We Hear Your PACE: Passive Acoustic Localization of Multiple Walking Persons

Article *in* Proceedings of the ACM on Interactive Mobile Wearable and Ubiquitous Technologies · June 2021

DOI: 10.1145/3463510

CITATIONS
0

READS
9

5 authors, including:

Henglin Pu
University of Michigan
7 PUBLICATIONS **30** CITATIONS
SEE PROFILE

Jun Luo
Nanyang Technological University
149 PUBLICATIONS **6,238** CITATIONS
SEE PROFILE

We Hear Your PACE: Passive Acoustic Localization of Multiple Walking Persons

CHAO CAI, Nanyang Technological University, Singapore
HENGLIN PU, University of Michigan, United States
PENG WANG, Huazhong University of Science and Technology, China
ZHE CHEN, Nanyang Technological University, Singapore
JUN LUO, Nanyang Technological University, Singapore

Indoor localization is crucial to enable context-aware applications, but existing solutions mostly require a user to carry a device, so as to **actively** sense location-discriminating signals. However, many applications do not prefer user involvement due to, e.g., the cumbersome of carrying a device. Therefore, solutions that track user locations **passively** can be desirable, yet lack of active user involvement has made passive indoor localization very challenging even for a single person. To this end, we propose Passive Acoustic loCalization of multiple walking pErsons (PACE) as a solution for small-scale indoor scenarios: it passively locates users by pinpointing the positions of their footsteps. In particular, PACE leverages both structure-borne and air-borne *footstep impact sounds* (FIS); it uses structure-borne FIS for range estimations exploiting their acoustic dispersion nature, and it employs air-borne FIS for Angle-of-Arrival (AoA) estimations and person identifications. To combat the low-SNR nature of FIS, PACE innovatively employs domain adversarial adaptation and spectral weighting to ranging/identification and AoA estimations, respectively. We implement a PACE prototype and extensively evaluate its performance in representative environments. The results demonstrate a promising sub-meter localization accuracy with a median error of 30 cm.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: Passive indoor localization, acoustic sensing, ranging, angle-of-arrival, domain adversarial adaptation, user identification

ACM Reference Format:
Chao Cai, Henglin Pu, Peng Wang, Zhe Chen, and Jun Luo. 2021. We Hear Your PACE: Passive Acoustic Localization of Multiple Walking Persons. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 55 (June 2021), 24 pages.
https://doi.org/10.1145/3463510

1 INTRODUCTION

As a key enabling technology for context-aware applications, the market size of indoor localization is predicted to reach 17 billion by 2025 [31]. This promising future has motivated many indoor localization developments in the last three decades, if we label the starting point by seminal proposals such as Active Badge [54] and RADAR [4].

Authors' addresses: Chao Cai, chris.cai@ntu.edu.sg, Nanyang Technological University, Singapore; Henglin Pu, henglinp@umich.edu, University of Michigan, United States; Peng Wang, somewap@hust.edu.cn, Huazhong University of Science and Technology, China; Zhe Chen, chen.zhe@ntu.edu.sg, Nanyang Technological University, Singapore; Jun Luo, luojun@ntu.edu.sg, Nanyang Technological University, Singapore.

2474-9567/2021/6-ART55 $15.00
https://doi.org/10.1145/3463510

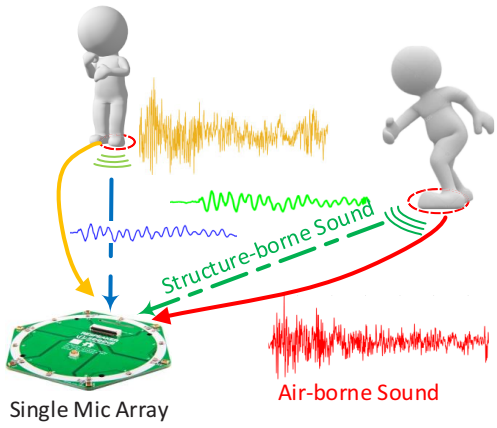


Fig. 1. Using a single microphone array, PACE leverages both structure-borne and air-borne sounds produced by footstep impacts to passively locate and identify multiple walking persons. Essentially, it utilizes the structure-borne and air-borne sounds to estimate range and angle-of-arrival, respectively, and it also exploits the rich features in the air-borne sounds to identify persons. As a passive solution, PACE is cost-effective and readily deployable.

With all these efforts, existing indoor localization solutions have converged into two categories, namely *device-based* (or *active*) and *device-free* (or passive). *Active localization* often relies on a device (e.g., smartphone) held by a target user. Using this device to actively sense ambient signals (either artificial or natural), this approach exploits the critical spatial information embedded in these signals to infer the user's location. Such location-discriminating signals may include, among others, light [57, 62], sound [25, 28, 29], magnetism [46, 61], and radio [51, 58, 60]. Recently appeared *passive localization* lifts the "burden" off users by passively tracking disturbances caused by user presence or motion. Such disturbances can be sensed by monitoring Wi-Fi communications [1, 19, 20, 39] or ambient fields/signals [17, 34, 45].

However, there are small-scale indoor applications that existing solutions fail to handle; we illustrate such applications by two scenarios. In one (elderly and child care) case, Alice leaves her old father Bob and young child Charlie temporarily at home, but she would like to get warned if at least one of them tend to move out of their "safe zones" (e.g., both of them to a slippery floor area or Charlie to a desk corner). In another (workspace management) case, Dave, as the leader of a sensitive project, wishes to keep track of the locations of his team members, so as to maintain a safe distance under the COVID-19 like circumstances and also to avoid "intruders" from interfering the project. The active localization approach (e.g., [4, 23, 25, 54, 62]) is largely infeasible for both cases, as the users may feel cumbersome to carry a device or be unwilling to get tracked, so a passive approach is apparently preferred. Unfortunately, passive localization systems based on Wi-Fi and PIR [19, 20, 34, 39] may suffer severe co-channel interference (especially the communication function of Wi-Fi devices), while Platypus [17] and VoLoc/Symphony [45, 53] require either a heavy sensing infrastructure or user voice to perform localization. Therefore, the open question is: for small-scale indoor scenarios under a short coverage radius, *can we passively track multiple users without heavy infrastructural support and user involvement?*

In order to answer the above question, we propose PACE (Passive Acoustic loCaliza-tion of multiple walking pErsons), a novel concept of tracking users' footsteps for the purpose of passive multi-user localization. PACE adopts a compact microphone array [41] to monitor the *footstep impact sounds* (FIS) produced by user walking, as shown in Fig. 1; it relies on the FIS to locate and identify multiple users simultaneously. Such a light-weight system incurs minimal infrastructure requirements and deployment costs, especially suitable for indoor spaces such as home, office, museum, and library. However, implementing PACE faces two practical challenges. On

one hand, indoor environments normally incur a heavy multi-path effect and also have a strong acoustic noise background; these have made pure *angle of arrival* (AoA) enabled solutions (e.g., [45]) insufficient to locate footsteps. On the other hand, while it is intuitive that human hearing can clearly identify acquainted footsteps, this identification is not readily achievable via acoustic sensing.

In designing PACE, we leverage the existence of two propagation components of FIS (i.e., *structure-borne* and *air-borne*) and their complementarity to tackle the above challenges. First of all, we use the sharp difference in propagation speeds to differentiate the two components. Secondly, we employ structure-borne FIS for range estimation exploiting their acoustic dispersion nature, but we rely on air-borne FIS for AoA estimation and user identification leveraging their rich features. Thirdly, we apply domain adversarial adaptation [14] for both ranging and user identification, aiming to extract domain independent features and thus enable generalizable functionalities, so as to handle unseen users and environments without relying on extensive training samples. Finally, though adopting a model-based approach for AoA estimation, we introduce a novel spectral weighting technique to sharpen peaks in a correlation spectrum and hence to achieve a finer resolution. To summarize, our paper makes the following major contributions:

- We propose PACE as the first acoustic localization system for passively tracking user footsteps.
- We study the complementarity of structure-borne and air-borne FIS, providing foundations for PACE design.
- We propose a novel spectral weighting technique to combat the low-SNR nature of air-borne FIS, thus improving the AoA estimation accuracy.
- We design a deep neural network with domain adversarial adaptation; it performs ranging and user identification without relying on extensive training samples.
- We implement a PACE prototype and extensively evaluate its performance in realistic settings. The results demonstrate a median localization error of only 30cm. we will open our source codes after paper acceptance.

In the following, we first provide motivations in Section 2, then present technical details of PACE design in Sections 3 and 4. We report our extensive evaluations in Section 5, discuss literature and limitations in Sections 6, and finally conclude our paper in Section 7.

2 BACKGROUND AND MOTIVATION

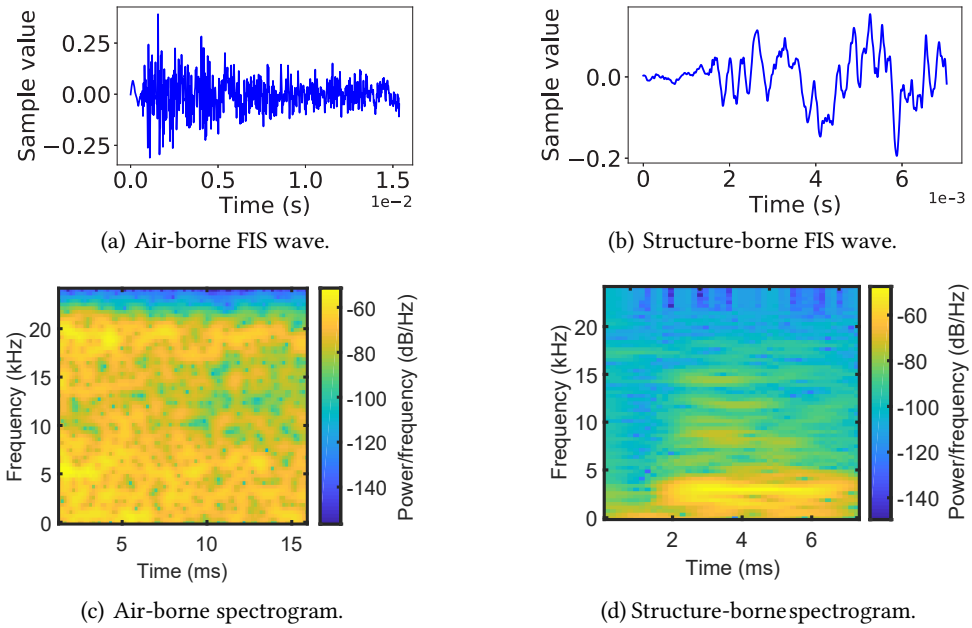
In this section, we explain the background of impact sounds, and motivate the PACE design via brief measurement studies with a 6-mic array [41] put on the floor.

