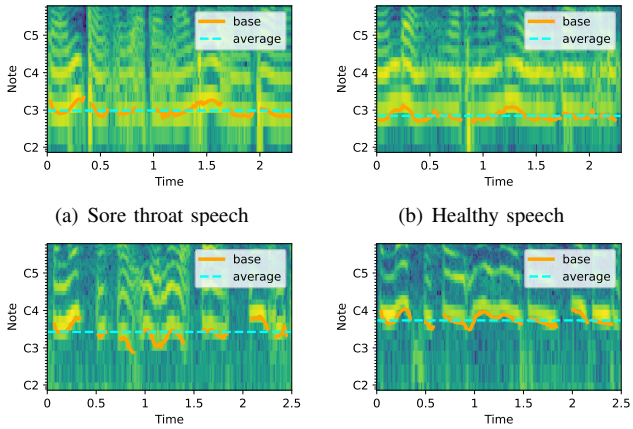


Fig. 7. The *SFF* curves of an identical utterance of six speakers, corresponding to S1-S6.



(a) Sore throat speech (b) Healthy speech

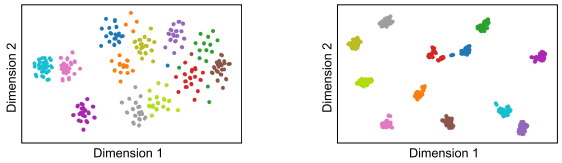
(c) Nasal sound speech (d) Healthy speech

Fig. 8. (a), (b) represent sore throat and healthy speech of a participant uttering Chinese "Fa Duan Xin Gei Wo De Nan Peng You" (i.e., Text to my boyfriend). (c), (d) represent another participant's nasal sound and healthy speech with the Chinese context "Jin Tian Tian Qi Zen Me Yang?" (i.e., How is the weather today).

audio files, and further processed into *SFF*. We perform these processing methods on several original and pitch-shift (±1 semitone) audio samples of different speakers. We utilize the t-SNE toolkit for visualization, where the difference between MFCC, classical DL-based X-vectors, and SFF are depicted in Fig. 20. We can observe that SFF outperforms others and has a high discrimination ability between speakers even when original and pitch-shift audio are mixed. We envision that SOFTER can outperform the X-Vectors also because of its unified model for all speakers. For instance, to achieve pitch insensitivity for a classical ASV system, the speaker data needs to be augmented when training the model, e.g., using both normal and pitch-modified speech. It may cause features of different speakers to overlap and make the model difficult to converge. Nevertheless, GMM models trained for individual users can still converge stably. We also compared a given utterance's SFF curves at five pitch levels, and found they are significantly close to each other. However, the SFF curves of six different speakers are distinct. We can conclude that SFF is relatively insensitive to pitch variation while keeping speaker discrimination ability.

C. Soundfield of Pathological Speech

As illustrated before, pathological speech can change pitch or introduce noise during articulation. We focus on three



(a) X-vectors (b) soundfield

Fig. 9. The t-SNE of X-Vector and *SFF*, the pathological and healthy speech of each participant were mixed together for feature extraction.

common symptoms that affect vocalization, i.e., nasal sound, sore throat (e.g., laryngitis), and both symptoms coincide. We collected the pathological and healthy speech respectively from each participant, who used to suffer nasal sound, sore throat, or both of them. We conducted quantitative analysis by comparing two audio's pitch. We start with performing pitch computation to obtain the pitch contours (i.e., fundamental frequency estimation, F0), adopting the pYIN algorithm [1]. Notably, as fundamental frequency dose not exist in unvoiced regions, thus those regions are automatically excluded from being labeled with pitch contours (in orange). As comparing two audio via irregular pitch contours is difficult, we average the multiple pitch contours' values of a given audio sample thus derive its base frequency. Results illustrate that the average base frequency of pathological speech differs from that of healthy speech by about 0.5-2 semitones. Among them, Fig. 8 shows two pairs of two participants' speech with different physical conditions. Moreover, we discover that the impact of symptoms on tone varies among people. We assume that the same symptoms originating from different etiologies will lead to different changes in speakers' vocal tracts. Our subsequent experiments in Sec. V also indicate that, when performing speaker authentication on pathological speech (i.e., with pitch change, laryngitis-related noise), classical ASVs suffer more significant performance degradation even if the enrolled and verified text are identical. Notably, all those ASVs' performances dropped a bit more when the mismatch from enrolled utterances existed. In contrast, the soundfield maintained acceptable performance regardless of symptoms that affect vocalization. In Fig. 9(b), the *SFF*s of 12 participants are visually clustered after dimension reduction with t-Distributed Stochastic Neighbor Embedding (t-SNE), compared to the X-vectors in Fig. 9(a), presents a great superiority.

D. Soundfield Observation

Based on SOFTER's usability, practicality and discrimination of speakers, our investigation focused on three aspects, conducted necessary experiments, and obtained the following insights:

- **Distance-sensitive.** Due to the complexity of the near field and the difficulty of fixing the position of the person and the phone, the distance-sensitivity of soundfield must be solved to achieve reliable speaker verification.
- **Pitch & Pathology-insensitive.** We revealed the similarity of pitch and symptom impacts on articulation. We also find that soundfield can still robustly represent speakers in cases of pitch variation and pathological speech than classical DL-based features.

IV. System Design

A. Design Goals and Challenges

We aim to design a robust speaker verification system that can achieve two goals: 1) performing reliable speaker verification regardless of the speakers' physical conditions with the advantages of pitch & pathology-robustness rendered in Sec. III. 2) enabling distance-agnostic verification to achieve user-friendliness and effectiveness. While there still are some challenges to be overcome:

- The user-microphone distance can vary, resulting in the mismatch of the enrolled and verified soundfield, thus degrading the system performance.
- To ensure the practicality of SOFTER, we can only utilize the off-the-shelf sensors on smartphones rather than the complex setups used in prototype systems to meet performance requirements.
- Noises and silence in speech signals should be removed due to not reflecting the soundfield.
- How to accurately build the speaker models of each participant with only several speech samples for enrollment?

B. Overview

To meet the above goals as well as tackle the challenges, we introduce the design of SOFTER in Fig. 10. It consists of the enrollment and verification stages, and there are four critical parts of the system: **1) Distance Sensing:** measure the range between the user's mouth area and smartphone. **2) Soundfield Reconstruction:** obtain the sound signal similar to the enrollment stage. **3) Soundfield Extraction:** derive the soundfield representing the speaker's identity. **4) Model Training and Inference:** gain the user-specific model and verify the user at the inference stage.

C. Echo Distance Ranging

To address the issue that soundfield varies with user-microphone distance, we envision recovering the current soundfield to the enrolled one. Fig. 4 depicts the common user-phone posture [24], in which the microphone and main speaker at the bottom of the smartphone conduct distance ranging via our designed acoustic signal. Acoustic ranging offers two advantages: 1) it does not rely on additional hardware and delivers a greater range than proximity sensors. 2) the echoes are so sensitive to the relative distance between the user's mouth and phone, which can meet the precision requirement.

Acoustic Signal Design. The signal design needs to meet several prerequisites. First, it allows the part of interest (i.e. echoes from the mouth) to be easily isolated from interferences from other transmission paths, such as some of the sound waves will directly propagate to the microphone, as shown in Fig. 4. Second, it should be imperceptible enough to users to minimize disturbance, e.g., ideally over 20 kHz. In contrast, some microphones on smartphones have poor high-frequency responses and need the designed signal to be

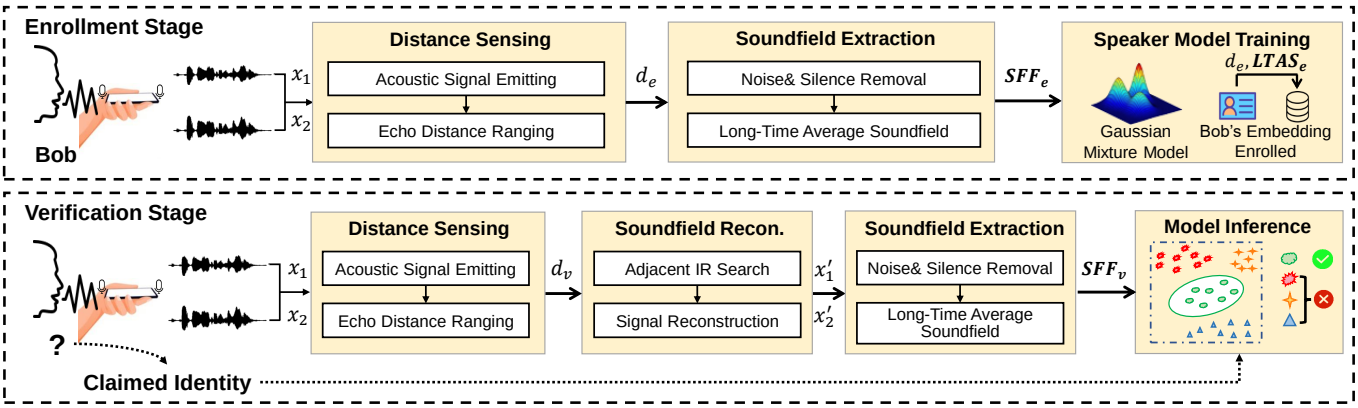


Fig. 10. Workflow of SOFTER. 1) Enrollment Stage: perform a series of processes on two-channel speech signals to obtain the enrolled soundfield features (SFF_e), gain the speaker model, and save the enrolled user-phone distance (d_e). 2) Verification Stage: *Sounfield Reconstruction* ensures the soundfield approximates the enrolled ones to achieve a distance-agnostic verification

separated from the human voice and noise bands to enable noise removal (e.g., via band-pass filters). Third, it should balance the short duration and robustness, as [29] adopts a near instant pulse (i.e., 1ms). Nevertheless, a signal pulse is so short that sometimes the hardware cannot respond, resulting in an unsuccessful launch.

Considering the aforementioned facts, we craft a signal consisting of 5 short 0.25ms-duration 12kHz single frequency (a.k.a, chirp) and 10ms interval between each chirp as shown in Fig.11 (a). Our design is based on several reasons: 1) The user-phone distance is generally 5-35cm, and to ensure reliable ranging, the chirp of the direct path cannot interfere with the echo path. Hence we must ensure that the earliest echo from the user's mouth back to the microphone is still later than the latest arrival time of the direct path. It is necessary to meet the following inequality: $T + \frac{D_d}{c} < \frac{2*D_e}{c}$, where the echo-path D_e is at least 5cm, the direct-path D_d is usually less than 1.5cm, thus it can be derived that the chirp duration T is about 0.25ms. 2) We set the interval between each chirp, which on the one hand, enables the hardware to respond in time and emit all pulses stably. On the other hand, the interval of 10ms is enough to separate different echoes, and the ranging results of 5 chirps help to reduce errors. We also apply the Hanning window to reshape the chirps' envelop to increase its peak-to-side lobe ratio as shown in Fig.11 (b), thus producing higher SNR for echoes.

Signal calibration. The raw signal first goes through an 11-13kHz Butterworth band-pass filter to remove background noises so that noises will not bury echoes from the human mouth. Except for the desired echoes, the signal also contains the direct-path component shown in Fig. 4. We cannot simply assume that it can be located with a constant delay due to the variable duration of hardware and software processing in both signal emitting and recording. In order to focus only on the echo path and exclude the direct path when sensing the distance, we recorded the direct transmission signal as a "template" in a quiet surrounding for future cross-correlation, where no reflector is within a two-meter distance in front of the bottom microphone. Notably, the difference between the template direct-path signal shown in Fig.11 (c) and the ideal

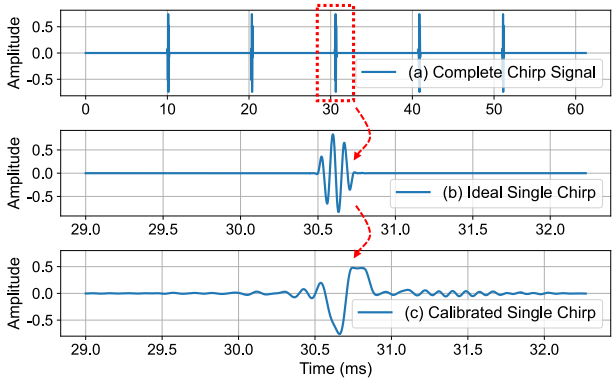


Fig. 11. Illustration of (a) distance ranging signal sequence and (b) ideal single chirp (c) calibrated single chirp is the ideal chirp's playback.

chirp is apparent due to the hardware imperfection.

Echo Decision. The preprocessed signal consists of direct- and echo-path components with energy only around 12kHz, shown in Fig.12 (a). A straightforward way to locate the echo-path part is to find the cross-correlation peak location [30] after the direct-path peak location. To tackle the multiple cross-correlation peaks challenge introduced by residual noise, we perform a *peaks identification strategy*: we denote the received signal after band-pass filtering as $e(t)$ and calibrated chirp signal as $s(t)$. Thus, to determine the exact direct- and echo-path signals' starting points, firstly, the calibrated signal $s(t)$ slides across the $e(t)$ sequence, and the correlation is calculated as follows:

$$F_{corr}(t) = \int_{-\infty}^{\infty} e(\tau)s(\tau - t)d\tau \tag{4}$$

Secondly, to capture the correlation trend changes of $F_{corr}(t)$, we derive the envelope of $F_{corr}(t)$ using the envelope detection strategy [31] and denote it as $E(t)$. The peaks within envelope $E(t)$ can be used as candidates to identify both direct- and echo-path beginning points. Thirdly, we adopt a local maximum identification method [32] that removes interfering cross-correlation peaks brought by the residual noise. It examines the envelope sequence $E(t)$ using a sliding window with a hop length of one sampling point. If the

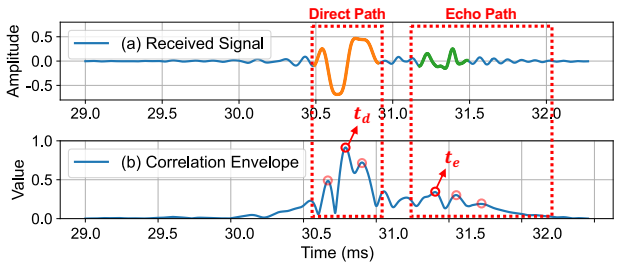


Fig. 12. Illustration of (a) the received single chirp signal and (b) cross-correlation peak location strategy.

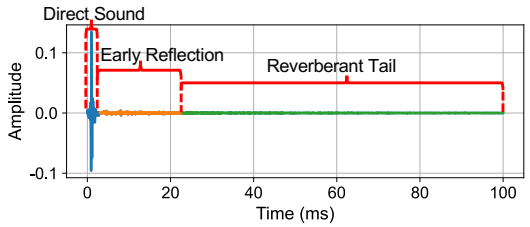


Fig. 13. Illustration of the recording IR sound.