2.1 Basics of Impact Sound

When an object impacts a surface, it causes vibrations at the impact point and thus radiates energy via both air and the solid medium behind the surface in acoustic waveforms. These acoustic waveforms (*impact sounds*, or IS) contain two major components. The *air-borne* IS has a constant speed c_a and is non-dispersive, so these waveforms retain their shape regardless of how long they propagate. The *structure-borne* IS traversing in solid media exhibits *acoustic dispersion*; in other words, high frequency components travel faster than low frequency ones. The certain speed c_f of a specific frequency component f could be defined as [42]:

$$c_f = \sqrt[4]{\frac{Ehf^2}{12\rho\left(1-v_p^2\right)}}, \tag{1}$$

where v_p is the phase velocity, and E, ρ, h are constants that characterize a medium: E quantifies elasticity, ρ characterizes stiffness, and h represents thickness. Therefore, when observing the structure-borne IS at different distances from the impact point, the resulting waveforms exhibit distinctive features, which can be leveraged to



(a) Air-borne FIS wave.

(b) Structure-borne FIS wave.

(c) Air-borne spectrogram.

(d) Structure-borne spectrogram.

Fig. 2. Air-borne and structure-borne FIS incurred by a footstep impact: time-domain waveforms (a) and (b), time-frequency representations (c) and (d).

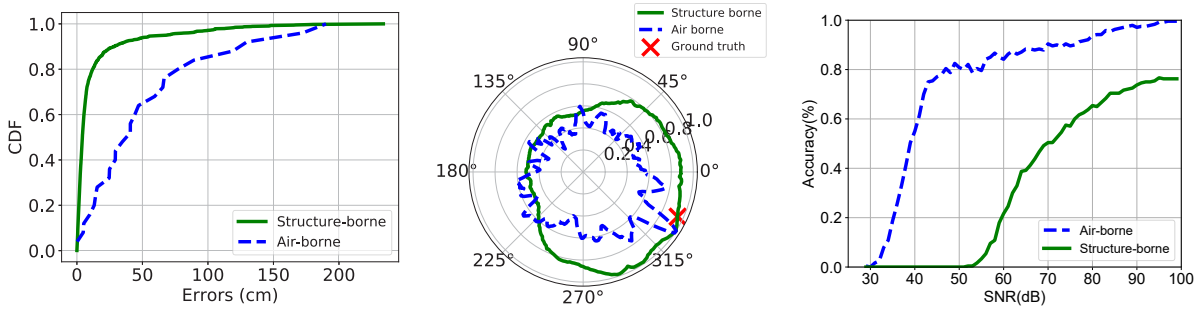
conduct accurate ranging. With a sampling rate of 192 kHz to preserve waveform details, we give two typical segments of both FIS components in Fig. 2. The very different time and frequency features of these two components are clearly visible, which motivates us to make the best use of them respectively.

2.2 Complementarity of Air-borne and Structure-borne FIS

As locating and identifying multiple walking persons can be a very challenging task, conventional acoustic localization schemes (e.g., relying only on AoA estimations) certainly do not work. Fortunately, the measurement study we present in this section explain that there exists a complementarity between air-borne and structure-borne FIS, which can be exploited to complete this task.

In the first measurement, we show that using structure-borne FIS can achieve a more accurate ranging than using air-borne FIS. With air-borne FIS, we get no choice but to use a path loss based model [10], as the signal strength appears to be the only available feature. To achieve a robust performance under temporal signal fluctuations, we use the mean signal energy within a sliding window to derive the signal strength of the path loss model. For structure-borne FIS, ranging requires a novel mechanism to leverage the acoustic dispersion (detailed proposal will be presented in Section 4.1). The results shown in Fig. 3 (a) clearly demonstrate that ranging based on structure-borne FIS is noticeably more accurate than that with air-borne FIS.

In our second measurement, we compare the performance of AoA estimation using these two signals. We run a delay-and-sum beamforming algorithm (details presented in Section 3.3) with a compact 6-mic array [41] to estimate AoA. Fig. 3(b) depicts the corresponding AoA spectra for respective signals. It is observable that with air-borne FIS, the beam pattern is much sharper and hence more robust to background interference. The reason for the inferior AoA estimation performance of structure-borne FIS is twofold: i) the signal is transient (see

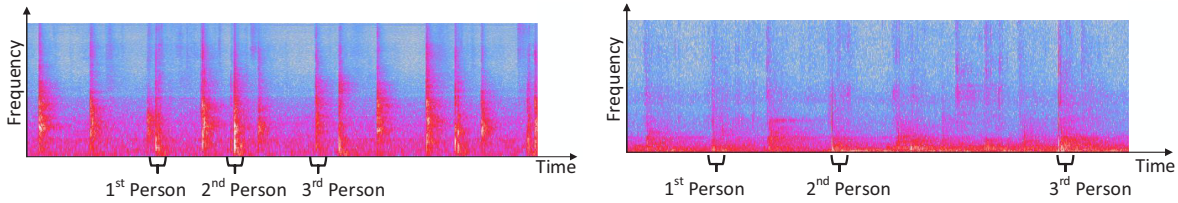


(a) Ranging performance. (b) AoA estimation performance. (c) User identification accuracy.

Fig. 3. Performance comparisons between structure-borne and air-borne FIS produced by footstep impact events, in terms of (a) ranging, (b) AoA estimation, and (c) user identification.

Section 2.3 for details), and ii) the acoustic dispersion may cause waveform distortions at different microphones in the array. This second reason renders correlation-based AoA estimation virtually invalid.

In our third measurement, we record the FIS of multiple users walking. The results shown in Fig. 4 present Mel-Frequency Cepstra (MFC) [32] of FIS by three users walking simultaneously on two different floors; which clearly demonstrates that FIS profiles produced by different users are separable in time, as far as they do not have synchronized paces: two steps are separated by at least 0.1 s. We also collect a set of FIS samples and use these samples to train a Gausian mixture model (GMM) [37], so as to identify other FIS samples whose mutual distances range from 0.025 m to 3.6 m. According to Fig. 3(c), the superiority of air-borne FIS is apparent: it requires less than 50 dB to achieve a 80% recognition accuracy, which can never be achieved by structure-borne FIS. Nonetheless, these results also show that GMM is incapable of performing user identification in reality, as 50 dB SNR is rarely attainable in practice; similar situations apply to other conventional processing methods summarized in [11].



(a) MFC for three users walking on a wood floor. (b) MFC for three users walking on a concrete floor.

Fig. 4. MFCs under different scenarios, where FIS belonging to three users are partially labelled only for exemplary purposes. They show that features produced by different users are separable in time domain.

2.3 Separating Air-borne FIS from Structure-borne FIS

In all aforementioned experiments, we implicitly assume that air-borne and structure-borne FIS are separable, now we justify this assumption. It is known that structure-borne sound travels at a speed around 3000 m/s (depending on the specific medium) while air-borne sound travels at 340 m/s (at a temperature of 15°C) [7]. This sharp propagation speed difference gives us a chance to separate these two types of FIS in time domain, given an adequate sampling rate. For instance, assume that the impact happens at 1 m distance from a microphone receiver

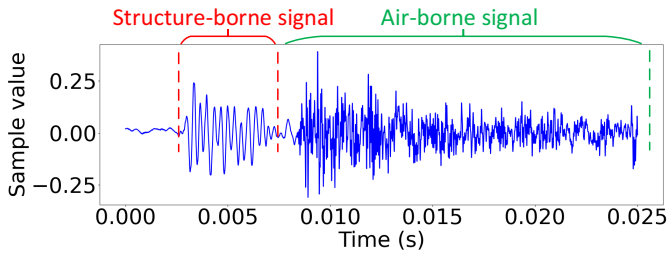


Fig. 5. Structure-borne and air-borne FIS appear in order, due to their distinctive propagation speeds.

whose sampling rate is $f_s = 192\text{kHz}$, and the structure-borne and air-borne propagation speeds are $c_s = 3000\text{m/s}$ and $c_a = 340\text{m/s}$, respectively, then a clean set of unpolluted structure-borne FIS should last for $f_s\left(\frac{1}{c_a} - \frac{1}{c_s}\right) \approx 500$ samples, or a equivalent of 2.6 ms. Therefore, the structure-borne component can be obtained by extracting samples within 2.6 ms at the beginning of individual FIS, and the remaining samples can be categorized into the air-borne component. Fig. 5 depicts the full-length waveform produced by a *footstep impact event* (FIE), clearly demonstrating the separable nature of the two components due to their propagation speed difference.

3 MODEL-BASED SIGNAL PROCESSING FOR PACE

As shown in Fig. 6, PACE mainly consists of four modules: signal detection, beamformer, ranging, and identification. The signal detection module extracts legitimate FIS for further processing. After splitting FIS into structure-borne and air-borne signals, the beamformer and identification modules utilize air-borne signals for AoA estimation and user identification, respectively. And the structure-borne signals are exploited for ranging. Integrating AoAs and ranges allows PACE to acquire accurate user locations, while user identification enables PACE to differentiate them. In the following, we focus only on the model-based signal processing modules, but leaving the model-free modules to be discussed in Section 4.

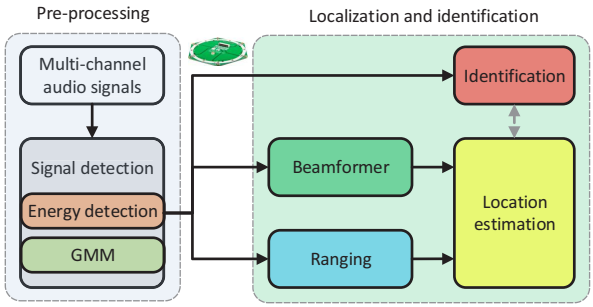
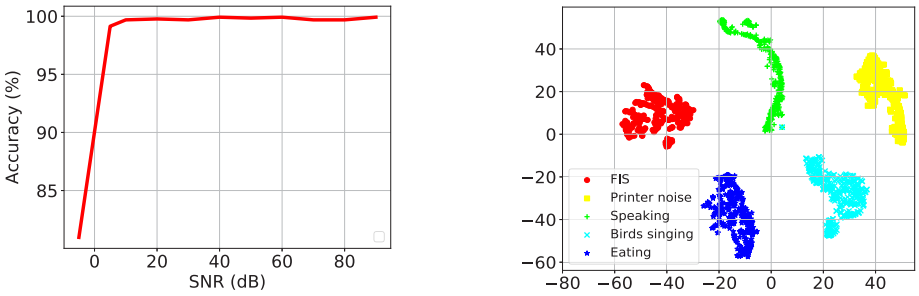


Fig. 6. PACE system architecture.

3.1 Signal Detection

As individual FIS have very short time duration (less than 0.1 s), a remote microphone always detects a sequence of abrupt energy changes. In PACE, we quantify this energy using Root Mean Square (RMS). Taking $\mathbf{x} = \{x_1, x_2, \cdots, x_L\}$ as a received acoustic frame within a certain time window, then its energy is defined as $E_{\text{RMS}}(\mathbf{x}) = \sum_{i=1}^{L} \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_L^2}{L}}$. One could signal the detection of an impact event if E_{RMS} goes above a certain threshold.



(a) Accuracy vs SNR. (b) Visualization

Fig. 7. FIS detection performance.