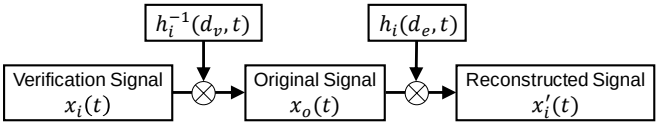


Fig. 14. The workflow of signal reconstruction.

extremum (i.e., peak) is less than the maximum value in the current window, such a peak is considered an interfering item and removed from the set of peaks. Therefore, as shown in Fig.12 (b), this strategy facilities our system to locate the start points of both direct and echo paths more precisely. Given that the energy of echoes is much weaker than the direct-path signal, we predefine the sliding window length for cross-correlation as 6 samples. It involves only the central portion with significant amplitudes during calculation, enhancing the adaptation to residual noise. Finally, the distance between a user and phone can be formalized as $d = c \cdot \frac{t_e - t_d}{2}$, where $t_e - t_d$ denotes the time difference between the echo and direct path.

D. Soundfield Reconstruction

We have learned about the performance degradation caused by distance sensitivity of the soundfield (i.e., location mismatch of enrollment and verification stage). Therefore, we aim to reconstruct the verification signals close to the enrollment phase. Specifically, we adopt the impulse response (IR), which enables the acoustic characteristics of a location to be captured and is considered a transfer function between the sound source and microphone [33]. IRs are also widely applied in numerous fields due to their ability to well characterize the physical world, e.g., data augmentation in speech/speaker recognition systems, speech enhancement, and physical adversarial attacks [34], [35].

Build IR Database. There are many approaches to deriving microphones' IRs, among which those near-instantaneous sounds are commonly adopted as the sound source for existing IR datasets [36], [37], e.g., hand clapping and gunshots, because of their convenience and effectiveness. However, keeping the impulse pattern generated by clapping hands always consistent is almost impossible. In contrast, the impulses generated by a stable sound source are more constant, such as high-fidelity loudspeakers. We employ an exponential sine sweep signal with aperiodic deconvolution formalized as Eq. 5, which also removes the artifacts caused by noise, nonlinear behavior of the speakers and time-variance [38].

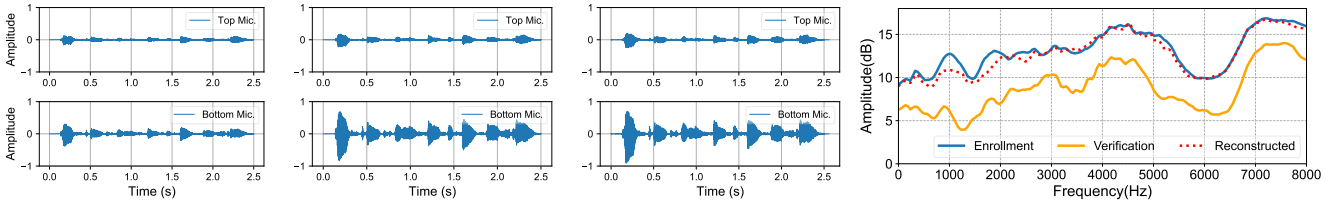
$$s(t) = sin[\frac{\omega_1 \cdot T}{ln(\frac{\omega_2}{\omega_1})} \cdot (e^{\frac{t}{T} ln(\frac{\omega_2}{\omega_1})} - 1)] \tag{5}$$

where $s(t)$ denotes a sweep that starts and ends at angular frequency ω_1, ω_2, respectively, taking T seconds. Specifically, we emit the sweep signal (i.e., from 0-24kHz, lasting 4 seconds) via JBL while the top and bottom microphones on smartphones record at different user-phone distances (i.e., 1cm intervals). Given $h_i(d,t), i \in [1,2]$ as the impulse response of respective top and bottom microphones and $s_i^{'}(d,t)$ as the recorded sweep signals, at difference distances, we have $s_i^{'}(d,t) = h_i(d,t) * s(t)$. Hence we obtain multiple IRs $h_i(d,t)$ of each user-phone location, using the deconvolution technology, i.e., $s_i^{'}(d,t)$ deconvolutes the time-reversal mirror of $s(t)$ [39]. Fig. 13 depicts an IR at a specific distance, which involves three parts: direct sound, early reflection, and reverberant tail. In order to ensure our system operate effectively in environments with varying sizes and shapes, e.g., office, lounge, balcony, we intercept the initial 3ms of IRs (i.e., direct sound part) to exclude multiple sound reflections and reverberation patterns.

In addition, we preset the IR database configured as a built-in library of application by the smartphone manufacturers, and users are also flexible to establish their own database following the above instructions. We establish respective IR databases for distinct smartphones instead of a unified database due to their microphone models and layout differences. Notably, SOFTER can maintain effective on varying smartphones, which is given in Sec. V. This may attribute to that embedded microphones are calibrated to comply with the PTSN (Public Switched Telephone Network) standard, therefore audible speech shares similar patterns on different devices [40]. Besides, unlike classical monophonic spectral features, *SFF* is intrinsically unaffected by microphones' gain and speaker's volume due to Eq. 1.

Adjacent IR Search. When users register their soundfield, user-phone distance d_e is given by echo sensing. Thus the adjacent IR $h_1(d_e,t)$ and $h_2(d_e,t)$ corresponding to the top and bottom microphones, i.e., the closet IR to the current location, can be derived by querying the IR database. Similarly, in the verification stage, we obtain the distance d_v and IRs $h_1(d_v,t)$ and $h_2(d_v,t)$.

Signal Reconstruction. Fig. 14 shows the signal reconstruction workflow. The verification speech signal recorded by the top and bottom microphones, shown in Fig. 15 (a), are denoted as $x_i(t), i \in [1,2]$. We derive the rebuilt signal in Fig. 15 (c), by sequentially convolving the received signal with inverse verification-stage and enrollment-stage transfer function [41]



(a) Verification Signal (b) Enrollment Signal (c) Reconstructed Signal (d) Soundfield Comparison

Fig. 15. The diagram of the effect of soundfield reconstruction. The reconstructed signal (c) obtained from a verification signal (a) at uncertain position, is very close to the actual enrolled signal (b), and the soundfield (d) is reconstructed almost identically to the enrolled one.

as follows:

$$x_i^{'}(t) = x_i(t) * h_i^{-1}(d_v, t) * h_i(d_e, t) \tag{6}$$

Fig. 15 (d) clearly denotes the rebuilt signal's soundfield almost identical to the enrolled signal in Fig. 15 (b), suggesting the signal reconstruction is valid.

E. Soundfield Extraction

Noise & Silence Removal. Noise can result in our acquisition of audio data in a low signal-to-noise ratio (SNR), impairing soundfield patterns and leading to poor speaker verification performance. Existing speech denoising algorithms are usually designed to improve audio audibility and intelligibility, where multi-channel solutions combine rich temporal and spatial information to eventually output denoised single-channel audio. To conduct denoising but not impair the intrinsic spatio-temporal information of audio, we propose a respective channel denoising method by making both microphones "sense" their environments for 0.5 seconds before the user speaks up, i.e., sampling the ambient noise so as to represent the real-time noise situation. When the user's voice is detected, we utilize the NoiseReduce toolbox [42], [43] to perform fast denoising via spectral gating. We also adopt voice activity detection (VAD) [44] to eliminate the influence of unvoiced portions (silent pauses and breaks within the utterance), which do not reflect the speaker's soundfield.