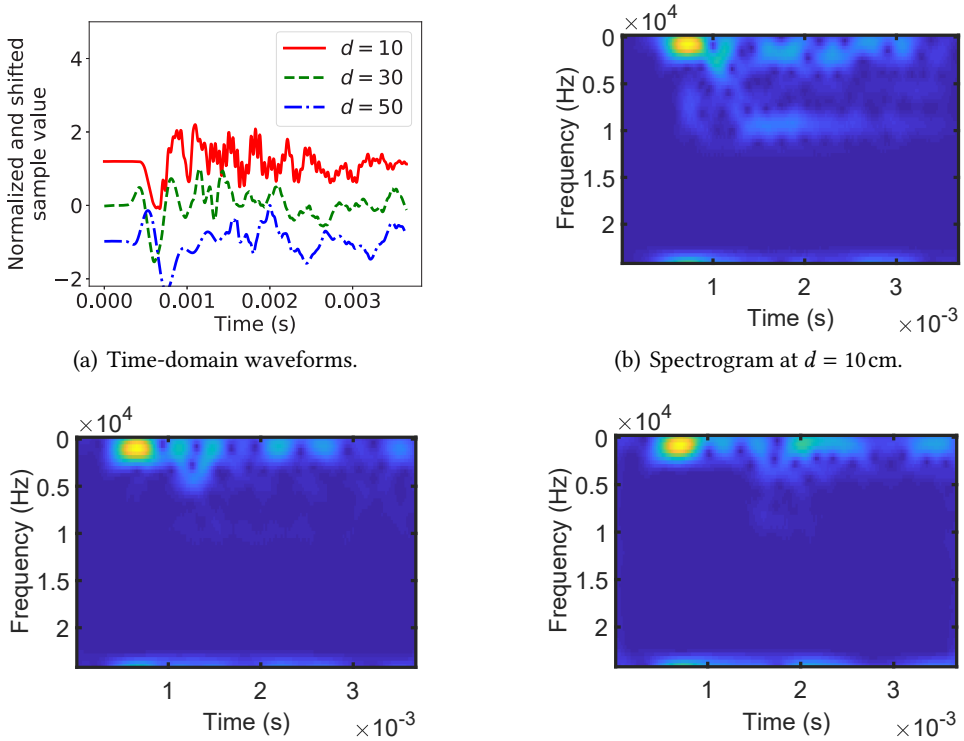
However, many background noises such as people speaking or cutlery collisions can also exhibit high E_{RMS} that even go beyond that of FIS. Therefore, FIS detection purely relying on signal energy is not robust due to its high false detection rate.

To address the above problem, we utilize GMM [37] to further identify whether a detected sound indicates an FIE. The GMM module makes the energy-based detection strategy less sensitive to the threshold, resulting in a performance more robust to low SNR. We train our GMM model against common background noise, which leads to an almost perfect performance as shown in Fig 7. These results are obtained with as few as 30 training samples collected at 3 different locations. More detailed configurations for this GMM model can be found in Section 5.1. Unfortunately, directly applying GMM for user identification is not feasible for the required high SNR, as discussed in Section 2.2. Since GMM cannot fully characterize temporal-spectral dynamic patterns for non-speech signals, we resort to a deep learning approach that will be elaborated in Section 4.2.

3.2 Ranging Is Challenging

As explained in Section 2.1, the acoustic dispersion of the structure-borne FIS may be exploited to estimate the distance from an FIE to a microphone. This potential is also confirmed by Fig. 8, which shows a clear trend of signal variation in distance. In order to leverage the divergence of the propagation speeds of different frequency components for ranging, we first simplify Eqn. (1) to $c = kf^{\frac{1}{2}}$, where k is a constant. An intuitive solution is to transmit several modulated signals at certain known distance to calibrate the constant k [21]. Consequently, when an FIE happens, we just need to separate N different frequency components from structure-borne FIS, say using Wiener-Ville Distribution to profile their relative arrival time $t_i, i \in [1, N]$. Then, for a certain distance d, we have $\frac{d}{c_{f_1}} - \frac{d}{c_{f_2}} = t_1 - t_2$, where c_{f_1} and c_{f_2} are the propagation speeds for frequencies f_1 and f_2, respectively. Therefore, $d = \frac{t_1 - t_2}{\frac{1}{c_{f1}} - \frac{1}{c_{f2}}}$ and we can involve more frequency components to improve the estimation accuracy.

However, the aforementioned ideal model faces two major issues. On one hand, our measurements reveal that different frequency components of structure-borne FIS are overlapped in a relatively short duration (mostly less than 5 ms). As it is already difficult to obtain an accurate frequency spectrum with very limited samples, separating mixed frequency components could be more challenging. This challenge could cause inaccuracy in estimating both frequency components and arrival times, thereby significantly affect the ranging performance. On the other hand, solid media often cause non-quantifiable attenuation that drastically increases in frequency, so FIS received at a further distance tend to lose their high frequency components, as shown in Fig. 8. These two problems indicate that a model-based approach is highly unlikely to handle this ranging problem well, so we instead adopt a deep learning based approach in Section 4.1.



(a) Time-domain waveforms.

(b) Spectrogram at $d = 10$ cm.

(c) Spectrogram at $d = 30$ cm.

(d) Spectrogram at $d = 50$ cm.

Fig. 8. Structure-borne FIS captured at different distances: time-domain waveforms at three distances (a), and the respective frequency representations (b)–(d).

3.3 AoA Estimation with Spectral Weighting

Unlike ranging, AoA estimation can be tackled by a model-based approach thanks to the diversity offered by the microphone array. In the following, we first introduce the basics and weakness of beamforming for AoA estimation. Then we present our spectral weighting method to combat the weakness, and we also explore the elevation dimension to handle multi-path interference.

Basics and Weakness of Beamforming. The widely used *delay-and-sum beamforming* estimates an AoA by searching the maximum energy formed by multiple microphones over a spherical grid [2, 40]. Essentially, beamforming for an AoA ϕ can be formulated as a maximum likelihood estimation:

$$\phi = \arg\max_{\theta} e\left(\boldsymbol{\tau}|\theta\right), \tag{2}$$

where e denotes the output energy, and $\boldsymbol{\tau}$, depending on a possible incident angle θ, represents the delays among multiple microphones. When $\theta = \phi$, the signals received by multiple microphones are made in-phase via cross-correlation, thereby maximizing the output energy. However, this algorithm is vulnerable to interference and often has low resolution when using unmodulated signals due to the fact that their cross-correlation peaks are not sharp enough [9, 40]. To achieve a better performance, it requires the input signals to have good pulse compression properties [26], which, unfortunately, do not hold for FIS. Consequently, the obtained beam pattern has very low resolution, and is hence susceptible to background interference.

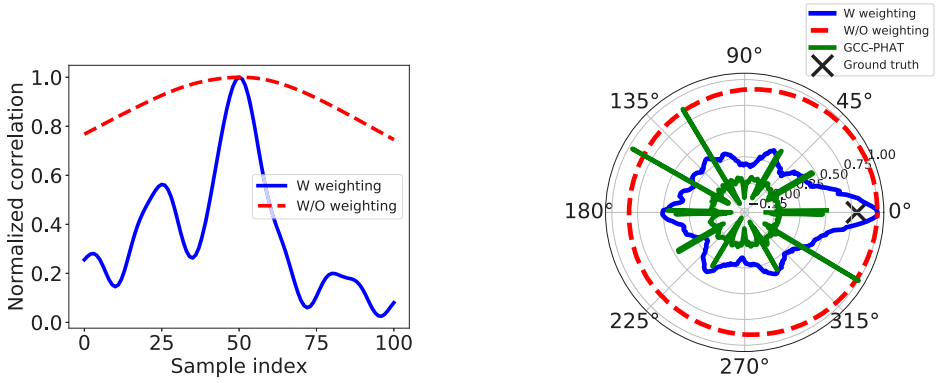
To better illustrate the issue, we conduct measurements using a circular array [41] sampling at 192 kHz. As the array has a diameter of 9 cm, the maximum delay is 2.6×10^{-4} s, equivalent to 50 samples. This implies that, in order to get better results, the correlation peak should have a steep gradient with a maximum of 50 samples in both negative and positive time delays. However, this could hardly be achievable by the basic beamforming method if taking FIS as input signals. As shown in Fig. 9(a), the gradient around the maximum peak is quite flat, leading to a rather wide beam pattern shown in Fig. 9(b) though the AoA is correct.

Spectral Weighting. To tackle the above issue, we design a new spectral weighting function: a multiplier applied to the frequency domain because FFT is used to speed up the correlation computations. Common spectral weighting techniques (e.g., GCC-PHAT [22]) yield sharp correlation peaks for both signals and noises, possibly leading to wrong AoA estimations shown in Fig. 9(b). To prioritize the contribution of each frequency component $X(f)$, we introduce a new weighting function adapting to SNR as a multiplier for $X(f)$:

$$W(f) = G(f)\,|X(f)|^{-\rho}\,, \tag{3}$$

where $\rho = \max\left\{\beta, \frac{X(f)-\alpha X_\sigma(f)}{X(f)}\right\}$, $X_\sigma(f)$ is the mean spectral power of noise (estimated in the absence of source signals), $\alpha \leq 1$ is a coefficient quantifying how conservative the estimated noise power is (default value 0.9), and β is a threshold normally set to 0.4. The term $\frac{X(f)-\alpha X_\sigma(f)}{X(f)}$ (hence ρ) approaches 1 when SNR is high, in which case $|X(f)|^\rho$ becomes the spectral magnitude. Otherwise, ρ is reduced to increase $W(f)$ and in turn to compensate $X(f)$, but noises (whose SNR below β) is not further compensated. In other words, unlike GCC-PHAT that blindly equalizes every frequency component, we prioritize the contributions of these components based on their respectively SNR. In particular, $G(f)$ is the Wiener function of *a priori* SNR ξ, utilized to preserve performance at low SNR and is estimated for the current n-th frame by $G_n(f) = \xi_n\,(\xi_n + 1)^{-1}$, where ξ_n is an estimate of the *a priori* SNR and could be estimated using decision-directed approach [12]:

$$\xi_n = \frac{\gamma\,[G_{n-1}(f)]^2\,|X_{n-1}(f)|^2 + (1-\gamma)\,[X_n(f)]^2}{\mathbf{E}\left\{\left[X_{\sigma,n}(f)\right]^2\right\}},$$



(a) Correlation spectra between two arbitrary microphones, w/ and w/o spectral weighting.

(b) Beam patterns with all 6 microphones, w/, w/o spectral weighting, and GCC-PHAT.

Fig. 9. Performance comparisons: with spectral weighting, peaks become significantly sharper than those without, in both (a) correlation spectra and (b) radical beam patterns, showcasing the effectiveness of our proposed weighting function.

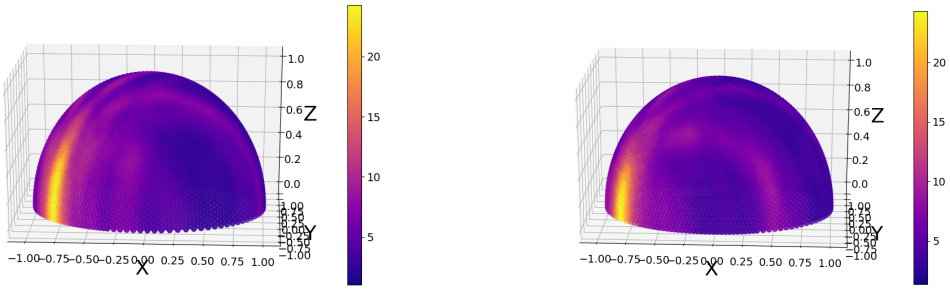
where γ is a constant normally set to 0.9, and $\mathbf{E}\left\{\left[X_{\sigma,n}(f)\right]^2\right\}$ is updated for every new frame. The final weighting adaptation to SNR $W(f)X(f)$ is translated back (via IFFT) to the correlation spectra shown in Fig. 9: because SNR is higher along the signal arrival direction (i.e., AoA) than other directions, the superiority of our proposed spectral weighting technique in distinguishing the correct peaks is evident.