Profiling the Soundfield. Combined with the modeling approach for the soundfield proposed in Sec. III, we start with a temporal and frequency processing of the two-channel signals, respectively, i.e., using the short-time Fourier transform (STFT), where each frame is 25ms duration, the overlap between consecutive frames is 12.5ms (smoothing out the frequency variation between phonemes), and we reduce the spectral leakage by applying Hanning window to each frame. Then we calculate the logarithm of the two-channel ratio according to Eq. 1. Notably, it reduces the influence of the user's loudness, especially considering that a person may not speak as loudly as he usually does when in unhealthy state. Finally, we obtain the n-dim *SFF* vector by performing long-time average normalization according to Eq. 3. In our experiments, the frequency resolution is 24kHz/512=46.8Hz. Given the typical applications such as online chatting and phone call with a sample rate not over 16kHz, we select the first 171 dimensions out of 512 if not stated otherwise.

F. Speaker Model

Compared to traditional spectral features such as Fbank and MFCC, *SFF* has the advantages of consistency and distinctiveness in characterizing identities. We believe that applying effective modeling can help authenticate users more reliably. The Gaussian mixture speaker model (GMM) can well fit the user's soundfield vectors to form the corresponding speaker model. Specifically, the distribution of feature vectors (i.e., soundfield) extracted from a person's speech signals is modelled by a Gaussian mixture density. For a n-dimensional soundfield vector, the mixture density for speaker s is defined as

$$p\left(SFF \mid \lambda_s\right) = \sum_{i=1}^{M} w_i^s p_i^s(SFF) \tag{7}$$

where M is the number of Gaussian components, and w_i denotes the mixing weight of the i_{th} Gaussian component $p_i(SFF)$ that parameterized by an $n \times 1$ mean vector μ_i and an $n \times n$ covariance Δ_i^s.

$$p_i(SFF) = \frac{1}{(2\pi)^{\frac{n}{2}} \left|\Delta_i^s\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(SFF-\mu_i)'(\Delta_i^s)^{-1}(SFF-\mu_i)} \tag{8}$$

The mixing weight w_i have the property of $\sum_{i=1}^{M} w_i^s = 1$, and the speaker model is $\lambda_s = \{w_i^s, \mu_i^s, \Delta_i^s\}, i = 1, \cdots, M$.

In the enrollment stage, given several user utterances, the speaker model parameters are estimated and converged using the iterative Expectation-Maximization (EM) algorithm. In the verification stage, the claimed identity's speaker model serves as a likelihood function to obtain a similarity score with input soundfield vectors compared with the predefined threshold. If the score exceeds the threshold, the speaker's identity is verified. Otherwise, it is rejected.

V. Evaluation

In this section, we evaluate the performance of SOFTER on speaker verification, i.e., how well it can distinguish a valid speaker from a stranger's voice while meeting distance-, pitch & pathology-agnostic requirements. To conduct a comprehensive evaluation, we first collect two speech datasets as listed in Sec. V-A. Second, we compare SOFTER with 6 state-of-the-art classical ASVs, i.e., DeepSpeaker [45], Pyannote [46], SpeakerNet [47], X-vectors [22], I-vectors [17], and commercial IFlytek [48] APIs under the same experiment settings. We implement and evaluate SOFTER on a server with Intel Xeon(R) Gold 5117 CPU, NVIDIA GeForce RTX3090 GPU, and 64-bit Ubuntu 18.04 LTS operating system.

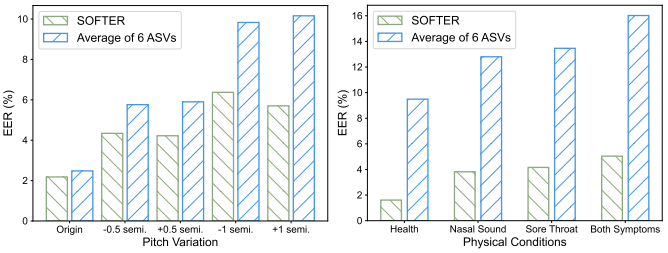
A. Experiment Setup

Data Collection. As previous work and our background (Sec. II-B) & investigation (Sec. III) have shown, symptoms affect vocalization is strongly related to pitch. We first evaluate SOFTER's pitch-insensitivity on a simulated English speaker recognition dataset (Voice-1). Since we foresee the robustness to pitch of the speaker verification system is the key to reliable verification in pathological speakers. Thus we collected a real pathology dataset (Voice-2) accessing each speaker's pathological and healthy speech, covering three cases that impact the voice (nasal sound, sore throat, and both symptoms simultaneously).

Voice-1: We find the VCTK corpus [28] well suited for our simulation, which is widely used in voice conversion and speech synthesis thanks to its high recording quality and sample rate (48kHz). Compared to most speaker recognition datasets sampled at 16kHz, VCTK retains the maximum property of the speakers' original speech signals and includes adequate native speakers of English (110 participants), therefore also commonly used for speaker-related tasks. We leverage the pyroomacoustic toolkit [49] that facilitates simulating the propagation between sound sources and microphones in 2D or 3D rooms, matching the idea of soundfield well. Specifically, We constructed an $8 \times 6 \times 3$ (meter) room where speakers stood on one side of the room and restricted each speaker to nearly the same position, with their height randomly set from 1.5 to 2m. Two microphones placed horizontally and 15cm (typical phone length) apart in the front of the virtual speaker's mouth were used to imitate a user holding the phone, as shown in Fig. 4. The user-phone distance for each speaker was randomly chosen between 5 and 35cm, considering different people's habits in practice. Furthermore, we filtered out 52 representative samples with almost the same content, out of 400 sentences of each speaker in VCTK, according to four types of word numbers (i.e., less than 5 words, 6 to 8 words, 9 to 11 words, and 12 words or more). Besides, we also modify each speaker's uttering pitch at 4 different levels to simulate the pathological speech. In sum, we obtained 28,600 simulated two-channel audio samples from 110 speakers.

Voice-2: In order to examine the effectiveness of SOFTER in real-world scenarios, we first gathered the utterances from 24 individuals in pathological conditions (i.e., nasal sound, sore throat, and both of them) by crowdsourcing, including 15 females and 9 males aged from 12 to 75[1]. Second, we kept in touch with them and performed additional utterance recordings when the subjects recovered vocalizing normally. Their voices were recorded by the participants' smartphones, while each held it horizontally in front of their mouth, within a range of 5 to 35cm, depending on their habits. Each participant was informed about uttering 50 Chinese commands under healthy conditions, 10 of which were used for registration, as well as these 50 identical commands under voice-affected conditions for verification. Specifically, the given utterances list for each participant is randomly selected from a translated version of ok-google.io, containing common interaction commands. We totally obtained 2,400 samples. The recording was performed in offices/dorms with slight background noises such as keyboard striking, talking, walking, and HVAC noises.

[1]We followed the Institutional Review Board (IRB) regulations to protect the rights of human participants.



(a) Voice-1 Result (b) Voice-2 Result

Fig. 16. Performance comparison of SOFTER with the average EER of 6 classical ASVs under each setting.