Dealing with Multi-path Effects. To make the AoA estimation robust to multi-path interference, we perform 3D beamforming and utilize the elevation angles to filter out AoAs produced by multi-path effect. Essentially, we extend our objective function to 3D as:

$$\phi_a = \arg\max_{\theta_a} e\left(\boldsymbol{\tau}|\theta_a, \theta_e\right), \tag{4}$$

where θ_a and θ_e are the azimuth and elevation angles, respectively, ϕ_a is the optimal solution to the maximum likelihood estimation. The previous 2D-beamforming can be regarded as a special case where $\theta_e = 0$.

Given air-borne FIS radiating hemispherically from its source on a floor, reflected signals project more energy onto the elevation whose azimuth angle coincides with that of the direct signal, as signals incident from other azimuth angles are more severely diffused. Therefore, if our 2D-beamforming obtains more than one ϕ_a, we shall search over the corresponding elevation angles θ_e in the 3D beamforming results. We remove any ϕ_a whose e sharply decreases in θ_e, as it is the AoA of some reflected signals. This effect, illustrated in Fig. 10, helps us to handle the multi-path interference.



(a) Direct signal.

(b) Multi-path signal.

Fig. 10. 3D heatmaps to illustrate multi-path handling. A direct-path AoA spectrum (a) differs significantly from a multi-path AoA spectrum (b), in terms of projected energy along the elevation dimension.

4 ASSEMBLING 'PACE' TOGETHER

In this section, we focus on model-free approaches to ranging and identification, so as to compensate what cannot be solely achieved by model-based techniques and thus to complete PACE.

4.1 Ranging Reloaded

According to Section 3.2, the function relation between FIS (structure-borne) and distance is too complicated to be explicitly evaluated. Consequently, we formulate the ranging problem as a *regression*: given a specific FIS waveform $\mathbf{x}, \mathbf{x} \in X$, we aim to learn a function $\mathcal{G} : X \rightarrow D$, where X and D denotes input (FIS waveform) and distance spaces, respectively. In reality, an input FIS waveform $\mathbf{x}$ is sampled from a joint distribution $P(\mathbf{x}, d, s)$, where $d \in D$ and $s \in S$, and S contains *domain* specific properties incurred by scene settings such as user diversity and environment dynamics. Apparently, only features characterizing the joint distribution $P(\mathbf{x}, d)$ are expected

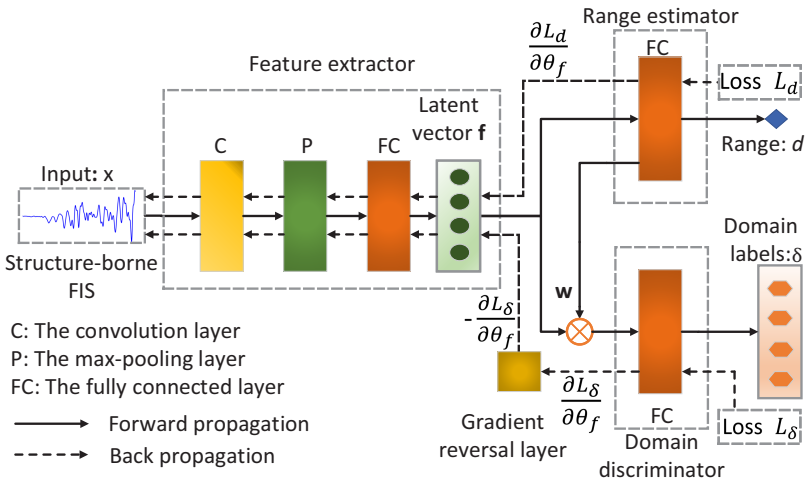


Fig. 11. Network architecture of R-Net.

for our ranging purpose but features induced by s should be eliminated. To this end, we use a deep neural network (DNN) to approximate $\mathcal{G}(\mathbf{x})$ and adopt a domain adversarial adaptation [14, 36] to exclude the impact from s.

Our R-Net, a DNN shown in Fig. 11, consists of three parts, namely feature extractor, range estimator, and domain discriminator. The feature extractor transforms an input FIS waveform $\mathbf{x}$ into $\mathbf{f} \in \mathbb{R}^Q$, a lower-dimensional feature vector. Using $\mathbf{f}$, the range estimator aims to infer an accurate distance, while the domain discriminator tries to identify different domains. The input to the domain discriminator is a weighted $\mathbf{f}$ whose weights are extracted from the range estimator. The ultimate goal of this network is to obtain a *domain independent* vector $\mathbf{f}$ so as to i) achieve an accuracy range estimation and ii) cheat the domain discriminator to fail its task. It is this domain independent $\mathbf{f}$ (with only range-specific features) that enables a cross-domain generalization of R-Net.

Most DNNs tackle acoustic signals in the format of spectrogram or transformed features such as MFCC or GFCC [8]. However, our R-Net takes the raw time-domain structure-born FIS as inputs, because spectrogram or MFCC reveals only magnitude information and thus loses phase information that is critical for ranging. As demonstrated in [48, 52, 59], phase contains finer-grained temporal information than any other features. The rationale is that phase preserves time associated metrics in a continuous form while other features often sample these metrics in a discrete manner. In training R-Net, we provide three sets of training data: i) the raw structure-borne FIS set X, ii) its corresponding ground truth range set D, and iii) the domain labels set Δ. The feature extractor G_f has one convolution layer followed by a pooling layer, a dropout layer, and a fully connected (FC) layer. It maps the input $\mathbf{x}$ to a lower-dimensional feature vector $\mathbf{f}$:

$$\mathbf{f} = G_f(\mathbf{x}; \boldsymbol{\theta}_f),$$

where G_f is parameterized by weight vector $\boldsymbol{\theta}_f$. The range estimator G_r consists of one FC layer parameterized by $\boldsymbol{\theta}_r$; it infers range by $\hat{d} = G_r(\mathbf{f}; \boldsymbol{\theta}_r)$ with the feature vector $\mathbf{f}$. Given the range training set D, we use mean square error as the loss function to learn the parameters $\boldsymbol{\theta}_r$:

$$\mathcal{L}_r = \frac{1}{|D|} \sum_{i=1}^{|D|} \left| d_i - \hat{d}_i \right|^2. \tag{5}$$

The domain discriminator utilizes a weighted feature vector as its input. This is due to the fact that waveforms of structure-borne FIS generated at different ranges exhibit distinctive features; if we do not account for this

range information, the domain discriminator may treat range as a special domain, which is contradictory to our ultimate goal of identifying only scene settings. To exclude range information for classifier, R-Net lets the domain discriminator take a weighted feature vector as its input: $\mathbf{k} = \sum_{i=1}^{Q} w_{\theta_r,i} f_i$, where $\mathbf{w}_{\theta_r}$ represents weight vector from the hidden layer of the range estimator G_r.

The domain discriminator G_δ also consists of an FC layer but with a softmax activation function, projecting input $\mathbf{k}$ into a *predicted probability* $\hat{\delta}$:

$$\hat{\delta} = G_\delta(\mathbf{k}; \boldsymbol{\theta}_\delta), \tag{6}$$

where $\boldsymbol{\theta}_\delta$ is the network parameter for G_δ. Here we make use of categorical cross-entropy as the loss function:

$$\mathscr{L}_\delta = -\frac{1}{|X|} \sum_{i=1}^{|X|} \sum_{j=1}^{|\Delta|} \log(\hat{\delta}_{i,j}), \tag{7}$$

where $\hat{\delta}_{i,j}$ is the predicted probability indicating the relation between the i-th FIS sample and the j-th domain.

Based on the aforementioned design and analysis, our final loss function is formulated as:

$$\mathscr{L} = \mathscr{L}_r - \alpha \mathscr{L}_\delta. \tag{8}$$

When training, $\boldsymbol{\theta}_r$ and $\boldsymbol{\theta}_\delta$ aim to minimize their respective loss functions $\mathscr{L}_r$ and $\mathscr{L}_\delta$, so their objectives are "adversarial" to each other due to the minus sign in Eqn. (8). While $\boldsymbol{\theta}_f$ also aims to minimize $\mathscr{L}_r$, the gradient reversal layer (active only during backpropagation) enables $\boldsymbol{\theta}_f$ to cheat G_δ by maximizing $\mathscr{L}_\delta$ and thus minimize $\mathscr{L}$. The outcome of this adversarial adaptation ensures that the feature extractor learns to extract only range-specific features and to neglect those induced by scene settings. Consequently, R-Net can readily handle FIS samples taken from unseen domains.

4.2 User Identification

As explained in Section 2.2, PACE exploits feature-rich air-borne FIS for user identification. Note that this function is only meant to differentiate users for the sake of multi-user localization, we shall extend it for user authentication purpose in an ongoing work [6]. In order to "filter out" the interference from scene settings, our identification network I-Net has a similar architecture to R-Net (so shared symbols and concepts shall be reused later), but I-Net differs from R-Net in three aspects: i) range estimator is replaced by a categorical classifier for identifying users, ii) range information, together with environment dynamics, become domain specific properties; they should be eliminated via the same adversarial adaptation procedure as for R-Net, iii) most importantly, as it is impractical to pre-collect air-borne FIS from all users, we require I-Net's feature extractor to learn features that are sufficiently discriminative and generalizable for identifying unseen users. To achieve this last objective, we introduce a center loss [56] in training the classifier:

$$\mathscr{L}_C = \frac{1}{2} \sum_{i=1}^{|X|} \left\| \mathbf{f}_i - \mathbf{c}_{y_i} \right\|_2^2, \tag{9}$$

where $\mathbf{c}_{y_i} \in \mathbb{R}^Q$ is the center for the y_i-th class deep features, $\mathbf{f}_i \in \mathbb{R}^Q$ is the i-th deep feature, and the summation is performed over the input set $|X|$. We update $\mathbf{c}_{y_i}$ in a mini-batch (with size m) manner where the gradient of $\mathscr{L}_C$ with respect to $\mathbf{f}_i$ and the update to $\mathbf{c}_j$ are calculated as:

$$\begin{aligned} \frac{\partial \mathscr{L}_C}{\partial \mathbf{f}_i} &= \mathbf{f}_i - \mathbf{c}_{y_i}, \\ \mathbf{c}_j &= \mathbf{c}_j + \frac{\sum_{i=1}^{|X|} \mathscr{I}(y_i = j)(\mathbf{c}_j - \mathbf{f}_i)}{1 + \sum_{i=1}^{|X|} \mathscr{I}(y_i = j)}, \quad \forall j, \end{aligned} \tag{10}$$

where $\mathcal{I}(y_i = j)$ is an *indicator function* whose value is 1 if $y_i = j$; otherwise 0. We combine the center loss $\mathcal{L}_c$ with a categorical cross-entropy loss $\mathcal{L}_s$ to train the user classifier:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_s + \beta \mathcal{L}_c \\ &= -\sum_{i=1}^{|X|} \log \frac{e^{\mathbf{w}_{y_i}^T \mathbf{f}_i + \mathbf{b} y_i}}{\sum e^{\mathbf{w}_j^T \mathbf{f}_i + b_j}} + \frac{\beta}{2} \sum_{i=1}^{|X|} \left\| \mathbf{f}_i - \mathbf{c}_{y_i} \right\|_2^2, \end{aligned} \tag{11}$$

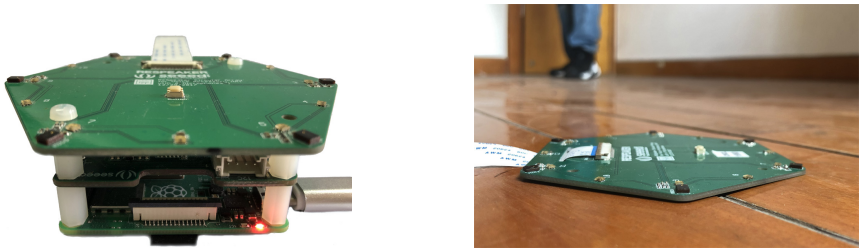
where $\mathbf{w}$ and $\mathbf{b}$ are the FC parameters, β is a scalar that balances the two losses. This loss function enables I-Net to maximize inter-class margins and minimize intra-class distances, thereby improving its generalizability.

5 SYSTEM EVALUATION

In this section, we present extensive evaluation studies on our PACE prototype. We refrain from system-level comparisons with other up-to-date passive indoor localization proposals (e.g., [38, 45]), because their application scenarios are very different as explained in Section 1.

5.1 Prototype and Experiment Settings

We implement our PACE prototype using a circular array with six microphones [41] running on a Raspberry Pi 4, as depicted in Fig. 12(a). We configure the sampling rate as 192 kHz with a 32-bit resolution, the highest achievable configurations on this hardware platform. The model-based approaches in PACE, i.e., signal detection and beamformer, are implemented in C. The model-free methods, i.e., ranging and identification, are implemented using Tensorflow [49]. Our localization experiments involve 3 indoor spaces with different floor materials (concrete, engineered wood, and solid wood) and 5 users (3 men and 2 women) each having three different choices on footwear (slippers, sneakers, and dress shoes), but the later user identification experiments in Sec. 5.2.3 further involve 3 extra users (1 men and 2 women) with other conditions remaining the same. The dimensions of the 3 spaces are $3.6 \times 8\,\mathrm{m}^2$, $4 \times 5\,\mathrm{m}^2$, and $3 \times 4\,\mathrm{m}^2$.



(a) Hardware platform. (b) Experiment setting.

Fig. 12. Implementing PACE with a microphone array (a) and corresponding experiment setting for evaluation (b).

In the signal detection module, the window length for the energy detection is set to 1 ms and the threshold for the detector is 0.1. For the GMM-based filter, we decimate audio samples with a factor of 12 to recognize FIS against various background noises and to achieve computational efficiency as well. Our GMM filter leverages MFC coefficients for recognition and has the following empirically set parameters: 16 mixture components, 20 ms duration for a phoneme, as well as 20 filters and MFC coefficients. We use one of the indoor spaces to train our GMM modules under 5 common background noises (bird sounds, human voice, phone ringings, musics, and printer noises); the remaining two indoor spaces are then used for testing. We also make sure that the samples concerning each noise type collected in one indoor room have a quantity of at least 30. Details of the model-free

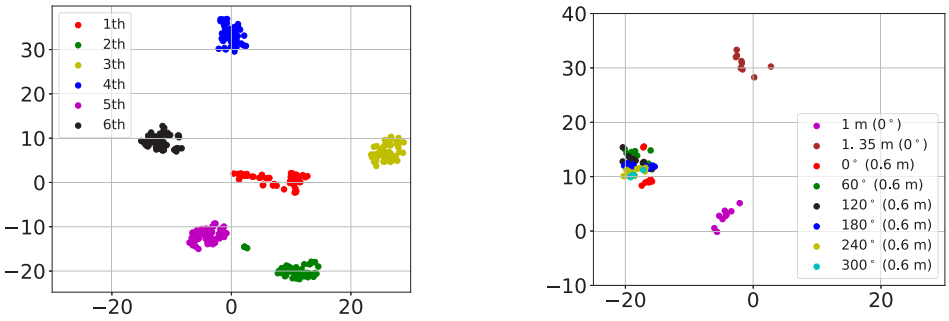
modules (including DNNs and their respective training datasets) are published on GitHub and will be made open-source after paper acceptance.

To collect data for training neural networks and evaluations, we conduct measurements under various indoor settings; one of them is shown in Fig. 12(b). To minimize the impact of the noises on training, we adopt a spectral subtraction technique [5] to remove them. To gather location ground truth along with FIS data collection, we randomly distribute 10 piezoelectric sensors stuck on the floor and pre-measure their ground truth locations. These piezoelectric sensors are connected via an ESP32 [13] wireless transmitter that is synchronized with PACE through a local WLAN. The sensitivity of these piezoelectric sensors are carefully tuned so that, only when a foot steps on it, the sensor output may saturate and signal PACE to mark this FIS profile as having a ground truth location. We let the 5 users randomly walk in each indoor setting, until we obtain at least 60 FIS samples for each user at every ground truth location. We re-deploy the sensors and repeat the above data collection process 10 times. In order to perform domain adversarial adaptation, we need two choices for each of the three factors defining an experiment setting (i.e., floor material, user, and footwear), so we use 8 settings for training. For each chosen setting, we further split the collected samples into training and testing ones with a ratio of 3:1. The labels for the training and testing samples are made different, in order to fairly evaluate the prediction performance. Later experiment statistics are all obtained by repeating each testing case with at least 1,000 testing samples.

5.2 Microbechmarks

We first evaluate the performance of individual components in PACE.

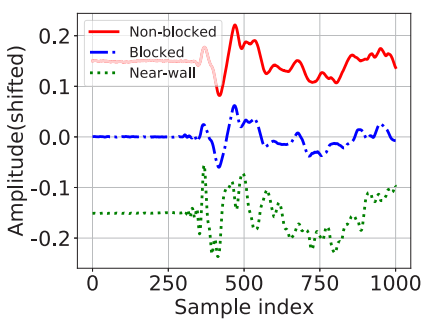
5.2.1 Ranging. To start with, we first confirm that structure-borne FIS indeed carry distance information. In other words, we verify if structure-borne FIS at different locations exhibit distinctive features, and if signal features generated at the same location are consistent. In this measurement, we collect FIS at six different locations with 10 cm interval. At each location, we collect over 60 FIS samples. We then use t-SNE [50] to reduce the data dimension to 2D and visualize the results in Fig. 13(a). The figure clearly shows that structure-borne FIS generated at different locations exhibit distinctive features, while they share similar features at the same location. This observation confirms the feasibility of utilizing structure-borne FIS for ranging. We also conduct measurements to check if structure-borne FIS have directional radiation patterns, as otherwise ranging can be interfered by arrival directions. To our relief, the results shown in Fig. 13(b) reveal that signal properties from different angles share similar features, hence arrival directions cause no interference.



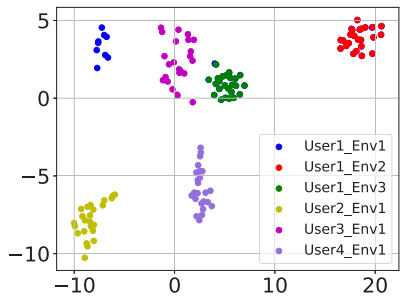
(a) Impact of range.

(b) Impact of angle.

Fig. 13. Structure-borne FIS offer discriminative features for ranging (a), which are not interfered by arrival directions (b).



(a) FIS waveforms under three different cases.



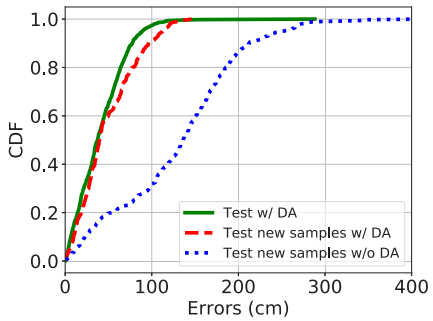
(b) 2D embeddings of FIS under different scene settings.

Fig. 14. Structure-borne FIS are barely altered by in-path blockage and reflections (a), but those from different scene settings exhibit distinctive features (b).

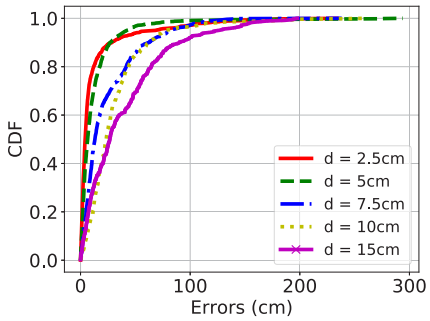
We also explore whether in-path blockages can incur scattering or absorbing effects and thus distort the identifiable signal properties. In this experiment, we first record an FIS clip 3 m away from an FIE where the line-of-sight path is not blocked. For comparison, we let another user standing in the middle, and record another FIS clip, and we also record an FIS clip when the microphone is placed close to a wall. Fig. 14(a) depicts these time-domain waveforms: in-path blockage introduces no noticeable impacts on FIS waveforms as the red and green curves shown in Fig. 14(a) are almost identical. This observation implies that it is feasible to have multiple users and furniture in a room even if some of them are blocked by others from time to time. Although the waveforms may slightly change when placing the microphone close to walls, this impact of reflection can be readily handled by the adversarial learning strategy adopted to train the R-Net. We finally collected FIS from different scene settings, namely different users or environments, at the same distance and visualize their low-dimensional embeddings in Fig. 14(b). The results confirm that FIS exhibit distinctive features incurred by different scene settings, strongly indicating the need for the domain adaptation.

After the aforementioned studies, we now present the overall ranging performance. We firstly demonstrate that R-Net is capable of extracting domain-independent features for ranging and thus can work across heterogeneous scene settings. To this end, we evaluate R-Net under three configurations. First, we involve data from all scene settings for training and testing, which we refer to as *Test with Domain Adaptation* (or Test w/ DA). Second, we randomly select data from one scene setting for testing, but excluding them from training, which case is denoted by *Test new samples with Domain Adaptation* (or Test new samples w/ DA). Third, we randomly use the data from one scene setting to train a new network that has the same architecture as R-Net having only the range estimator. We then test this trained network on data from another scene setting; the results are referred to as *Test new samples without Domain Adaptation* (or Test new samples w/o DA). The corresponding results are shown in Fig. 15(a). It can be observed that, with domain adaptation, a median ranging error of only around 25 cm can be achieved even with data from other scene settings that have never participated in the training process. Without domain adaptation, the median error reaches up to 1 m, potentially leading to an even larger localization error. These results strongly confirm that R-Net offers a robust cross-domain ranging performance.