Evaluation Metrics. In our evaluation, we choose some classical metrics in speaker recognition, i.e., false acceptance rate (FAR), false rejection rate (FRR), equal error rate (EER), and accuracy. FAR: it characterizes the rate at which a cheater is wrongly accepted by the system and considered a registered user. FRR: it characterizes the rate at which the system falsely rejects a registered user. EER: it shows a balanced view of the FAR and FRR and is defined as the rate at which the FAR equals the FRR. AUC is widely applied in machine learning due to its insensitivity to data imbalance. A higher AUC score indicates that the system has more consistent performance even with changes in physical conditions. Of all these metrics, we are more concerned about the FRR because we hope that SOFTER can correctly authenticate registered users even if their voices are influenced by pathology.

B. Overall Performance

We use Voice-1 & Voice-2 to evaluate the performance of 6 aforementioned ASVs and ours. Note that SOFTER does not rely on a large amount of training data as classical ASVs do, it only requires 10 utterances per speaker in healthy conditions for enrolling, and the rest are used for verification. As for 5 local ASVs, we also reproduced them with two large speech datasets. To eliminate performance errors due to language differences, we train these models with Voxceleb2 when evaluating the English dataset Voice-1, and for the Chinese dataset Vocie-2, we train these ASVs with Aishell-1. For the commercial IFlytek system, we directly call the API to perform the same operation.

Voice-1 Results: The AUCs and EERs demonstrate that SOFTER is slightly inferior to the state-of-the-art SpeakerNet and X-Vectors in cases where the utterances' pitch is consistent with the registered, shown in Tab. I (the row noted as "original"). Nevertheless, it still surpasses the widely used DeepSpeaker, Pyannote, I-vectors and IFlytek, indicating the effectiveness of soundfield. Moreover, the EERs under all 5 pitch levels show that SOFTER is more insensitive to pitch changes than the well-trained DL-based ASVs, even if the pitch of VCTK speakers was changed by ascending 1 semitones, SOFTER still keeps the EER below 6.4%. In contrast, the best EER of other models already exceeds 9.23%.

TABLE I
THE OVERALL PERFORMANCE OF SOFTER AND OTHER SIX ASV MODELS ON VOICE-1 AND VOICE-2 DATASETS.

Dataset & Model		DeepSpeaker		Pyannote		SpeakerNet		X-vectors		I-vectors		IFlytek		SOFTER	
		AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Voice-1	original	99.63	2.45	99.57	2.61	99.76	1.86	99.76	1.92	99.45	3.13	99.43	3.20	**99.65**	**2.18**
	-0.5 semitones	98.06	7.66	98.57	6.07	99.10	4.39	99.05	4.53	98.53	6.19	98.52	6.30	**99.12**	**4.34**
	-1 semitones	96.62	11.15	97.26	9.79	97.46	9.33	97.17	9.99	97.03	10.33	97.49	9.23	**98.49**	**6.37**
	+0.5 semitones	98.15	7.34	98.48	6.34	98.85	5.23	98.89	5.05	98.78	5.44	98.39	6.62	**99.13**	**4.22**
	+1 semitones	96.69	10.99	97.82	8.40	98.00	7.84	95.87	12.32	95.10	13.26	97.58	9.03	**98.69**	**5.70**
Voice-2	health	99.32	4.17	97.99	7.23	98.82	5.55	92.58	16.16	93.04	15.18	97.09	8.65	**99.82**	**1.61**
	nasal sound	97.85	7.48	95.54	10.95	96.50	9.55	91.09	18.61	90.63	19.12	95.42	11.09	**99.42**	**3.82**
	sore throat	96.35	9.77	96.61	9.36	96.28	9.88	88.46	20.63	89.06	20.30	95.22	11.41	**99.31**	**4.16**
	both symptoms	96.08	10.16	94.06	13.26	94.33	12.77	85.51	23.37	85.32	23.58	94.23	12.97	**99.03**	**5.04**

Furthermore, we also compare the relative performance improvement of SOFTER compared to the other 6 models on average. We obtain the EERs improved by 12.11%, 24.67%, 28.50%, 35.19%, 43.92% respectively for these 5 settings, shown in Fig. 16 (a). Overall, the advantages of SOFTER over the other models are pronounced.

Voice-2 Results: On the recorded data of pathological speakers, SOFTER shows a clear superiority in real scenarios compared to other systems. First of all, the EER of our system verified in healthy conditions at 1.61% is close to the result of Voice-1 experiment, despite language and content differences between the two datasets, suggesting the system is language- & content-agnostic. At the same time, it remarkably outperforms the rest classical models with an average EER of up to 9.49%. Notably, we find that SOFTER has significantly weaker performance degradation than other models in three symptoms, regardless of nasal sound, sore throat, or both, with only 10 utterances registered in health. Specifically, the average EERs of 6 classical ASVs are 9.49%, 12.80%, 13.46%, and 16.02% in 4 physical conditions, shown in Fig. 16 (b), where ours achieve at least 68.53% improvements. On the one hand, we consider it related to the habit of people holding their phones away from their mouths. On the other hand, the actual soundfield is more complex, except for the direct-path sound waves, components affected by the head and torso, etc., also introduce the individual's distinct information. Therefore the adverse effects caused by the affected voice source are weakened.

C. Impact Factors on Performance

In this section, we concentrate on evaluating the robustness of SOFTER as a speaker verification system against the influence factors in common scenarios. We conduct the sample rate and content length experiments based on Voice-1, and we evaluated SOFTER based on Voice-2 considering the impacts of real-world noise, location, and recording devices.

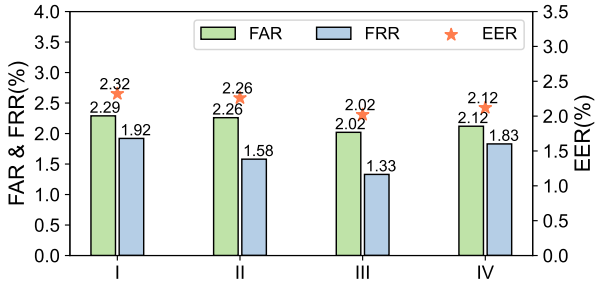
Impact of Sample Rate. When we record or transmit the audio, those with high sample rates can always retain more information than the audio with low sample rates. The missing information in the recording process may influence speaker verification. Therefore, we experimented with four typical sample rates (8kHz, 16kHz, 44.1kHz and 48kHz) to explore the effect of sampling rate on SOFTER. We enroll and evaluate speaker models using audio samples with the same content and four different sampling rates. As Fig. 17 (a) shows, our system can keep high performance at audios with all sample rates, where the EER of 8kHz (2.51%) is slightly higher. We assume that a low sample rate may result in a loss of high-frequency information and lead to performance degradation. However, when the sample rate increases above 16kHz, SOFTER maintains stable performance regardless of the sample rate variation. The results verify that SOFTER can distinguish speakers well under the standard sample rate settings.

Impact of Content Length. We should validate whether SOFTER is content-agnostic compared to multiple text-independent ASV models. A reasonable intuition is that the performance correlates with the content length. Therefore, we divided our audio into four parts according to the content length. Each group contains sentences of 0-7, 8-11, 12-19, and over 20 words. Fig. 17 (b) depicts that our system maintains performance well in all four groups. Even though the FARs of relative short utterances are slightly higher at 2.29% and 2.26% than long utterances, the FRRs and EERs still keep in an acceptable range at about 2.26%. The results imply that SOFTER has the capability of distinguishing users regardless of their command length.

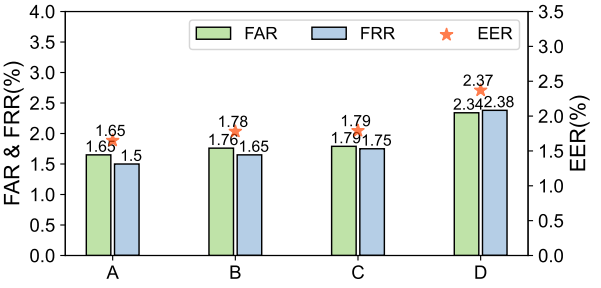
Impact of Noise. When users speak commands for verification in practical scenarios, noise is an inevitable and essential factor we should take into account. We experimented with 4 representative noises in daily scenarios to quantify our system's robustness to resist noise's impact, including office (keyboard striking, 60dB), home (frying food, 65dB), cafeteria (people whispering, 70dB), white noise (75dB). As Fig. 17 (c) shows, the performance of three groups with typical noises of keyboard striking, frying food, and people whispering, is nearly the same as the original group. This result shows that SOFTER can resist the natural noises well with the EERs of 1.65%, 1.78%, and 1.79%, respectively. As for the random noise, our system suffers a slight impact while the performance is still acceptable with an EER of 2.37%. We assume that the randomly generated noise will cause some unnatural changes to soundfield. The results verify that our system is robust enough to resist the impact of noise.

Impact of Position. Sec IV discusses and validates that SOFTER can achieve distance-agnostic by reconstructing the speech signals of different locations. Through our reconstruction, the soundfields of various distances will be rebuilt

(a) Impact of Sample Rate



(b) Impact of Content Length



(c) Impact of Noise

Fig. 17. Result of experiments of three impact factors. (a): impact of sample rate. (b): impact of content length. I, II, III, and IV respectively represent the sentences with 0-7, 8-11, 12-19, and over 20 words. (c): impact of noise. A, B, and C are noises of keyboard striking, frying food, and people whispering. D denotes the white noise.

highly similar to the registered soundfield. We first gather the participants to enroll their voices at a distance of 5cm. Then we record their voice at a distance from 5cm to 30cm for ASV. SOFTER will reconstruct this audio and then take the reconstructed audio as input. To better represent our system's performance, we also conduct an experiment with the original soundfield without reconstruction. Fig. 18 shows that traditional soundfield will be affected by the distance, and the EER of the distance at 30cm reaches up to about 20%. In comparison, SOFTER can perform well at different distances with stable and reasonable EERs of 2.82%, 2.89%, 2.96% and 2.42% at 4 locations respectively. As the distance increases, the gap between reconstructed and original soundfields will be more pronounced. It verifies that SOFTER is a distance-insensitive ASV system.

Impact of Recording Devices. There is a concern about whether SOFTER can keep high performance in different recording devices because the generalization of devices also matters a lot. Therefore, we further evaluate on five mainstream smartphones, including Redmi K40, Google Nexus,

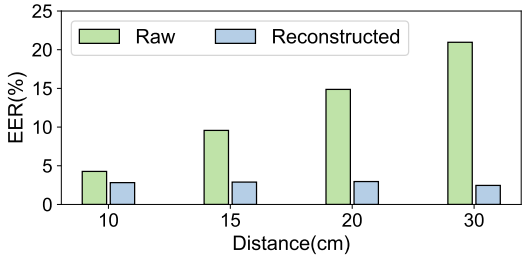


Fig. 18. Results of four different verification distance between original and reconstructed soundfield.

Huawei P30, Google Pixel, and OPPO Reno5. We fixed the same position of these phones and collect 15 speakers' speech samples. The EERs of these five additional devices are 1.38%, 1.64%, 1.42%, 1.79% and 1.25%, respectively. This result illustrates that SOFTER can maintain comparable performance even in different recording devices, indicating the generalization and utility of our system in further applications.

D. Runtime Overhead

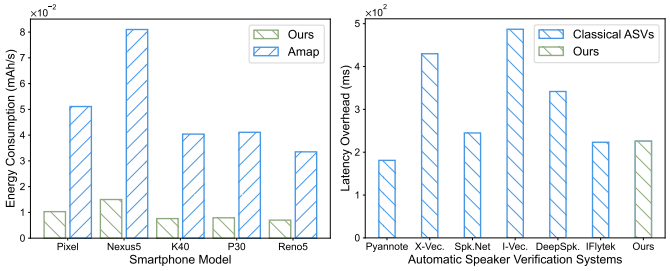
As a service for handheld devices, the structure of SOFTER is similar to other voice assistants such as Siri. We split and deploy it on the smartphone (device-side) and server (server-side), respectively. The device-side conducts location sensing and captures speech signals (i.e., soundfield). Moreover, the server-side executes signal reconstruction, soundfield extraction, and speaker model inference. The runtime overhead of our system is mainly divided into the energy overhead brought by the device-side APP operating, the latency overhead from audio file transferring between client and server, server-side processing & inference, and result returning.

Energy Overhead: We loop the application for echo sensing and recording with each lasting 3s, running on five different smartphones (i.e., Google Pixel, Google Nexus5, Redmi K40, Huawei P30, and OPPO Reno5) for an hour. Thus, we obtain the average power consumption from these devices at 9.6e-3 mAh/s (mAh per second). By comparison, the power consumption of navigation APPs, i.e., Amap [50], on these devices is significantly higher in Fig. 19(a), around 49.4e-3 mAh/s. Notably, SOFTER only consumes power when invoked. Combined with its low consumption, the impact on users' daily experience is negligible.

Latency Overhead: We also compare the overall latency of SOFTER with other ASVs by performing the same number of test samples and obtaining the average latency. Fig. 19(b) shows that SOFTER's latency of 225.7ms, meeting the latency requirements of commercial voice assistant APIs (480ms [51]), of which the audio uploading delay, server-side processing & inference delay, and result feedback delay are 77.8, 142.1, and 5.8 ms, respectively. There are only Pyannote & IFlytek slightly outperforms ours, and SOFTER is significantly faster than X-vectors & I-vectors. We envision deploying all the functions on the device-side can reduce the latency efficiently.

VI. Discussion

In this section, we discuss SOFTER from three aspects: application scenario, potential improvements, and comprehensive



(a) Energy Overhead (b) Latency Overhead

Fig. 19. Runtime overhead comparison of SOFTER and counterparts.

datasets in this area.

Application scenario. Our system focuses on handheld device scenarios, especially for users who speak with a smartphone held. While our systems may not perform well in several corner cases, such as the user being far from the smartphone or speaking in a narrow space. Specifically, the nature of acoustic propagation theoretically makes measuring the soundfield at an extreme distance difficult. As the distance increases, the sound waves change more faintly, and the sound pressure is closer at different locations. It is too challenging to profile the soundfield based on only two microphones. Besides, in a narrow space, acoustic propagation becomes very complex, in which multiple reflections make the soundfield subject to many interfering factors and difficult to measure. Apart from the handheld devices, we envision that SOFTER can be extended to more scenarios in the future.

More microphones for authentication. After investigating various manufacturers and their prototypes of smartphones, we discovered that most smartphones had been equipped with three or even four microphones. These extra microphones are usually installed on the back of the phone and used for recording when taking videos. However, most existing smartphone recording applications only support exporting a two-channel audio file, which limits the capability of collecting and profiling the soundfield. We believe that in the future, with the opening of the APIs, we can use more microphones to improve the accuracy of our system.