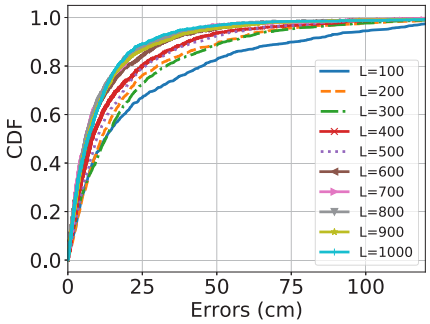
We next study the ranging performance under different parameter settings. We first inspect the impact of FIS sampling interval (i.e., the closest distance among labels of individual FIS in the training data) and the results are shown in Fig. 15(b). Clearly, the ranging performance gains a noticeable improvement if the sampling interval reaches down to 5 cm. We further use Fig. 15(c) to show that the ranging performance can be improved



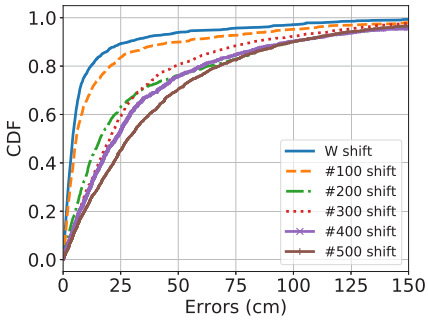
(a) CDF of ranging performance under different settings.



(b) Impact of sampling interval on ranging performance.

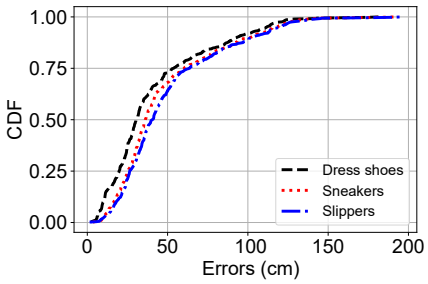


(c) Impact of sample length.

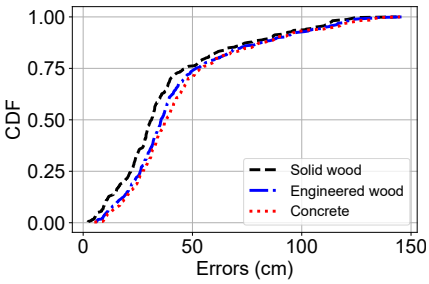


(d) Impact of starting point.

Fig. 15. Ranging performance with R-Net.



(a) Ranging performance for different footwear on engineered wood floor.



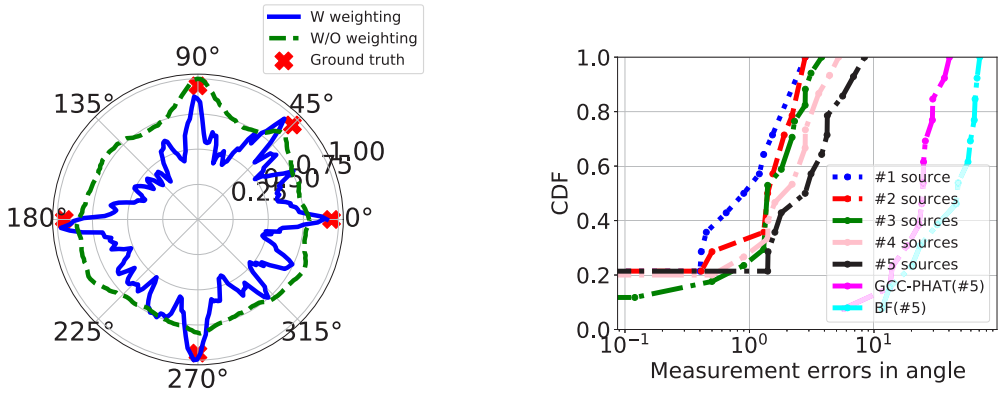
(b) Ranging performance on different floor materials with dress shoes.

Fig. 16. Footwear (a) and floor material (b) cause virtually unnoticeable impacts on the ranging performance.

when the number of samples L increases. However, this improvement becomes marginal when L gets over 500. Consequently, PACE utilizes only 500 samples for ranging. We then verify the impacts of detected starting point of FIS (in terms of the number of shifted samples against what is indicated by our detection strategy) on the ranging performance. As shown in Fig. 15(d), shifting away from our detected starting point can only deteriorate

the ranging performance, firmly proving the correctness of our detection strategy. We finally explore the impact of footwear and floor materials on the ranging performance. Here we differentiate the results of Fig. 15(a) (the red curve) according to the 3 footwear and floor materials explained in Section 5.1. As revealed by the respective performances shown in Fig. 16, footwear and floor materials virtually introduce no impact on ranging. [1]

5.2.2 AoA Estimation. Recall that we adopt air-borned FIS for AoA estimations via beamforming, but we propose a special spectral weighting function to sharpen the correlation peaks. To demonstrate the improved performance of AoA estimation with our spectral weighting function, we conduct an experiment under a rather challenging case where five users walk around the microphone array. Fig. 17(a) depicts a snapshot of the AoA spectrum under different methods: the spectrum with spectral weighting clearly shows five sharp peaks correctly indicating the ground truth AoAs, whereas that without spectral weighting fails to perform correct estimations (e.g., almost no peaks at 0° and 180°).



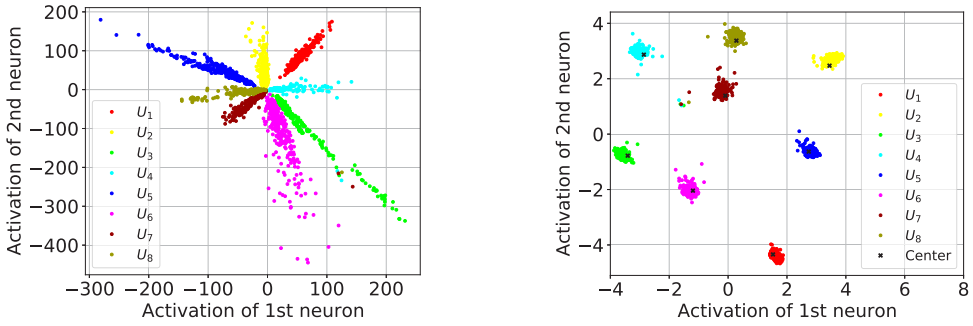
(a) AoA spectrum. (b) CDF of estimation errors.

Fig. 17. AoA estimation performance: (a) AoA spectrum with our proposed weighting function yields much sharper peaks, (b) AoA estimation errors under different number of FIS sources and other solutions.

To further verify its performance, we conduct extensive AoA measurements under different number of FIS sources; the CDF of the measurement errors are shown in Fig. 17(b). It can be observed that the number of sound sources imposes negligible impacts on the performance: the algorithm can still achieve an 80-percentile error of less than 4° in AoA estimation with 5 users, which is 16× improvement over that without spectral weighting (BF) and 7.25× over GCC-PHAT. In comparison, passive RF approach [38] only achieves 80% error of 18° and acoustic solution [45] yields a median error of around 10°. The better performance of PACE can be partially explained by the good correlation property of FIS.

5.2.3 User Identification. As mentioned in Section 4.2, to improve the I-Net's generalization ability so as to discriminate unseen users, we add a center loss to the commonly used cross-entropy loss. The effectiveness of this additional loss term is demonstrated in Fig. 18: by minimizing the intra-class distances while separating the inter-class boundaries, the center loss has greatly enhanced the network's discrimination ability for the 8 involved users. We further present the identification performance with and without center loss under different sampling rates; the results are summarized in Table 1. These results strongly demonstrate that center loss can

[1] We admit that certain extremely cases (e.g., soft carpets that result in very low-strength FIS) may cause troubles to PACE, in which case other non-acoustic methods have to be used as a complement.



(a) Output without center loss.

(b) Output with center loss.

Fig. 18. Activation output of the last hidden layer (b) with and (a) without center loss.

substantially improve the user identification performance by up to 10% on both testing and unseen data. In addition, raising the sampling rate can improve the identification performance for the network trained without center loss, but it appears to have much less impact on that trained with center loss. Note that the identification accuracy of I-Net is achieved using only one step, but Footprintid [35] requires up to 10 steps for identification. Consequently, I-Net significantly reduces the identification latency. Meanwhile, as it is rather unlikely that two users share similar physical conditions and an individual has a very unique foot motion style, footsteps generated by different users always carry distinctive features, rendering I-Net feasible in practice.

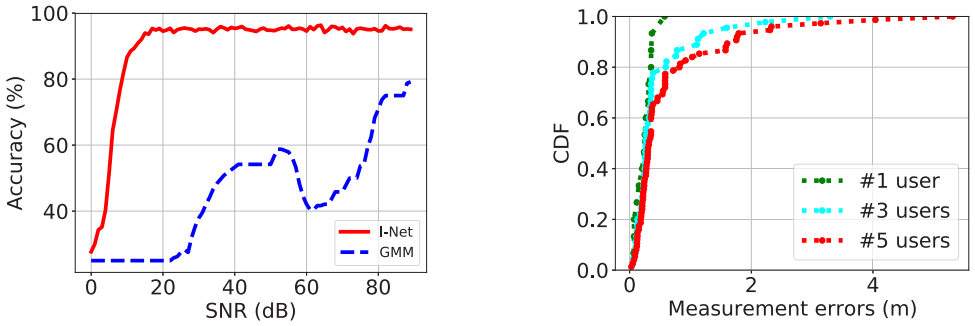
Table 1. Identification accuracy at different sampling rates: the center loss significantly enhances the performance for both observed and unseen users.

Accuracy	**48 kHz**	**96 kHz**	**192 kHz**
Test accuracy w C	92.19%	91.9%	96.02%
Test accuracy w/o C	82.92%	86.01%	88.28%
Unseen data w C	90.7%4	92.11%	91.56%
Unseen data w/o C	78.54%	83.44%	87.93%

We finally compare the identification performance between I-Net and GMM under different SNRs. In order to emulated different SNRs, we superimpose additive white Gaussian noise above FIS captured under realistic settings. The results shown in Fig. 19(a) clearly demonstrate that I-Net achieves significantly better performance than GMM. Since the FIS are captured under common background noise, the overall SNRs of these artificial signals should be worse than what are indicated by the x-axis labels, possibly causing the non-monotonic identification accuracy.

5.3 Localization

After evaluating individual components, we now report the overall localization performance under 5 walking users. We first summarize the localization performance (in terms of error CDF) in Fig. 19(b). The results demonstrate a median error of around 30 cm under a single user, which is comparable to (or even better than) the state-of-the-art passive acoustic and RF solutions (e.g., [45] at 50 cm and [39] at 75 cm). When the number of users increases, potential "collisions" among FIS from different users may lead to worse performance, but the 80-percentile error is still less than 1 m even with 5 users. We then present the tracking performance by showing the traces of two



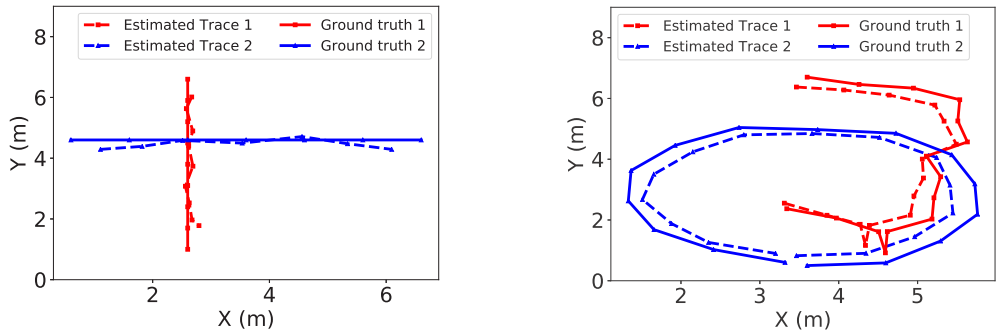
(a) Identification performance under different SNR.