Comprehensive datasets. The evaluation on our system requires both healthy and pathological speech samples of the subjects, making it difficult to collect data due to the more extended period and smaller patient samples than traditional ASV tasks. To the best of our knowledge, existing research on pathological speech mainly focuses on detecting whether a given utterance is pathological or not. There is still no comprehensive or authoritative dataset for such a goal. To evaluate SOFTER as well as the classical ASVs on more participants, we believe that in the future, a larger corpus could help and accelerate the related research in this area.

Intervals of IR database. To balance the database's effectiveness and the manufacturers' burden, we investigate when the number of different distances is enough for constructing such a database. As described in Sec. IV-D, we initially collected IRs from 5cm 35cm at 1cm intervals (31 distances involved), taking around 6∼8 minutes per device. Besides, we constructed IR databases in 3cm (11 distances involved) and 5cm intervals (7 distances involved). We assume the user enrolls the soundfield at 5cm and verifies at uncertain positions. We also select two distances: 12.5cm is selected for the 5cm-interval group because it reaches a maximum deviation of 2.5cm from both 10cm and 15cm in the database, and should have the most significant reconstruction error. Similarly, 21.5cm is chosen for the 3cm-interval group.

TABLE II
EUCLIDEAN DISTANCES OF ENROLLED AND RECONSTRUCTED SIGNALS

Group	Euclidean Distance (12.5 cm)	Euclidean Distance (21.5 cm)
1cm-interval	13.84	18.14
3cm-interval	14.96	21.24
5cm-interval	25.69	29.10

Without soundfield reconstruction, 38.81 and 77.81 are the Euclidean distances between enrolled (in blue) and verification signals (in yellow) at 12.5cm and 21.5cm, respectively. Table. VI shows the Euclidean distances between enrolled and reconstructed signals. We found that a larger interval corresponds to a larger Euclidean Distance (i.e., reconstruction error). In addition, a larger Euclidean distance (3cm-interval: 14.96 vs. 1cm-interval: 13.84) between the enrolled and reconstructed signals only brings a slight performance decrease (i.e., slightly higher EER). We envision that manufacturers can balance the IR database establishment burden and reconstruction effectiveness according to their security and usability requirements. For instance, as for a scenario with higher security requirement, the manufacturer should choose 1cm-interval or even more refined intervals. As for scenarios where usability is more important, we think 3cm-interval can meet the requirements.

VII. Related Work

Voice authentication. Voice authentication is an essential branch in biometric authentication technology (e.g., face, voiceprint, fingerprint) and can be divided into two categories: active and passive. "Active" refers to the authentication method of obtaining user characteristics by emitting and receiving designed signals, and "passive" denotes performing a series of processing and authentication for users' speech signals only. For active authentication, VocalPrint [52] uses mmWave to sense tiny vocal vibrations near the user's throat, but the cost is prohibitive due to the introduction of expensive equipment. EarPrint [53] obtains the sound conduction characteristics of the user's body for identification by sending a swept signal in the ear canal and receiving it. VocalLock [54] emits FMCW-like acoustic signals to characterize the static vocal tract shape and dynamic motion of the vocal tract during speech. In passive authentication, deep learning-based methods are widely used. At the same time, the lack of pathological speech data hinders its prosperity in this direction due to the prerequisite of considerable training data. [55] uses inward-facing microphones to collect bone-conducted sounds of dental occlusion in binaural canals to achieve authentication. [56] captures the dynamic movements of the lip based on the Doppler effect, which adversaries can easily

imitate. CaField [12] enables capturing soundfield by commercial smartphones to distinguish humans from loudspeakers to defend against spoofing attacks. [57] implements speaker verification and liveness detection by leveraging the additional high-sample rate microphone, which is impractical in most scenarios. However, existing works did not consider a non-negligible authentication case that speech attribute changes because of vocal tract sickness. SOFTER investigates that the soundfield is insensitive to the human physical condition and proposes distance sensing and soundfield reconstruction to address the distance-sensitive challenge. Besides, we do not require any additional hardware or modification, suggesting its practicality, usability, and effectiveness.

Pitch-variable & pathological speech. Previous works have uncovered the significant impact of pitch on ASVs [9], [58]. A speaker's voiceprint would inevitably change with the pitch as a fundamental speech attribute. Based on the property, attackers can easily disguise ASVs by only modifying the pitch [10], [11]. Correlatively, intra-speaker pitch variation may also be caused by ageing or pathology. [59] proposed a multi-task learning strategy to improve the pitch-variable singer identification performance, which might be a corner case in ASV. We envision new biometrics—soundfield can theoretically fix the issue. We also demonstrate that soundfield is pitch-insensitive and generalize it for pathological speaker verification. Whereas existing works mainly focus on exploring the characteristics of pathological speech and diagnosis of the specific etiology. [7] investigates the audible effects of cold on a phonetic level and shows that pathological speech introduces pitch changes and additional noise. [60], [61] look into the aspect of detecting etiology, such as how to detect Parkinson's disease via speech. [8] collects a dataset of 40 days and proposes a dual model strategy to resist the degradation caused by unhealthy physical conditions, while pathological speech usually suddenly appears and is rare. It requires more data to update the models smoothly, suggesting it lacks practicality. In contrast, our work aims to robustly authenticate users on both sides (health and unhealth), requiring only a few utterances for enrollment.

VIII. Conclusion

In this paper, we propose SOFTER based on soundfield to tackle the challenges of poor speaker verification performance when the speech is pitch-changed, which is also inevitable when a speaker is in unhealthy physical conditions. Our investigation demonstrates that soundfield delivers advantages of pitch & pathological speech-insensitiveness. In addition, for the mismatch between soundfield and the registration phase due to the user-phone distance change during authentication (i.e., distance sensitivity of the soundfield), we propose a two-stage mechanism: distance ranging and soundfield reconstruction to tackle the challenge. We also collected two datasets, including 31,000 utterances recorded from different physical conditions of 134 speakers. Results show that SOFTER can maintain the speaker discrimination capability well regardless of the pitch variation or speaker's physical conditions, suggesting its effectiveness.

References

[1] S. Jajodia and H. C. van van Tilborg, *Encyclopedia of Cryptography and Security: L-Z*. Springer, 2011.

[2] A. Klapuri, "Introduction to music transcription," in *Signal processing methods for music transcription*. Springer, 2006, pp. 3–20.

[3] S. Team, "Personalized hey siri," https://machinelearning.apple.com/research/personalized-hey-siri.

[4] R. Liu, C. Cornelius, R. Rawassizadeh, R. Peterson, and D. Kotz, "Vocal resonance: Using internal body voice for wearable authentication," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–23, 2018.

[5] R. He, X. Ji, X. Li, Y. Cheng, and W. Xu, "Ok, siri" or" hey, google": Evaluating voiceprint distinctiveness via content-based prole score," in *Proceedings of the 31th USENIX Security Symposium*, 2022.

[6] R. G. Tull and J. C. Rutledge, "Analysis of "cold-affected" speech for inclusion in speaker recognition systems." *The Journal of the Acoustical Society of America*, vol. 99, no. 4, pp. 2549–2574, 1996.

[7] J. Wagner, T. Fraga-Silva, Y. Josse, D. Schiller, A. Seiderer, and E. André, "Infected phonemes: How a cold impairs speech on a phonetic level," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, F. Lacerda, Ed. ISCA, 2017, pp. 3457–3461. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1066.html

[8] H. Ai, Y. Wang, Y. Yang, and Q. Zhang, "An improvement of the degradation of speaker recognition in continuous cold speech for home assistant," in *International Symposium on Cyberspace Safety and Security*. Springer, 2019, pp. 363–373.

[9] B. O'Brien, C. Meunier, and A. Ghio, "Evaluating the effects of modified speech on perceptual speaker identification performance," in *Interspeech 2022*, 2022.

[10] L. Zheng, J. Li, M. Sun, X. Zhang, and T. F. Zheng, "When automatic voice disguise meets automatic speaker verification," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 824–837, 2020.

[11] L. Tavi, T. Kinnunen, and R. G. Hautamäki, "Improving speaker de-identification with functional data analysis of f0 trajectories," *Speech Communication*, vol. 140, pp. 1–10, 2022.

[12] C. Yan, Y. Long, X. Ji, and W. Xu, "The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 1215–1229.

[13] S. Furui, "Speaker recognition," http://www.scholarpedia.org/article/Speaker_recognition, 2008.

[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[15] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE signal processing letters*, vol. 13, no. 5, pp. 308–311, 2006.

[16] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.

[17] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[18] P. Kenny, "Bayesian speaker verification with, heavy tailed priors," *Proc. Odyssey 2010*, 2010.

[19] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Twelfth annual conference of the international speech communication association*, 2011.

[20] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 1695–1699.

[21] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[22] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[23] N. Lavan, A. M. Burton, S. K. Scott, and C. McGettigan, "Flexible voices: Identity perception from variable vocal signals," *Psychonomic bulletin & review*, vol. 26, no. 1, pp. 90–102, 2019.

[24] L. Zhang, S. Tan, and J. Yang, "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*, 2017, pp. 57–71.

[25] G. Zhang, X. Ji, X. Li, G. Qu, and W. Xu, "Eararray: Defending against dolphinattack via acoustic attenuation." in *NDSS*, 2021.

[26] B. E. Treeby and B. T. Cox, "k-wave: Matlab toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of biomedical optics*, vol. 15, no. 2, p. 021314, 2010.

[27] S. D. I. Software, "Sound fields: Free versus diffuse field, near versus far field," https://community.sw.siemens.com/s/article/sound-fields-free-versus-diffuse-field-near-versus-far-field, 2020.

[28] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

[29] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 321–336.

[30] B. Zhou, M. Elbadry, R. Gao, and F. Ye, "Batmapper: Acoustic sensing based indoor floor plan construction using smartphones," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 42–55.

[31] Y. Ren, C. Wang, Y. Chen, J. Yang, and H. Li, "Noninvasive fine-grained sleep monitoring leveraging smartphones," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8248–8261, 2019.

[32] Y. Ren, P. Wen, H. Liu, Z. Zheng, Y. Chen, P. Huang, and H. Li, "Proximity-echo: Secure two factor authentication using active sound sensing," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[33] J. Merimaa, T. Peltonen, and T. Lokki, "Concert hall impulse responses pori, finland: Reference," *Tech. Rep*, 2005.

[34] J. Deng, Y. Chen, and W. Xu, "Fencesitter: Black-box, content-agnostic, and synchronization-free enrollment-phase attacks on speaker recognition systems," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022.

[35] J. Ze, X. Li, Y. Cheng, X. Ji, and W. Xu, "Ultrabd: Backdoor attack against automatic speaker verification systems via adversarial ultrasound," in *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2023, pp. 193–200.

[36] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor, "The ace challenge—corpus description and performance evaluation," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2015, pp. 1–5.

[37] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *2009 16th International Conference on Digital Signal Processing*. IEEE, 2009, pp. 1–5.

[38] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio engineering society convention 122*. Audio Engineering Society, 2007.

[39] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Audio Engineering Society Convention 108*. Audio Engineering Society, 2000.

[40] X. Li, X. Ji, C. Yan, C. Li, Y. Li, Z. Zhang, and W. Xu, "Learning normality is enough: A software-based mitigation against the inaudible voice attacks," in *Proceedings of the 32nd USENIX Security Symposium*, 2023.

[41] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, 1979.

[42] T. Sainburg, "timsainb/noisereduce: v1.0," 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139

[43] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, p. e1008228, 2020.

[44] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2021.

[45] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[46] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech 2021*, 2021.

[47] N. R. Koluguri, J. Li, V. Lavrukhin, and B. Ginsburg, "Speakernet: 1d depth-wise separable convolutional network for text-independent speaker recognition and verification," *arXiv preprint arXiv:2010.12653*, 2020.

[48] iFlytek Cloud, "Iflytek speaker recognition," https://console.xfyun.cn/services/ivp, 2022.

[49] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018, pp. 351–355. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461310

[50] A. Cloud, "Introduction of amap," https://www.alibabacloud.com/zh/customers/autonavi.

[51] K. Kumar, C. Liu, Y. Gong, and J. Wu, "1-d row-convolution lstm: Fast streaming asr at accuracy parity with lc-blstm." in *INTERSPEECH*, 2020, pp. 2107–2111.

[52] H. Li, C. Xu, A. S. Rathore, Z. Li, H. Zhang, C. Song, K. Wang, L. Su, F. Lin, K. Ren *et al.*, "Vocalprint: exploring a resilient and secure voice authentication via mmwave biometric interrogation," in *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 2020, pp. 312–325.

[53] Y. Gao, Y. Jin, J. Chauhan, S. Choi, J. Li, and Z. Jin, "Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–25, 2021.

[54] L. Lu, J. Yu, Y. Chen, and Y. Wang, "Vocallock: Sensing vocal tract for passphrase-independent user authentication leveraging acoustic signals on smartphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 2, pp. 1–24, 2020.

[55] Y. Xie, F. Li, Y. Wu, H. Chen, Z. Zhao, and Y. Wang, "Teethpass: Dental occlusion-based user authentication via in-ear acoustic sensing," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 1789–1798.

[56] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1466–1474.

[57] H. Guo, Q. Yan, N. Ivanov, Y. Zhu, L. Xiao, and E. J. Hunter, "Supervoice: Text-independent speaker verification using ultrasound energy in human speech," in *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, 2022, pp. 1019–1033.

[58] W. Apple, L. A. Streeter, and R. M. Krauss, "Effects of pitch and speech rate on personal attributions." *Journal of personality and social psychology*, vol. 37, no. 5, p. 715, 1979.

[59] Y. Gao, X. Li, C. Li, W. Sun, X. Ji, and W. Xu, "Varasv: Enabling pitch-variable automatic speaker verification via multi-task learning," in *2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2)*. IEEE, 2021, pp. 3240–3245.

[60] R. Tadeusiewicz, "Application of neural networks in the diagnosis of pathological speech," in *Artificial Neural Networks in Biomedicine*. Springer, 2000, pp. 141–150.

[61] C. Botelho, F. Teixeira, T. Rolland, A. Abad, and I. Trancoso, "Pathological speech detection using x-vector embeddings," *arXiv preprint arXiv:2003.00864*, 2020.

APPENDIX

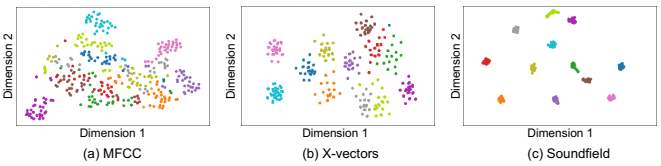


Fig. 20. The t-SNE of MFCC, X-Vector, and *SFF*, the pitch-shifted and original speech of 12 speakers were mixed together for feature extraction.