(b) Localization performance under different number of users.

Fig. 19. User identification (a) and multi-user localization (b) performance.

simultaneous walking users under two cases (simple and complex) in Fig. 20. In both simple and complex cases, the estimated traces and ground truth ones are highly consistent. Besides excellent localization performance, the results also show PACE can correctly recognize the user behind each trace: traces do not get messed up even after intersections.

We also study the time complexity of different modules of PACE in Table 2. The results reveal that it costs less than 30 ms to simultaneously locate and identify one user, and this cost is strictly proportional to the number of users. This salient run-time performance is achieved via attentive code optimization. Specifically, we optimize the code for model-based modules using ARM Neon technology [3], a SIMD (single instruction, multiple data) architecture to accelerate hardware run-time performance. This optimization has brought us more than 10× improvements. The sampling rate can affect the overall time cost but the impacts are marginal. When the sampling rate increases, only the time cost of the signal detection module grows, whose upper limit is around 6 ms. The beamformer is sampling rate agnostic as its input has a fixed size. The inputs of R-Net are already obtained with the highest sampling rate 192 kHz. The inputs for I-Net are the STFT of air-borne FIS and have two dimensions, namely time and frequency. Suppose the sampling rate is doubled, the time dimension is also doubled as more



(a) Tracking performance for simple traces.

(b) Tracking performance for complex traces.

Fig. 20. Tracking under multiple users.

Table 2. Computational time cost for each module.

Module	Signal detection	Beam-former	R-Net	I-Net
Time cost (ms)	0.16	18.17	5.29	3.25

samples are involved, but the frequency dimension is halved as we fix the STFT length to 128 points and only take the frequency component below 24kHz. As a result, the input size remains almost unchanged, making the inference time for I-Net also sampling rate agnostic. To conclude, PACE is certainly feasible to conduct real-time localization for multiple users.

6 RELATED WORK AND DISCUSSION

In this section, we briefly survey the closely related literature, so as to strengthen our novel contributions. We also discuss certain limitations upon which PACE may consider to improve in a future work.

6.1 Related Work.

Recent years have witnessed a plethora of indoor localization solutions; they all exploit various signals that carry spatial information to enable localization [1, 15, 17, 23, 25, 29, 38, 39, 51, 57, 60, 62]. Most of these proposals are active (i.e., requiring users to carry a device for sending or receiving some form of signals) [23, 24, 29, 47, 51, 57, 60], but they might not be suitable for certain practical applications (e.g., elderly care or security monitoring) that require passive localization to avoid involving users. Passive RF solutions locate a user via its location-dependent backscattered signals [1, 19, 20, 27, 38, 39], yet they can barely discriminate users. Platypus [17] utilizes human body induced electric potential for passive localization and identification, but it involves a rather heavy infrastructure to cover the concerned area. In the following, we focus on discussing acoustic solutions that have potential to enable passive localization that simultaneously tracks and identifies multiple users.

Although earlier acoustic solutions involve an infrastructure for sending acoustic beacons (i.e., modulated acoustic waveforms) in a synchronized manner and user-held devices to process these beacons [25, 26, 29], recent proposals have advanced in locating users without actively involving them. For instance. CovertBand [33] transmits beacons from an amplified speaker driven by a smartphone, and locates multiple users (even their actions) by analyzing the acoustic reflections received by the phone. Mao et al. [30] adopt a similar approach to CovertBand but use a different modulation technique and a microphone array as the receiver. It utilizes 2D-MUSIC [55] and RNN [16] to track multiple users and to identify their gestures in a room scale. However, by mimicking the function of an active sonar (i.e., transmitting some form of sounds to probe the users), both [33] and [30] are not passive localization solutions, hence not applicable to the indoor scenarios we have in mind.

The most recent proposals VoLoc [45] and Symphony [53] both aim to locate human voice so as to execute location-dependent voice control. They exploit both the direct-path AoA and those of reflect paths to locate a user (voice). While VoLoc targets sequentially identifying these paths by an iterative identification-cancellation procedure to gradually "peel off" all paths, Symphony innovates in leveraging the reflection path to create a virtual microphone array and hence locates multiple users via reverse ray-tracing. While VoLoc and Symphony may well support location-dependent voice control, they are still not purely passive as they require user voice. In comparison, our PACE is among the first passive acoustic solution that can simultaneously locate and identify multiple users. PACE is totally passive as it neither involves users (certainly without user-held devices) nor actively transmits signals to probe. To the best of our knowledge, this joint localization and identification in multi-user scenarios have never been achieved by existing passive localization solutions.

6.2 Discussions.

PACE exploits both model-based and model-free approaches for the localization purpose. One would argue that why not adopt the model-free (i.e., deep neural network) approach to tackle all the problems. The underlying reason is that deep neural network, albeit more powerful, often needs a heavy training process that entails substantial efforts in collecting and labeling data. On the contrary, when a mathematical model is clearly behind a physical phenomenon (e.g., AoA estimation), model-based approaches can be more effective and also easy to implement. As a result, PACE fuses these two methods to strike a balance between model derivation and parameter calibration. Nonetheless, we are actively considering a unified data processing framework for improve the overall efficiency of PACE.

PACE adopts a microphone array, so it has a potential to tackle the situation where an FIE collision (i.e., totally synchronized pace) happens, as many signal processing algorithms (e.g., FastICA [18] or NMF [43]) can decompose mixed sources. However, these algorithms assume that the spectral energies of mixed sources are sparse and not entirely overlapped. Also, each source should have distinct statistical properties over its time duration. Such requirements make these algorithms infeasible to disentangle totally overlapped FIS caused by FIE collision, because FIS have concentrated energy in a rather short duration. Fortunately, this short duration also makes FIE collision less likely to occur; PACE can separate two steps as far as they are 0.1s apart. Since we believe that employing deep learning techniques may better handle this collision, we have planned it as a future study.

PACE is supposed to work in small-scale indoor environments, as explained in Section 1. Its coverage, based on our experience from experiments, can be up to $8 \times 8\text{m}^2$ when the array sits in a corner. Under such an application scenario, user are few (around 5) and they walk independently; these are essential to guarantee the 0.1s separation. In other words, PACE is not applicable to places such as shopping malls or airports, where people may move in groups [44] and hence having synchronized paces. For small spaces larger than the $8 \times 8\text{m}^2$ or having non-rectangular shapes, one can partition them into subareas with each covered by one PACE array.

One rational behind PACE using both air-borne and structure-borne FISs is that these two signal components almost always appear together since their existences are dictated by physical laws. Under extreme case where one of them is missing (due to, for instance, temporary shadowing), we can use adjacent values (precedent and subsequent values) to interpolate missing data. Because i) PACE is meant to locate mobile users who can produce FISs, ii) no object is allowed to cover the PACE array (e.g., a carpet), and iii) a single PACE array does to cover more than one wall-separated area, constant shadowing are rather unlikely to take place. For static users, PACE has to resort to their latest generated FISs for localization.

The whole PACE prototype (including the Raspberry Pi and microphone array) costs less than 120USD. It can be made even cheaper if integrated into, say, Amazon Echo. In our prototype, the array involves 6 microphones, and we use them all for AoA estimations. However, we suspect a 2-microphone array (e.g., Google Home) should also work at a cost of a higher error. With this prototype and considering the average human foot length of 20 to 30cm, the median error of 30cm achieved by PACE is sufficiently accurate for practice use.

7 CONCLUSION

In this paper, we present PACE, the first pure passive acoustic localization system for small-scale indoor scenarios. PACE leverages footstep impact sounds (FIS) to simultaneously locate and identify multiple users without their active involvement. Specifically, PACE builds adversarially adapted deep neural networks to exploit structure-borne FIS for range estimation and air-borne FIS for user identification. Moreover, PACE utilizes feature-rich air-borne FIS for AoA estimation, and it fuses all these information to achieve localization and identification simultaneously. PACE requires no sophisticated hardware design but only a commodity microphone array. It entails a lightweight deployment and achieves a sub-meter level localization accuracy. Therefore, we deem PACE as an effective and efficient solution for our envisioned small-scale indoor applications.

ACKNOWLEDGMENTS

We are grateful to anonymous shepherd and reviewers for their valuable comments. This research is supported in part by AcRF Tier 1 Grant RG17/19 and AcRF Tier 2 Grant MOE2016-T2-2-022.

REFERENCES

[1] Fadel Adib, Zach Kabelac, Dina Katabi, and Robert C. Miller. 2014. 3D Tracking via Body Radio Reflections. In *Proc. of the 11st USENIX NSDI*. 317–329.
[2] X. Anguera, C. Wooters, and J. Hernando. 2007. Acoustic Beamforming for Speaker Diarization of Meetings. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 7 (2007), 2011–2022.
[3] Arm. 2020. Neon Technology. https://developer.arm.com/architectures/instruction-sets/simd-isas/neon.
[4] P. Bahl and V.N. Padmanabhan. 2000. RADAR: An In-Building RF-based User Location and Tracking System. In *Proc. of the 19th IEEE INFOCOM*. 775–784.
[5] M. Berouti, R. Schwartz, and J. Makhoul. 1979. Enhancement of Speech Corrupted by Acoustic Noise. In *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*. 208–211.
[6] Chao Cai, Ruinan Jin, Peng Wang, Liyuan Ye, Hongbo Jiang, and Jun Luo. 2021. PURE: Passive mUlti-peRson idEntification via Deep Footstep Separation and Recognition. https://arxiv.org/abs/2104.07177.
[7] NDT Research Center. 2020. Temperature and the Speed of Sound. https://www.nde-ed.org/EducationResources/HighSchool/Sound/tempandspeed.htm. Accessed: 2020-07-21.
[8] Jagmohan Chauhan, Yining Hu, Suranga Seneviratne, Archan Misra, Aruna Seneviratne, and Youngki Lee. 2017. BreathPrint: Breathing Acoustics-based User Authentication. In *Proc. of the 15th ACM MobiSys*. 278–291.
[9] J. C. Chen, Kung Yao, and R. E. Hudson. 2002. Source Localization and Beamforming. *IEEE Signal Processing Magazine* 19, 2 (2002), 30–39.
[10] Krishna Chintalapudi, Anand Padmanabha Iyer, and Venkata N. Padmanabhan. 2010. Indoor Localization without the Pain. In *Proc. of the 16th ACM MobiCom*. 173–184.
[11] P. Connor and A. Ross. 2018. Biometric Recognition by Gait: A Survey of Modalities and Features. *Elsevier Computer Vision and Image Understanding* 167 (2018), 1–27.
[12] Y. Ephraim and D. Malah. 1985. Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 33, 2 (1985), 443–445.
[13] ESPRESSIF. [n.d.]. ESP32 Wireless SoCs. https://www.espressif.com/en/products/socs/esp32/overview. Accessed: 2020-08-08.
[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-Adversarial Training of Neural Networks. *The Journal of Machine Learning Research (JMLR)* 17, 1 (Jan. 2016), 2096–2030.
[15] Y. Gao, J. Niu, R. Zhou, and G. Xing. 2013. ZiFind: Exploiting Cross-Technology Interference Signatures for Energy-Efficient Indoor Localization. In *Proc. of the 32nd IEEE INFOCOM*. 2940–2948.
[16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.
[17] Tobias Grosse-Puppendahl, Xavier Dellangnol, Christian Hatzfeld, Biying Fu, Mario Kupnik, Arjan Kuijper, Matthias R. Hastall, James Scott, and Marco Gruteser. 2016. Platypus: Indoor Localization and Identification through Sensing of Electric Potential Changes in Human Bodies. In *Proc. of the 14th ACM MobiSys*. 17–30.
[18] A. Hyvarinen. 1999. Fast and Robust Fixed-point Algorithms for Independent Component Analysis. *IEEE Transactions on Neural Networks* 10, 3 (1999), 626–634.
[19] Kiran Joshi, Dinesh Bharadia, Manikanta Kotaru, and Sachin Katti. 2015. WiDeo: Fine-grained Device-free Motion Tracing using RF Backscatter. In *Proc. of the 12th USENIX NSDI*. 189–204.
[20] C. R. Karanam, B. Korany, and Y. Mostofi. 2019. Tracking from One Side: Multi-person Passive Tracking with WiFi Magnitude Measurements. In *Proc. of the 18th ACM/IEEE IPSN*. 181–192.
[21] Hyosu Kim, Anish Byanjankar, Yunxin Liu, Yuanchao Shu, and Insik Shin. 2018. UbiTap: Leveraging Acoustic Dispersion for Ubiquitous Touch Interface on Solid Surfaces. In *Proc. of the 16th ACM SenSys*. 211–223.
[22] C. Knapp and G. Carter. 1976. The Generalized Correlation Method for Estimation of Time Delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 24, 4 (1976), 320–327.
[23] Manikanta Kotaru, Kiran Joshi, Dinesh Bharadia, and Sachin Katti. 2015. SpotFi: Decimeter Level Localization Using WiFi. In *Proc. of the 43th ACM SIGCOMM*. 269–282.
[24] Manikanta Kotaru, Pengyu Zhang, and Sachin Katti. 2017. Localizing Low-Power Backscatter Tags Using Commodity WiFi. In *Proc. of the 13th ACM CoNEXT*. 251–262.

[25] Patrick Lazik, Niranjini Rajagopal, Oliver Shih, Bruno Sinopoli, and Anthony Rowe. 2015. ALPS: A Bluetooth and Ultrasound Platform for Mapping and Localization. In *Proc. of the 13th ACM SenSys*. 73–84.
[26] Patrick Lazik and Anthony Rowe. 2012. Indoor Pseudo-ranging of Mobile Devices Using Ultrasonic Chirps. In *Proc. of the 10th ACM SenSys*. 391–404.
[27] Xiang Li, Shengjie Li, Daqing Zhang, Jie Xiong, Yasha Wang, and Hong Mei. 2016. Dynamic-MUSIC: Accurate Device-Free Indoor Localization. In *Proc. of ACM UbiComp*. 196–207.
[28] Hongbo Liu, Yu Gan, Jie Yang, Simon Sidhom, Yan Wang, Yingying Chen, and Fan Ye. 2012. Push the Limit of WiFi Based Localization for Smartphones. In *Proc. of the 19th ACM MobiCom*. 305–316.
[29] Kaikai Liu, Xinxin Liu, and Xiaolin Li. 2013. Guoguo: Enabling Fine-grained Indoor Localization via Smartphone. In *Proc. of the 11st ACM MobiSys*. 235–248.
[30] Wenguang Mao, Mei Wang, Wei Sun, Lili Qiu, Swadhin Pradhan, and Yi-Chao Chen. 2019. RNN-Based Room Scale Hand Motion Tracking. In *Proc. of the 25th ACM MobiCom*. 1–16.
[31] MarketsandMarkets. [n.d.]. Indoor Location Market - Global Forecast to 2025. https://www.marketsandmarkets.com/Market-Reports/indoor-location-market-989.html. Accessed: 2020-08-08.
[32] S. Nakagawa, L. Wang, and S. Ohtsuka. 2012. Speaker Identification and Verification by Combining MFCC and Phase Information. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 4 (2012), 1085–1095.
[33] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. CovertBand: Activity Information Leakage Using Music. *Proc. of the 17th ACM UbiComp* (2017), 1–24.
[34] S. Narayana, V. Rao, R. V. Prasad, A. K. Kanthila, K. Managundi, L. Mottola, and T. V. Prabhakar. 2020. LOCI: Privacy-aware, Device-free, Low-power Localization of Multiple Persons using IR Sensors. In *Proc. of the 19th ACM/IEEE IPSN*. 121–132.
[35] Shijia Pan, Tong Yu, Mostafa Mirshekari, Jonathon Fagert, Amelie Bonde, Ole J. Mengshoel, Hae Young Noh, and Pei Zhang. 2017. FootprintID: Indoor Pedestrian Identification through Ambient Structural Vibration Sensing. In *Proc. of ACM UbiComp*. 89:1–31.
[36] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-Adversarial Domain Adaptation. In *Proc. of AAAI*. 3934–3941.
[37] J. Portelo, M. Bugalho, I. Trancoso, J. Neto, A. Abad, and A. Serralheiro. 2009. Non-speech Audio Event Detection. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 1973–1976.
[38] Kun Qian, Chenshu Wu, Zheng Yang, Yunhao Liu, and Kyle Jamieson. 2017. Widar: Decimeter-Level Passive Tracking via Velocity Monitoring with Commodity Wi-Fi. In *Proc. of the 18th ACM MobiCom*.
[39] Kun Qian, Chenshu Wu, Yi Zhang, Guidong Zhang, Zheng Yang, and Yunhao Liu. 2018. Widar2.0: Passive Human Tracking with a Single Wi-Fi Link. In *Proc. of the 16th ACM MobiSys*. 350–361.
[40] B. Rafaely. 2005. Analysis and Design of Spherical Microphone Arrays. *IEEE Transactions on Speech and Audio Processing* 13, 1 (2005), 135–143.
[41] RaspberryPi. [n.d.]. ReSpeaker 6-Mic Circular Array Kit for Raspberry Pi. https://wiki.seeedstudio.com/ReSpeaker_6-Mic_Circular_Array_kit_for_Raspberry_Pi/. Accessed: 2020-08-08.
[42] Annie Ross and Germain Ostiguy. 2007. Propagation of the Initial Transient Noise From an Impacted Plate. *Journal of Sound and Vibration - J SOUND VIB* 301 (03 2007), 28–42.
[43] Mikkel Schmidt and Rasmus Olsson. 2006. Single-Channel Speech Separation using Sparse Non-negative Matrix Factorization. In *Proc. of the 9th ISCA INTERSPEECH*. 1652–1658.
[44] Rijurekha Sen, Youngki Lee, Kasthuri Jayarajah, Archan Misra, and Rajesh Krishna Balan. 2014. Grumon: Fast and Accurate Group Monitoring for Heterogeneous Urban Spaces. In *Proc. of the 12th ACM SenSys*. 46–60.
[45] Sheng Shen, Daguan Chen, Yu-Lin Wei, Zhijian Yang, and Romit Roy Choudhury. 2020. Voice Localization Using Nearby Wall Reflections. In *Proc. of the 26th ACM MobiCom*. 1–14.
[46] Y. Shu, C. Bo, G. Shen, C. Zhao, L. Li, and F. Zhao. 2015. Magicol: Indoor Localization Using Pervasive Magnetic Field and Opportunistic WiFi Sensing. *IEEE Journal on Selected Areas in Communications* 33, 7 (2015), 1443–1457.
[47] Yuanchao Shu, Kang G. Shin, Tian He, and Jiming Chen. 2015. Last-Mile Navigation Using Smartphones. In *Proc. of the 21st ACM MobiCom*. 512–524.
[48] Ke Sun, Ting Zhang, Wei Wang, and Lei Xie. 2018. VSkin: Sensing Touch Gestures on Surfaces of Mobile Devices Using Acoustic Signals. In *Proc. of the 24th ACM MobiCom*. 591–605.
[49] TensorFlow. [n.d.]. An End-to-End Open Source Machine Learning Platform. https://www.tensorflow.org/. Accessed: 2020-08-08.
[50] Laurens van der Maaten and Geoffrey Hinton. 2008. Viualizing Data Using t-SNE. *Journal of Machine Learning Research* 9 (11 2008), 2579–2605.
[51] Deepak Vasisht, Swarun Kumar, and Dina Katabi. 2016. Decimeter-level Localization with a Single WiFi Access Point. In *Proc. of the 13th USENIX NSDI*. 165–178.
[52] Anran Wang, Jacob E. Sunshine, and Shyamnath Gollakota. 2019. Contactless Infant Monitoring Using White Noise. In *Proc. of the 25th ACM MobiCom*. 1–16.

[53] Weiguo Wang, Jinming Li, Yuan He, and Yunhao Liu. 2020. Symphony: Localizing Multiple Acoustic Sources with a Single Microphone Array. In *Proc. of the 18th ACM SenSys*. 82–94.
[54] R. Want, A. Hopper, V. Falcao, and J. Gibbons. 1992. The Active Badge Location System. *ACM Trans. on Information Systems* 40, 1 (Jan. 1992), 91–102.
[55] M. Wax, Tie-Jun Shan, and T. Kailath. 1984. Spatio-Temporal Apectral Analysis by Eigenstructure Methods. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 4 (1984), 817–827.
[56] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. In *Proc. of ECCV*. 499–515.
[57] Bo Xie, Guang Tan, and Tian He. 2015. SpinLight: A High Accuracy and Robust Light Positioning System for Indoor Applications. In *Proc. of the 13th ACM Sensys*. 211–223.
[58] Moustafa Youssef and Ashok Agrawala. 2005. The Horus WLAN Location Determination System. In *Proc. of the 3rd ACM MobiSys*. 205–218.
[59] Sangki Yun, Yi-Chao Chen, Huihuang Zheng, Lili Qiu, and Wenguang Mao. 2017. Strata: Fine-Grained Acoustic-based Device-Free Tracking. In *Proc. of the 15th ACM MobiSys*. 15–28.
[60] Chi Zhang, Feng Li, Jun Luo, and Ying He. 2014. iLocScan: Harnessing Multipath for Simultaneous Indoor Source Localization and Space Scanning. In *Proc. of the 12th ACM SenSys*. 91—104.
[61] Chi Zhang, Kalyan P. Subbu, Jun Luo, and Jianxin Wu. 2015. GROPING: Geomagnetism and cROwdsensing Powered Indoor NaviGation. *IEEE Transactions on Mobile Computing* 14, 2 (2015), 387–400.
[62] Shilin Zhu and Xinyu Zhang. 2017. Enabling High-Precision Visible Light Localization in Today's Buildings. In *Proc. of the 15th ACM MobiSys*. 96–108.